European Journal of Operational Research 265 (2018) 263–276

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

Predicting adolescent social networks to stop smoking in secondary schools

Angelico Fetta, Paul Harper*, Vincent Knight, Janet Williams

School of Mathematics, Cardiff University, Cardiff CF24 4AG, United Kingdom

ARTICLE INFO

Article history:
Received 14 February 2017
Accepted 11 July 2017
Available online 12 August 2017

Keywords:
OR in health services
Behavioural OR
Social networks
Agent based simulation
Link prediction

ABSTRACT

Social networks are increasingly being investigated in the context of individual behaviours. Research suggests that friendship connections have the ability to influence individual actions, change personal opinions and subsequently impact upon personal wellbeing. This paper explores the effect of individual friendship selection decisions, and the impact they may have on the overall evolution of a social network. Using data from a large smoking cessation programme in secondary schools, an agent based simulation aiming to predict the evolution of the adolescent social networks is created. The simulation uses existing friendship selection algorithms from link prediction literature, along with a new approach to link prediction, termed PageRank-Max. This new algorithm is based upon the optimisation of an individuals eigen-centrality, and is found to be more successful than existing methods at predicting the future state of an adolescent social network. This research highlights the importance of eigen-centrality in adolescent friendship decisions, and the use of agent-based simulation to conduct behavioural investigations. Furthermore, it provides a proof-of-concept for targeted interventions driven by social network analysis, demonstrating the utility of using emerging sources of social network data for public heath interventions such as with tobacco use which is a major global health challenge.

1. Introduction

Investigation into individual behaviours in relation to social networks has experienced substantial growth in recent years. This is in part due to the availability of social network data as a result of social networking sites such as Facebook, Twitter and Google+, and the computing advancements that allow for the exploration of such large data sets (Kwak, Lee, Park, & Moon, 2010; Mislove, Koppula, Gummadi, Druschel, & Bhattacharjee, 2008; Salter-Townshend, 2012).

This paper is concerned with the individual decisions that cause social network evolution in adolescents, which is applied to data from a large smoking cessation programme in secondary schools. Smoking is a major global health challenge and tobacco use is said to kill 6 million people worldwide per year (World Health Organisation, 2015). More than 5 million of those deaths are the result of direct tobacco use while more than 600,000 are the result of non-smokers being exposed to second-hand smoke. Secondary schools are a common point at which people start smoking with, for example, two-thirds of smokers in the UK starting before the age of 18 (Action on Smoking & Health (ASH), 2016). Quitting smoking is notoriously difficult; among all current U.S. adult cigarette smokers, nearly 7 out of every 10 (68.8%) reported that they wanted to quit but were so far unable to do so (Centers for Disease Control & Prevention, 2016). Smoking increases the risk for serious health problems, many diseases, and death (Centers for Disease Control & Prevention, 2014).

The theory of friendship decisions amongst adolescents has been widely researched, with factors such as proximity (Festinger, Back, & Schachter, 1950), reciprocation (Parker & Seal, 1996) and similarity (McPherson et al., 2001) discussed as important. Often studies such as these are based on qualitative evidence, with scientific experts drawing conclusions based on retrospective analysis. Our research discusses the development of an Agent Based Simulation (ABS) model which allows for the testing of behavioural theory relating to friendship. Through the use of specifically selected algorithms, drawn from the link prediction literature, a predicted future state of a social network can be made. The predicted future social network may then be compared with the real social network for accuracy, with conclusions drawn around the implemented behavioural theory.

* Corresponding author.
E-mail address: harper@cardiff.ac.uk (P. Harper).

http://dx.doi.org/10.1016/j.ejor.2017.07.039

Simulation provides a tool to explore the evolution of a system, scrutinise theory and evaluate potential outcomes. Within the domain of OR, simulation is a core tool utilised for research – lending itself to applications such as manufacturing,defence and healthcare (Pidd, 2004). ABS is a particular paradigm of simulation, which aims to take an individualistic view of system evolution (An, 2012). ABS is a micro-simulation technique, modelling the individual behaviours of specific objects in a system to understand the emergent global phenomena (Niazi & Hussain, 2011).

ABS investigations related to social networks have covered a variety of topics. Epidemiology in particular has adopted ABS techniques to explore the spread of infectious diseases through networks, including HIV spread in Amsterdam (Mei, Sloot, Quax, Zhu, & Wang, 2010), Influenza in a metropolitan social network (Mao, 2014) and H1N1 on a Chinese university campus (Mei et al., 2010). ABS has also been used in the investigation of network structure, as opposed to its effects, although the number of papers in this area is far fewer. Pujol, Sanguesa, and Delgado (2002) uses agents to extract *reputation* in a social network topology, Han, Zhao, Hadzibeganovic, and Wang (2014) explores hierarchical geographical network structures and Bernstein and O'Brien (2013) uses ABS to generate 'realistic' social network data sets; however, these studies do not utlise empirical social network data. Given the individual perspective of ABS, and the ability to quantify the impact to a system as a result of the interactions of constituent parts, ABS appears an appropriate method to explore the behavioural factors influencing the evolution of adolescent social networks.

