Progress Report

Matt Smith

November 25, 2012

1 Project Description

This project, ultimately, aims to provide a simulation of individuals attempting to give up smoking in the situation of a social network. In particular, focus is given to how the network manipulates their behaviour and whether some social organisations and constructs are more suited to triggering the breaking of the smoking habit. To analyse this, a model of both a human and a social network will be created, through which a given period of time can be simulated - the result of this being data that can be analysed to find the previously mentioned social conditions conducive to give up smoking.

2 Progress Summary

On the whole, the project continued to be consistent with the project specification, insofar as rather than large changes having to be made, specific details and implementation methods are now known, and there is a greater understanding about the theoretical basis of the models. No development towards the final model has been started - to date, experimental code has been written as proof-of-concept programs to examine different approaches. These pieces of work will, however, form the basis for parts of the final model.

3 Model Research Progress

3.1 Human Representation

As modelling a human (also referred to as a *node*) in general is a difficult task, the scope of the project can be reduced significantly by ensuring that only aspects of human behaviour related to the analysis of smoking are used. Furthermore, some assumptions about the way in which nodes behave have to be made to prevent the project encompassing a large part of sociological and psychological theory. Due to this, it has been decided that the some human behaviours will be represented in a probabilistic manner - the reasons for this are twofold. Firstly, it provides a simple method of recreating some facets of actions. For example, a person could rate, on a scale of one to ten, how likely they are to start smoking in the next few days. Although this is not necessarily an accurate measure, it is enough for a simplistic representation. Example attributes include:

- *Susceptibility to peer-pressure* - peer-pressure may over-ride other considerations
- *Likeliness to smoke* - humans may have an internal opinion on smoking that guides them.
- *Willpower* - how likely they are to continue their giving-up effort.
- *Number of Cigarettes Per Day* - a higher number may make giving up harder.
- *Ability to Influence* - hub nodes may be influential, and as such be powerful.

Some other attributes, however, cannot be representing by probabilities. These include items such as how many cigarettes someone smokes, whether someone smokes at all and how likely they are to reach out

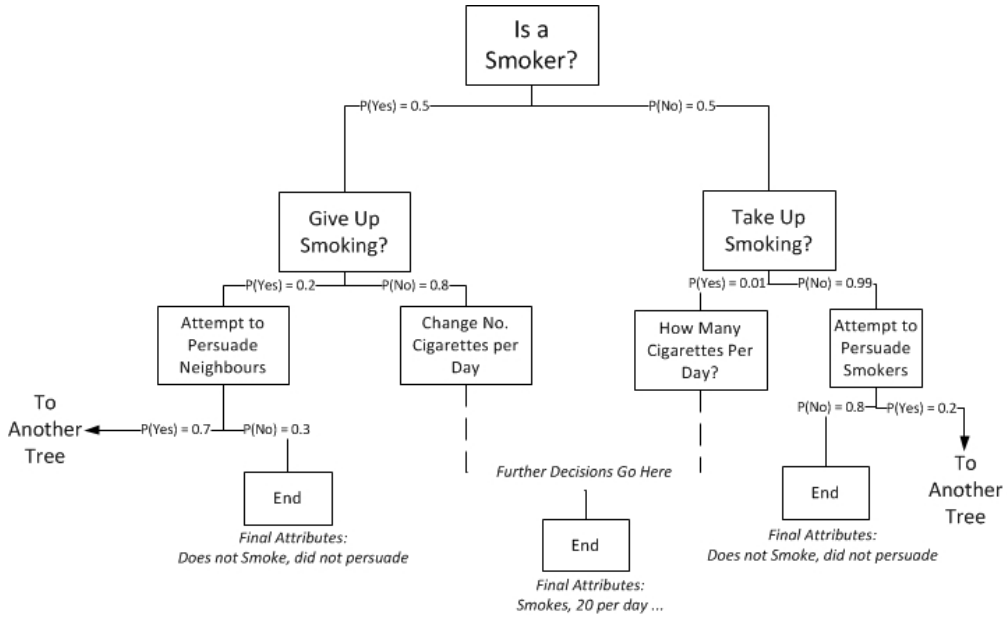

Figure 1: Example Decision Tree

beyond their local network. This kind of attribute will not be normalised, but instead will be grouped as necessary, or handled in a raw form.

Over time, decisions will be made that relate to these attributes, and to represent this series of decisions, a decision tree will be used. This is a structure that maps series of choices that the person can make to a tree, leading them to some state consisting of a collection of attribute values. The key advantages of this are that it allows a fixed number of choices per cycles, with the ability to force the structure of the choices (for example, important ones at the top). The specific structure and content of this tree is largely a developmental matter and needs to be tuned over time, but a concept tree can be seen in figure 1.

As such, the tree and its contents will be researched whilst the overall model development is in progress. Currently, the planned approach is to divide separate concerns (i.e. decisions on smoking would be a separate tree to decisions about whether to try and make new connections) and produce a tree for each of these. Certain series of decisions in one tree may trigger another, as seen in fig. 1, but this may be changed if it does not perform as expected.

In terms of the model-at-large, it is important for the network to be able to reconfigure itself over time, due to the notion of homophily [4] This is the concept of people with similar interests since people move towards to others with similar interests whilst at the same time, friends will develop similar interests within the relationship, leading to a split between influence and reconfiguration. The former is a major focus of this project, since understanding how the influence flows around a social network is key to concluding how the network impacts behaviour.

Up to this point, only the human as a single entity has been considered - to model effects of interacting with others, influence must be modelled. In the current plans, the *Linear Threshold* model will be used to implement influence - it is based upon the idea of the number of neighbours for a given node holding an attribute or performing an action will increase the chance of the node also taking on that attribute/performing the action. In addition, the previously mentioned decision tree can either incorporate or have a standalone tree for the impact of incoming influence from surrounding nodes. Once more, this will be done probabilistically, where the edge weights either dampen or emphasise the influence passing through them (see below).

Reconfiguration is also important, but is much harder to accurately model - to maintain a reasonable project scope, this will not be as formal as an implementation. It is likely that a decision tree will be created that will choose and negotiate relationships with other nodes, with the aim of building a model that will emulate the previously mentioned concept of homophily.

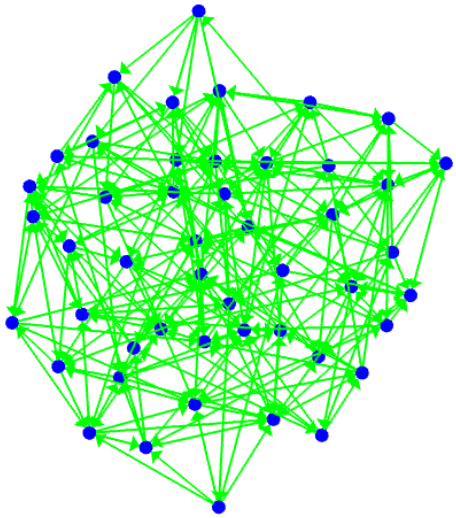

Figure 2: Small-World Network - 50 nodes, $\beta = 0.5$, K = 10 - Layout: Fruchterman-Reingold Algorithm

To add into the strictly rational aspect of the human model, some representation of irrationality will be researched and included. Elster [2] mentions how irrationality can often be a deviation from the rational course of action, which fits in well with the decision tree model. One possible programmatic approach is to work some random factors into the decision trees, triggering choices that would not usually be the rational decision. Obviously, this would be done in moderated way and tuned to avoid it becoming a regularity and hence causing it to no longer be irrational.

In terms of the items that propagate, attributes are a particular focus - the idea of others affecting your choices is key to this process. Example attributes that would propagate are whether someone smokes, and the number of cigarettes someone smokers, per day - if person's friend smokes heavily, then the person may increase to allow them to spend more time with their friend. As such, the probabilities used in the decision trees will be manipulated by the nodes connected to a given person. The overall aim is that these manipulations will represent the overt and subconscious actions of others that encourage someone to choose a particular action.

3.2 Network Representation

One of the key factors in the success of this project was to be able to generate a realistic social network, as without this, the model would be unable to produce meaningful results. To date, two methods have come out as useful in the research - *small-world* and *scale-free*, the latter being a type of *preferential attachment* network.

Small-world networks are created using the Watts & Strogatz model [5], and produce graphs similar to the one shown below in figure 2. In general, the method uses a degree averaging technique to give all members in the graph roughly the same number of edges. As implied by the name, this method is strongly related to the idea that most people know each other, or if not directly, through a small number of connections - as such, this is not entirely useful to this project as it is difficult to realistically model the environment of larger groups of people.