The motivation to adopt a quantitative simulation-based research approach to adolescent friendships, as presented in this paper, is that it appears to be an unexplored niche in social network literature. More specifically, the ability to implement link prediction methods in an ABS framework for adolescent social networks, provides a novel contribution to the literature. Furthermore we provide a proof-of-concept for targetted interventions driven by social network analysis, demonstrating the utility of using emerging sources of social network data for public heath interventions.

This research also contributes to the growing body of work in Behavioural Operational Research (BOR) which is defined as the study of behavioural aspects related to the use of OR methods in modelling, problem solving and decision support (Hämäläinen, Luoma, & Saarinen, 2013). BOR may broadly be considered within three categories: behaviour in models (methods), behaviour with models (actors) and behaviour beyond models (praxis) (Franco & Hämäläinen, 2017). Our work is firmly grounded in incorporating behaviour within models (methods). Furthemore, as comprehensive reviews of the application of OR to healthcare (Brailsford, Harper, Patel, & Pitt, 2009; Hulshof, Kortbeek, Boucherie, Hans, & Bakker, 2012) reveal, relatively little prior consideration has been devoted to behavioural aspects in this field. Hence this paper also aims to demonstrate the use of BOR for healthcare applications.

The remainder of the paper is structured as follows. In Section 2 we introduce the data from the smoking in schools programme. Section 3 outlines the chosen network structures utilised within this research, whilst link prediction methods are introduced in Section 4. The developed ABS is described in Section 5. A new method for link prediction, *PageRank-Max*, is proposed in Section 6, validated in Section 7, and compared against the other methods in the results in Section 8. Conclusions are made in Section 9.

2. Case study

There are signficant global challenges to reducing smoking from a public health perspective. The World Health Organziation (WHO) has created the Tobacco Free Initiative (TFI) which aims to "reduce the global burden of disease and death caused by tobacco, thereby protecting present and future generations from the devastating health, social, environmental and economic consequences of tobacco use and exposure to tobacco smoke" (World Health Organisation, 2016). Many of the TFI's actions are aimed at adolescents given that this is a common time in life at which people start smoking. It is therefore vital to intervene at this age given the addictive nature of tobacco and the longer-term health effects.

Our conceptual approach to the problem is in predicting social networks to help with more targetted interventions to reduce the uptake of smoking amongst adolescents. The case study data is taken from "A Stop Smoking in Schools Trial" (ASSIST) and explores the effects of social networks upon attitudes toward adolescent smoking, with a view to inform potential cessation proliferation methods. Formed through a joint venture between Cardiff University Institute of Society, Health and Ethics and The Department of Social Medicine at the University of Bristol, UK, ASSIST was designed as a peer-led intervention, formulated around the 'Gay Hero' work of Kelly et al. (1992). Schools from across the West of England and South Wales were recruited to the study, through stratified randomisation, following a cohort of Year 8 students (12–13 year olds) over the course of a three and a half year period (Holliday, 2006).

Three waves of social network data were collected at one year intervals for 18 schools in the study. Each participant was asked to name up to six other students with whom they shared a friendship. From this data, a school based social network may be constructed, describing friendship evolution over the course of the study. The students' ability to only identify up to six friendships may be considered a limitation of the study; however, the work of Kirke (1996) and Pearson and Michell (2000) suggest that friendships ranked below the top six connections do not carry equal significance. Additionally, the average number of friendship nominations in the data across the three time points was calculated as 3.8 (T_1), 4.3 (T_2) and 3.8 (T_3) – suggesting students often did not opt to maximise their number of friendship nominations. Given the objective of this research is to predict social network structure to identify future influence, the friendship nomination limit is unlikely to substantially impact the conclusions of this research.

From the 18 schools, 12 are classified as control and 6 as intervention. Identified socially prominent individuals in adolescent social networks within the intervention schools were given training to diffuse a 'stop smoking' message to their peers (Audrey, Cordall, Moore, Cohen, & Campbell, 2004). An example of the data from one school may be observed in Figs. 1 and 2 demonstrating the evolution of the social network over time (friendship network at year 1 and 2 for Figs. 1 and 2 respectively).

Figs. 1 and 2 show network patterns and evolutions that were seen in many of the control schools. That is, over time the prevalence of smoking increases and that smokers tend to cluster together as friends. The findings of Campbell et al. (2008) suggest a reduced smoking prevalence in intervention schools in the early stages of the trial. Overall, the researchers concluded that ASSIST was a success, providing a cost-effective method for increasing adolescent smoking cessation (Hollingworth et al., 2012).