In a larger graph, it is highly unlikely that most people know each other within a few connections, so the creation of a scale-free network (using the *Barabási-Albert* model [1]) succeeds here. This uses a *preferential*

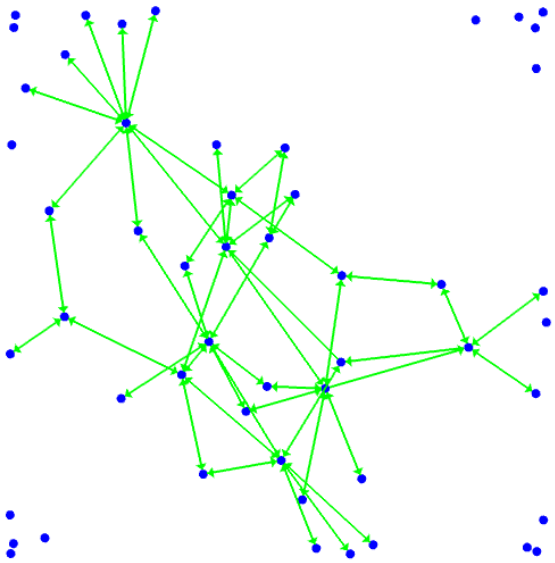

Figure 3: Scale-Free Network - 50 nodes, starting with 4 vertices, 3 edges - Layout: Fruchterman-Reingold Algorithm

attachment method - "the rich get richer" - to generate networks which instead have a number of 'hub' nodes, similar to popular figures in friendship circles. Additionally, the graph is significantly less connected, resulting in a seemingly more realistic network. As can be seen in figure 3, a basic scale-free graph of the same number of nodes exhibits features such as 'hubs' and groups of nodes. Furthermore, when increased to 250 nodes(fig. 4), these features are emphasised. With some modification of this algorithm, networks produced should approach realistic social networks. Note that there are a lot of outliers - this would have to be accounted for in modifications, for example by removing a certain percentage of them. Another major advantage of generating scale free networks is that they require a base network. As such, structures of interest can be used as this base so that it is known they are present as part of a network which can then be analysed for how they affect/are affected by the greater environment.

Both methods have advantages in a certain set of situations, so it is worth including the ability to generate and simulate upon either. Examples of a *small-world* network being useful is when examining tightly knit but larger networks, whereas *scale-free* networks are ideal for a more general network. The methods in general will be constantly adapted and tweaked to behave in similar ways to real networks, which will be measured through metrics such as betweenness, network diameter and average path length.

Alongside the generated networks, sampled networks will also be gathered - by combining these with a set of generated networks, developmental models of components such as the decision tree can be tuned with the ability to reproduce model behaviour. Working from this, larger data sets can be both gathered from real datasets and computer-generated to provide a wide range of networks usable in the final analysis.

Moving from the whole network to individual edges, relationships in a social network map well to a directed graph. This is because relationships are not necessarily symmetric - assuming they are removes the concept of role-models, an important feature of social networks, and also blocks another key concept, that of different people being influenced by others to a different degree.

Using this idea, it has been decided that edge-weight should represent the 'influence factor' of that relationship - if node A has an edge of weight 0.5 to node B, then any behaviours that traverse this edge

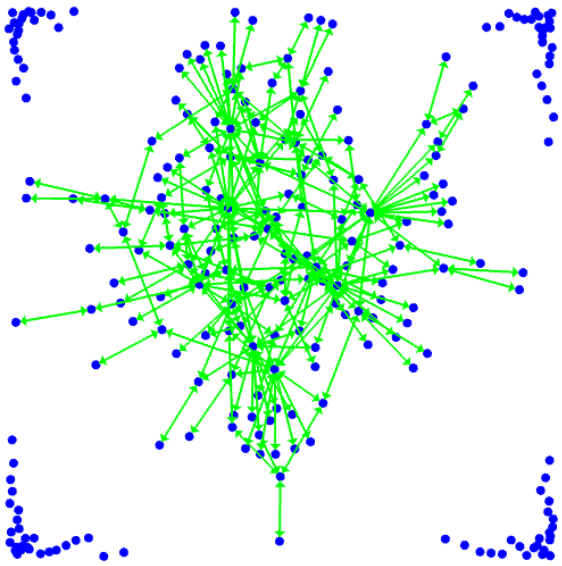

Figure 4: Scale-Free Network - 250 nodes, starting with 4 vertices, 3 edges - Layout: Fruchterman-Reingold Algorithm

will only do so with an impact of half the original. Firstly, this reinforces the idea of role-models (a person is highly influenced by another, but the other is not necessarily influenced, or even knows, the first). It also provides a more realistic approximation of how a friendship works, since although friends, one person may think more of the other and as such tend to be influenced more easily by them. Values between 0 and 1 will be used here to maintain the probabilistic approach.