3. Network structures

This section introduces the essential graph theoretic and network science definitions that are used to inform the research in the development of the ABS (Section 4). As our study is concerned with the investigation of social networks, and ultimately the development of a new algorithm to predict social network evolution, the relevant metrics to analyse and interpret network structures are required.

An undirected *Graph* is defined as a pair $G = (V, E)$ of sets such that E is a subset of the unordered pairs of V, where V is the set

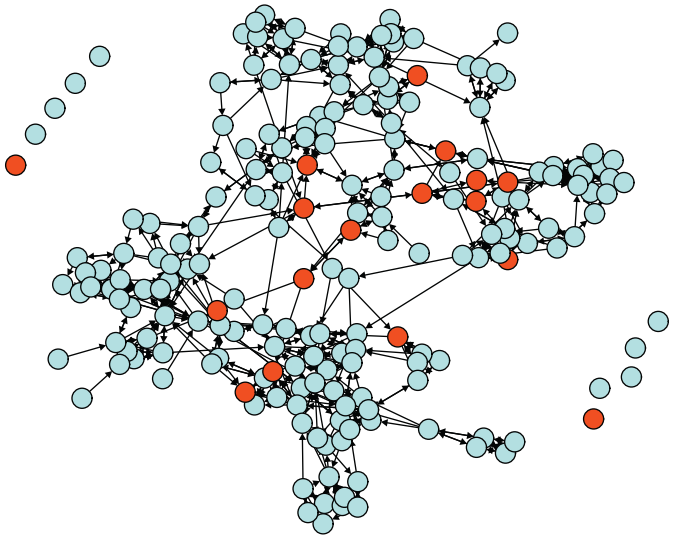

Fig. 1. Social network at T_1; dark nodes indicate smokers.

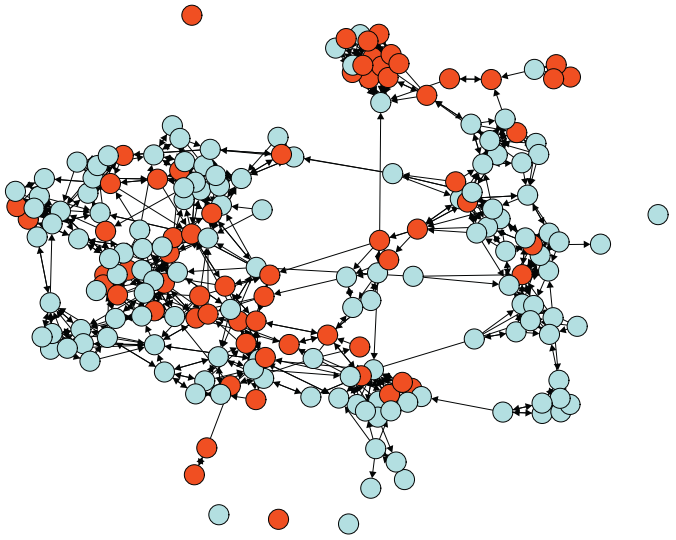

Fig. 2. Social network at T_2; dark nodes indicate smokers.

of vertices (or nodes) and E represents the set of edges (or links). A directed graph (or digraph) may be defined in the same manner, except that E is a subset of the ordered pairs of V.

The *order* of G is defined as the number of elements in the set of vertices V, denoted by $|G|$; thus $|G| = |V(G)|$ (Bollobas, 2013). For simplicity, the number of vertices for any particular graph G, shall be referred to as n. A social network may be represented as a *directed* or *undirected* graph. A directed graph offers a rich source of information, both in terms of the qualitative implications of friendship, and the quantitative metrics of network calculation.

For an undirected graph, an edge $\{i, j\}$ links the vertices v_i and v_j and may be represented by ij. A directed network edge preserves the order by which a link is made, such that an edge $\{i, j\}$ implies a link from v_i to v_j is denoted by $i \to j$, therefore it cannot be assumed the link $j \to i$ exists. A number of the metrics defined later require the maximum number of edges (e_{max}) of a graph; for an undirected graph, $e_{max} = \frac{n(n-1)}{2}$, and for an directed graph, $e_{max} = n(n-1)$.

With the basic elements of a graph defined, we next introduce four commonly used network metrics that we utilise (in the results Section 7) to compare the performance of predicting social network evolution under different link prediction measures (which will be defined in later sections).

3.1. Average degree

The *degree* of a vertex v_i is denoted as $deg(v_i)$ and represents the number of incident edges of v_i. In a directed network, these may be separated further in terms of *in-degree* $deg(v_i)_{in}$ and *out-degree* $deg(v_i)_{out}$, defined as the count of the inward links and outward links of v_i respectively (Newman, 2003).

In terms of network cohesion, and a representation of the graph as a whole, the *average vertex degree* may therefore be calculated by:

$$\frac{\sum_{i=1}^{n} deg(v_i)}{n} \tag{1}$$

where $deg(v_i)$ is replaced by the directed network equivalent (if required).

3.2. Reciprocity

A directed graph's in-degrees and out-degrees allows for incident edges to become unreciprocated. In terms of a social network, this could suggest the node v_i extending a link to v_j but the link $j \to i$ not being in existence. This provides a representation of network cohesion, termed reciprocity.

A *reciprocated* tie is one in which for the vertices v_i and v_j, the links $i \to j$ and $j \to i$ exist. The overall *reciprocity* of the directed graph G is said to be:

$$r = \frac{|L|}{|E|} \tag{2}$$

where L is the set of edges involved in reciprocal ties. As such, $r \in [0, 1]$, meaning that $r = 1$ signifies a fully reciprocated graph (Newman, Forrest, & Balthrop, 2002). Both average degree and reciprocity are used in this paper to measure network cohesion.

3.3. Transivity ratio

For a directed graph, a *transitive triple* is defined to be a sequence of edges such that $i \to j$, $j \to k$ and $i \to k$ exist (Wasserman & Faust, 1994). A *subgraph* is defined as $G' = (V', E')$ of $G(V, E)$ if $V' \subset V$ and $E' \subset E$. In an undirected graph, a *triangle* may be considered as a complete subgraph containing three nodes of G, where the number of triangles containing v_i is defined to be $\delta(v_i) = |\{\{v_i, v_j\} \in E : \{v_j, v_k\}, \{v_i, v_k\} \in E\}|$ (Schank & Wagner, 2005). The number of all possible triangles in G is denoted by $\tau(G)$, therefore the *transitivity ratio* $T(G)$ may be calculated by:

$$T(G) = \frac{\sum_{i=1}^{n} \delta(v_i)}{\tau(G)} \tag{3}$$

For a directed graph, edges are converted into undirected associations (Luce & Perry, 1949).

This measurement calculates the proportion of "closed triangles" of nodes, in relation to all connected triples of nodes. This gives a representation of how clustered the network is, offering an indication of mutual relations. Other interpretations of graph transitivity have been suggested; for example, the global clustering coefficient and the local clustering coefficient (Watts & Strogatz, 1998), both of which are said to suffer from bias (Soffer & Vázquez, 2005). Given its overall simplistic and effective nature, coupled with the avoidance of inherent bias associated with other methods, the transitivity ratio has therefore been selected as the metric of choice for quantifying network clustering within this research.

3.4. Average path length

Travelling a concourse of nodes via a graph's incident edges is described as navigating a *path*. A *path* is a graph P of form $V(P) = \{v_0, v_1, ...v_l\}$, with edges $E(P) = \{v_0v_1, v_1v_2, \ldots, v_{l-1}v_l\}$, denoted by $v_0v_1, \ldots, v_l$. The end vertices are v_0 and v_l, therefore the path may be denoted by $v_0 - v_l$. In a directed graph, the direction of the edges dictate the direction of the path (Bollobas, 2013).

The path of a network plays an important role in the description of reachability between nodes. For example, if a path exists between the nodes v_i and v_j then these nodes are said to be *reachable* (Holme, 2005). In a fully connected graph, every node is reachable. Social Networks are unlikely to ever achieve complete reachability, even less so if the network is directed (Barabási, Albert, & Jeong, 2000). To garner an overall picture of the reachability between paths of nodes, one must consider the *geodesic*, the shortest path connecting two vertices v_i and v_j (Harary, 1994).

The *Average Path Length* (APL) l_G for G is described as the shortest distance between the nodes v_i and v_j, denoted as $d(v_i, v_j)$, divided by the maximum possible number of edges (e_{max}) (Newman, 2001). A disconnected APL assumes $d(v_i, v_j) = 0$ if $v_i = v_j$ and $d(v_i, v_j) = n$ if v_i cannot reach v_j . Therefore:

$$l_G = \frac{\sum_{i \neq j} d(v_i, v_j)}{e_{max}} \tag{4}$$

APL is a robust measurement of network topology, often quoted as the main factor in the classification of network type (Fronczak, Fronczak, & Holyst, 2004).

4. Link prediction algorithms

Link prediction is the process of attempting to foresee connections that are yet to be established (Liben-Nowell & Kleinberg, 2007). Given a graph $G_t(V, E)$ of n nodes/vertices (V with vertices v_i) and a set of links/edges (E with edges e_i) at time t, an attempt is made to arrive at G_{t+1} through the evaluation of possible new edges, $e_{i,j}$ between vertices v_i and v_j.

Link prediction algorithms have a variety of applications, including: optimisation of website navigation (Zhu, Hong, & Hughes, 2004), the recommendation of content to web users (recommender systems) (Huang, Li, & Chen, 2005), and the acceleration of academic collaboration (Farrell, Campbell, & Myagmar, 2005). Methods employed in conjunction with the link prediction problem include machine learning (Goldenberg, Kubica, & Komarek, 2003; Hasan, Chaoji, Salem, & Zaki, 2006), Markov methods (Domingos & Richardson, 2007; Taskar, Wong, Abbeel, & Koller, 2004) and statistical inference (Popescul & Ungar, 2003). It is widely accepted that the task of accurately predicting links is difficult (Getoor, 2003; Taskar et al., 2004), in part due to the a priori probability of a link being small (Getoor & Diehl, 2005).