3.3 User Features

There are two main user features that are to be included in the project, to date. The first is the generation of networks based on loose statistics - this, when combined with both the previously mentioned network generators and the human model, will be a relatively trivial interface between the initialisation of the simulation environment and the user, so should cause no problems in the development phase.

The other feature is the ability to parse and export existing graphs - again, as mentioned above, work has gone into parsing datasets and outputting them as *GraphML* - to work this into the simulation setup should be relatively simple as it is based on code already written. A particularly useful feature would be to export the network to *GraphML* at given simulation cycle, to allow for static analysis of the network at a given time. Within this, there should be some way to indicate humans which are user triggered, in that the user can force certain attributes at certain points.

3.4 Commercial Model Integration

At this stage, it would not be suitable to consider whether the model will fit into a larger, commercial-level model as not enough development of the model itself has been completed. It should be noted, however, that up to now, there have been no issues that would appear to cause any problems.

3.5 Visualisation

Along with simulating the human and network behaviour, it is important to be able to view snapshots of the model to understand what is happening. So far, this has come via two avenues - *Gephi* and the *Repast GUI*. Gephi is particularly useful when it comes to generating standalone networks for analysis - examples of this include inspecting algorithm-generated networks for realism and checking if a sampling method is working correctly, producing a reasonable graph. The *Repast GUI* provides a similar service but is refreshed on each simulation tick. Due to this, the GUI effectively gives a live view of the network and how it is behaving - of course, this is has some more simplistic tools than *Gephi*, so works in a complementary way.

3.6 Validation & Analysis of Model

During development, the model must be validated against realistic figures. This will be done in a number of ways, involving comparing it to governmental statistics [3], other models (for example by SandTable), and methods that will arise during development. The general aim of validating the model is to ensure that, for known graphs/simulation starting points, it behaves in an expected manner, i.e. providing the same results as other runs within a degree of certainty.

Although final analysis methods are yet to be decided, great care has been taken to ensure that there is as larger range as tools as possible available for analysis, and as such stopping any implementation based restrictions on these tools. An example of this include building a *GraphML* serialiser for *JUNG* networks, to allow for *Gephi*/*Cytoscape* analysis of these graphs. More tools for analysis will be developed, whilst research into the ideal methods will be undertaken throughout the implementation. Considering final analysis methods, statistical methods are likely to form the basis of comparing realistic models to the model produced as part of this project - this will provide a numerical approach to evaluating the 'realism' of the networks and nodes involved.

4 Timetable

The timetable from the current point onwards can be seen in fig. 4. Note that certain aspects of the work will be developed continuously. These are:

- Network reconfiguration and generation, as this depends on the influence model and decision tree.
- Model validation against real statistics, such as NHS smoking figures.
- Documentation, by way of a development log and working towards the final report.
- Research into the causes of smoking, social network influence and rationality.

These aspects of the project drive or are affected by other sections so must continuously be developed and checked throughout.

Week No.	Week Beginning	Tasks
9	26th Nov	Research & experiment with decision trees, read into rationality.
10	3rd Dec	Experiment with network reconfiguration and work on tools.
11	10th Dec	Set up development environment and design model structure.
12	17th Dec	Build model structure, with focus on extendability using modular design.
13	24th Dec	Expand upon existing network generators, parsers, exporters and build automated setup tools.
14-15	7th Jan	Implement decision tree basics and the core of an influence model.
16	14th Jan	Analyse previous work (with focus on decision tree) and tune, as well as working more attributes into the human model. Begin to build analysis tools (i.e. graph dumps, node tracking, automated running, forcing specific structures).
17	21 Jan	Further tune decision tree, incorporating rationality/irrationality and user-specified trigger humans (i.e. certain user-given characteristics that cause them to give up). Continue analysis tool development, running some basic simulations and analysing these.
18	28th Jan	Tune and assess network reconfiguration on model so far. Begin to run simulations.
19-20	4th Feb	Run more simulations and analyse data.
21	18th Feb	Further data analysis and presentation preparation.
22	25th Feb	Presentation of project so far.
23-30	4th Mar	Write the final version of the report, with further data analysis if necessary.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[2] Jon Elster. *Explaining Social Behavior.*

[3] Health and Lifestyle Statistics Social Care Information Centre (HSCIC). Statistics on smoking: England 2012, August 2012.

[4] Charles Kadushin. *Understanding Social Networks.*

[5] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. 393:440–442, June 1998.