The seminal paper (Liben-Nowell & Kleinberg, 2007) discusses link prediction specifically applied to social networks, by applying a range of widely accepted methods to predict new academic collaboration data. This review was later augmented by Lü and Zhou (2011), expanding the algorithms tested and using alternative collaboration data. However, the motivations for academic collaboration may be different to the friendship selection methods of adolescents. Additionally, the analysis does not consider the impact of the links being formed and thus influencing the formation of other links in the network. The possibility to breaking links is also absent from the work of Liben-Nowell and Kleinberg (2007) and Lü and Zhou (2011), as the research was concerned only with the formation of new collaborations. Thus, the ability to implement link prediction methods in an ABS framework for adolescent social networks, provides a novel contribution to the literature.

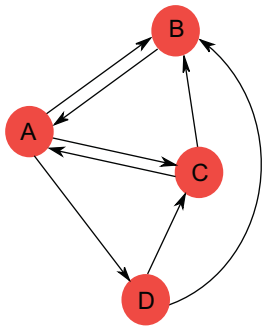

Fig. 3. Example network for illustration of link prediction algorithms.

Four prediction methods have been selected for the purpose of our research: Adamic/Adar (Section 4.1), Katz (Section 4.2), SAB Modelling (Section 4.3) and (PageRank (Section 4.4). These methods have been selected for comparison with a newly developed algorithm that we propose in Section 6, *PageRank-Max*. A summary of each method now follows aided by an example based upon the illustrative network shown in Fig. 3 (where appropriate).

4.1. Adamic/Adar

The Adamic/Ada (AA) method was originally developed to quantify how webpages were similar in terms of content, specifically focusing upon personal web pages; if the content between two pages is similar (Adamic & Adar, 2003) theorised that a connection between them is more likely to appear. The authors based their theory upon the notion that friends tend to be similar to one another (Carley, 1991; Feld, 1981), therefore making connections more probable.

To perform the AA method, the *neighbourhood*, $\Gamma(i)$, of each individual, i, is required; $\Gamma(i)$ being the set of individuals with whom i shares a connection. A score is calculated for each link (ij) that is not present (unobserved) in the network, such that:

$$Score[i, j] = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(z)|} \tag{5}$$

where z is a mutual connected vertex of both i and j.

The AA score for ij is therefore based upon the number of connections an individual z (who is a friend of both i and j) possesses. If z has a small number of connections, then having z as a common neighbour of both i and j is rarer than if z had a high number of connections. As such, rarer common neighbours increase Score[i,j] meaning that a link between i and j is more likely.

The following example illustrates the mechanism by which AA makes a link prediction:

- Taking the social network in Fig. 3, the unobserved links are identified as: $B \rightarrow C$, $B \rightarrow D$, $C \rightarrow D$ and $D \rightarrow A$.
- Taking the unobserved link $B \rightarrow C$, examining the friendships of B and C gives the neighbourhoods $\Gamma(B) = \{A\}$ and $\Gamma(C) = \{A, B\}$, respectively.
- As both $\Gamma(B)$ and $\Gamma(C)$ contain agent A, A is identifies as the only common neighbour of agents B and C.
- Agent A has three outward links, as such $|\Gamma(A)| = 3$ and therefore the Score [B,C] = 0.910 (3 d.p.).
- The scores for the remaining unobserved links ($B \rightarrow D$, $C \rightarrow D$ and $D \rightarrow A$) are also calculated. The resultant scores are ranked and the links with the highest scores are most likely to develop according to the AA link prediction method.

The example presented is conducted upon a directed network, however, the AA method does not consider the effect of reciprocation - a reciprocated tie being one in which the links $i \rightarrow j$ and $j \rightarrow i$

both exist, previously defined in Eq. (2). Returning to our example, the calculated Score[B, C] for the unobserved link $B \to C$ does not consider that the link $C \to B$ exists; this ignore the fact that agent B may wish to reciprocate the link with C, basing the strength of the "relation" purely upon the size of the neighbourhood of A.

4.2. Katz

Developed by Katz (1953) as a method to identify individuals of status within a group "free from the deficiencies of popularity contest procedures", the method examines not only the number of "popularity votes" an agent receives, but also the popularity of the voting individuals. As such, Katz argues that a more accurate perception of high status individuals in a group may be garnered. With respect to link prediction, the popularity votes referred to by Katz may be considered as connections in a network.

To perform the Katz method, the sociomatrix, X, of a network is required. It is well-known that the paths between individuals in a social network may be found by exploiting the powers of the relevant adjacency matrices (Festinger, 1949). For matrices with binary entries (such as X), non-zero elements x_{ij}^2 of the matrix X^2 indicate the number of paths of length two being present between agents i and j; similarly, a non-zero element x_{ij}^3 of the matrix X^3, indicates the number of paths of length three between agents i and j – higher powers having corresponding interpretations. In terms of link prediction, a score for an unrealised link between $i \to j$ is calculated as:

$$Score[i, j] = \sum_{l=1}^{n-1} \phi^l |path_{i,j}^{[l]}| \tag{6}$$

whereby $|path_{i,j}^{[l]}|$ represents the number of paths of length l between i and j, and ϕ is the selected *dampening* factor. The selection of ϕ must satisfy the condition $\phi < 1$, with $\frac{1}{\phi}$ being the smallest integer value greater than the largest eigenvalue of matrix X.

The Katz method, much like the AA method, assumes undirected network connections, with the underlying concept assuming that popular individuals are more likely to connect with one another – shortening the overall average shortest path length of the network. To illustrate the calculation of the Katz method, an example using the social network in Fig. 3 is as follows:

- For the calculation of the Katz method, the 4×4 sociomatrix X in Fig. 3 is required:

$$X = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

  Elements $x_{2,3}$, $x_{2,4}$, $x_{3,4}$ and $x_{4,1}$ are zero, indicating the potentially unobserved links.
- As the number of agents $n = 4$, the maximum path length for an indirect connection between agents is 3. Therefore the power of matrices to $n - 1$ are calculated:

$$X^2 = \begin{pmatrix} 2 & 2 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 2 & 1 & 0 & 0 \end{pmatrix}$$

$$X^3 = \begin{pmatrix} 3 & 3 & 2 & 2 \\ 2 & 2 & 1 & 0 \\ 2 & 3 & 2 & 1 \\ 1 & 2 & 2 & 2 \end{pmatrix}$$

- The value ϕ is selected by finding the maximal eigenvalue (λ) of X. As $\lambda = 1.950$ (3 d.p.), the value of $\frac{1}{\phi}$ is taken to be 2, allowing $\phi = 0.5$; this satisfies the requirements of $\phi < 1$ and $\frac{1}{\phi}$ being the smallest integer value greater than the characteristic root of X.
- Taking once again the unobserved link of $B \to C$, the Score[B,C] is calculated as:

$$Score[B, C] = (0.5)^1 \cdot 0 + (0.5)^2 \cdot 1 + (0.5)^3 \cdot 1 = 0.375$$

- The remaining unobserved link scores are calculated in the same manner and ranked accordingly. The links with the highest scores, are those which are most likely to occur at a subsequent timestep.

4.3. Stochastic Actor Based

The Stochastic Actor Based (SAB) modelling approach is not a static method such as those of AA and Katz. Rather, Snijders (1996) defines the SAB approach to be a class of models for longitudinal network data – 'actors' within the network utilising heuristics to optimise their individual goals, subject to a selection of constraints. Discrete observations of a network are explored, with the evolution of social ties from G_t to G_{t+1} a result of many small changes occurring between the specified time periods (Carrington, 2005) – the observed networks assumed to be the result of a Markov process in continuous time.

Consider T observations of a social network, represented as the adjacency matricies X_t for $t = 1, \ldots, T$, each observation containing the same set of n actors. Evolution of the network is solely modelled from the point of inception X_1, with the evolution to X_1 not being considered. The actions of actors within the network at t are simulated, changes in friendship ties based upon actor specific personal objective functions; the process attempting to model the micro-changes necessary to arrive at the network of $t + 1$. The complete SAB algorithm (Snijders, 1996) is rather detailed and complex in its implementation, so in the interests of space is not repeated here.

4.4. PageRank

The PageRank (PR) algorithm was developed by Brin and Page (1998), the founders of Googe. PR analyses the link structure of a network, taking into consideration not only the number of links to a node, but also the importance of the node sending the outward link. The PR (w_i) for each node i, is such that $w_i \geq 0$ and $w_j > w_k$ indicates j is a more important node than k. If $\bar{H}_i$ denotes the set of nodes that link to i, and H_i the set of nodes linked outwardly from i, then the PR w_i is calculated as:

$$w_i = \sum_{j \in \bar{H}_i} \frac{w_j}{|H_j|} \tag{7}$$

The calculation of w_i is recursive and can be initiated with any selected initial importance scores, iterating until convergence. The calculation of the PR may be interpreted as a random walk on a graph; in the context of the internet, a "random surfer" clicks on webpage links at random – the resultant probability of arriving at a page defined as its PR.

The "random surfer" calculation of PR is useful when importance scores are necessary for large graphs (such as the internet), whereby the adjacency matrix of connections X is unobtainable. However, if X is known, an adjusted matrix (M) may be calculated with $m_{ij} = \frac{1}{|H_j|}$ if the link $j \to i$ exists and $m_{ij} = 0$ otherwise. The PR calculation may then be expressed as a system of linear equations $Mw = w$, with the problem reduced to finding the principal eigenvector of the matrix M. Due to the properties of M, it is possible to find an eigenvalue $\lambda = 1$ which generates a *unique* positive eigenvector; this eigenvector being the vector of PageRanks (Page & Brin, 1999).

The matrix M is defined as column stochastic if each element $m_{ij} \geq 0$ and the sum of each column is 1, this ensures the existence of $\lambda = 1$. However, this does not guarantee the existence of a *unique* λ necessary for ranking, therefore other requirements of M need to be satisfied. From Perron–Frobenious theorem (Meyer, 2000), a column stochastic matrix M that is *irreducible* with $m_{ij} \geq 0$, generates:

- An eigenvalue $\lambda > 0$ with corresponding eigenvector $v > 0$.
- The existence of a dominant eigenvalue λ_1, such that $\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$.
- All eigenvectors ≥ 0 are a multiple of w.

Therefore, M also needs to satisfy the condition of irreducibility, whereby M cannot be placed into block-upper triangular form through a series of permutations. M may become reducible if disconnected clusters of nodes exist in the network. Furthermore, nodes with an inward link but no outward links, termed as "dangling nodes", also affect the necessary requirements for a unique vector of PageRanks (Ipsen & Selee, 2008).

To ensure the successful calculation of the PR vector, M is required to represent a *strongly connected* graph; a graph being strongly connected if a path from any given node i to j exists. Performing the PR calculation upon a strongly connected graph is not always possible, as is the case for both web pages and social networks. As such, calculation of a new matrix $\bar{M}$ is required:

$$\bar{M} = (1-d)Q + dM \tag{8}$$

where Q is the matrix of elements $\frac{1}{n}$ and d is the 'dampening factor', ensuring that $\bar{m}_{ij} \geq (1-d)Q$ which satisfies the required conditions; d is generally selected to be 0.85 (Bryan, 2006). The principal eigenvector of $\bar{M}$ is calculated, returning the required PR.

To illustrate PR, the following example is conducted upon the network in Fig. 3:

- The sociomatrix X of the network in Fig. 3 is :

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

  with the number of outward links for each agent: $|H_A| = 3$, $|H_B| = 1$, $|H_C| = 2$ and $|H_D| = 2$.
- The matrix M is calculated where $m_{ij} = \frac{1}{|H_j|}$ if the link $j \to i$ exists and $m_{ij} = 0$ otherwise, giving:

$$M = \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix}$$

- Taking $d = 0.85$ with $n = 4$, the $\bar{M}$ matrix is calulated as:

$$\bar{M} = 0.15 \cdot \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} + 0.85 \cdot \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix}$$

$$\bar{M} = \begin{pmatrix} \frac{3}{80} & \frac{71}{80} & \frac{37}{80} & \frac{3}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{37}{80} & \frac{37}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{3}{80} & \frac{37}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{3}{80} & \frac{3}{80} \end{pmatrix}$$

- The matrix $\bar{M}$ is in the form that allows for the calculation of the PR vector. The eigenvector of $\bar{M}$ corresponding to the dominant eigenvalue is found to be:

$$W = \begin{pmatrix} 0.36816 \\ 0.28796 \\ 0.20208 \\ 0.14181 \end{pmatrix}$$

- Hence, the PageRank of each node is found. As node A has the highest PageRank, it is therefore the most "important" node in the network.

5. Agent based simulation

This section describes the developed Agent Based Simulation (ABS), reported here using the revised Overview, Design concepts, Details (ODD) protocol for Agent Based Models (Grimm et al., 2010).

5.1. Purpose

The aim of the proposed simulation is to take the ASSIST data and simulate the evolution of the adolescent social networks over time, with an attempt to understand the process by which connections are modified.

5.2. Entities, state variables and scales

The simulation is Java based, which is an object-orientated programming (OOP) language. As such, the simulation is structured to have a 'Main' class, where the methods necessary for running the simulation are executed, and an 'Agent' class, whereby each instance of Agent represents an individual from the ASSIST data.

Each Agent object has a variable (an array list) relating to the individual's connections, with access to a global array (sociomatrix) containing the adjacency matrix of all links for the school being simulated. When an update occurs, the changing 'Agent' object (searching agent) updates its own link information variable and the global adjacency matrix.

Agents have other attributes, such as:

- Age – calculated from 'Date of Birth', and increases as time progresses.
- Sex – drawn directly from data.
- Smoking Level – the self reported smoking level classified on a scale from one to six, where a smoking value of one indicates 'never smoked' and six representing 'more that 6 cigarettes a week'. Again, drawn directly from data.

While these attributes are recorded in the simulation, the do not impact the progression of the model and have been included for ease in future work. Time is measured on a scale of weeks and there is no spacial representation at this stage, as only the change in connections is of importance.

5.3. Process, overview and scheduling

The following step-by-step guide describes how a link prediction is made in the ABS:

- On initialisation, a sociomatrix (X) and number of link changes (ϵ) are read from the database, giving a network rate of change $\rho = \frac{1}{\epsilon}$.
- At time t an event occurs, with the time between events being negatively exponentially distributed with parameter ρ.
- The event signifies that an agent must make a change to their outgoing links, the agent making the change being selected uniformly at random (termed as the 'searching agent').
- The randomly selected agent i (searching agent) receives a "message" telling them they must make a change, the change made being based upon the maximisation of i's personal objective function f_i.
- Agent i iterates through the link changes offered by the selected link prediction method (the 'testing agents') from amongst the 4 methods described in Section 4, finding their maximum f_i.
- Agent i makes one change to their outgoing links, updating X accordingly.
- The process repeats until stopping conditions are satisfied, subsequent agents making use of the updated links from previous agents to make their decisions.

The advancement of the simulation may therefore be interpreted as having a Discrete Event Simulation (DES) structure, as the system decides when events will occur and the selection of the agent who must make a change. An important deviation from the DES structure is that the changes made to the system are agent based decisions, the agents selecting the friendship option that most suits them (through their personal objective function). As a result, agent j must consider the changes made previously by agent i; this means agent j's decisions may be affected by those of i, potentially changing j's overall decision. Modelling friendship changes in the specified manner, means that individual decisions affect the system as a whole; individual connection decisions affecting future connections the network. The simulation may therefore be thought of as an ABS, with discrete event based timing ; a diagram of the logic is visible in Fig. 4.

5.4. Design concepts

The simulation makes use of six different approaches to link prediction:

- Random – the selecting agent randomly selects any other agent in the simulation for connection. If a connection already exists, the connection is broken. This method is included for the purposes of providing a baseline to the other link prediction methods evaluated.
- Adamic/Adar, Katz, SAB and PageRank – These methods have been described in Section 4, the associated algorithms implemented in the simulation.
- PageRank-Max – A novel approach to link prediction, based on the PageRank approach. A detailed discussion of this new algorithm may be found in Section 6.

5.5. Initialisation

On initialisation, the simulation is required to create multiple instances of the Agent class. The user must decide the school and timestep for prediction, the simulation accessing the ASSIST database and querying the relevant tables through the use of SQL. The sociomatrix X and number of changes ϵ are saved as global variables, with the number of Agent objects created based on the information within X. A separate data table is accessed, containing the properties of the individuals to be simulated (such as unique id); the information is then applied, giving each agent an identity.

Each agent then accesses the row in X that represents their connections, storing the agents to whom they send an outlink within a local variable. The network is then drawn for visualisation purposes, the graphics being able to update each time an agent makes a change. With the initialisation process complete, a representation of the school network (at the designated time) is present, the simulation being able to commence.

5.6. Input data

The ASSIST data provides multiple observations of a school social network, therefore, the predictions made may be assessed against real data at later time periods – gaining an insight into the accuracy of the predictions. Three waves of data are available (T_1, T_2 and T_3), as such, two predictions can be made – that of T_1 to T_2 and T_2 to T_3.

An Access database has been created for use with the simulation, holding information regarding friendship ties and basic student information. The database contains a separate table relating to the adjacency matrix of social ties, for each school at each time step; this allows individual schools to be modelled separately with ease.

5.7. Software

The software used for the ABS is Anylogic 6 (AnyLogic, 2002) given the ease in which it can connect to databases; this being a requirement when inputting the ASSIST data into the model to create the social network structures. Furthermore, AnyLogic offers the user the ability to expand its basic functionality with Java, which streamlines the coding of link prediction methods into the simulation. The source code is available at Fetta, Harper, Knight, and Williams (2017).

6. PageRank-Max

Given the potential importance of centrality in message diffusion within a social network, it stands to reason that centrality may also be of importance to the individuals comprising the social network. We propose a new link prediction algorithm, the PageRank-Max (PR-Max) method, which provides an individual perspective of centrality, a searching agent altering its connections based upon the personal optimisation of its own eigen-centrality.

The PR-Max method seeks to find the connection that may improve an agents own PR. On receipt of a message from the environment, the changing agent (i) begins iterating through all agents in the network as follows:

- Agent j is selected for testing.
- The connection from i to j is altered, either by forming a link or breaking an existing link.
- Agent i's PR is calculated and stored as $f_{i,j}$.
- The connection change is reversed.
- The process repeats.

Once all possible changes to i's connections are assessed, the greatest value of $f_{i,j}$ is selected – the associated connection change being made. The PR-Max method works much like the SAB method, testing the result of an actual change to the network; however, it does not require the creation of a model prior to use, as a transformation of the sociomatix is its sole requirement. The simplicity of the PR calculation means that PR-Max method also does not require two waves of network data, being able to predict

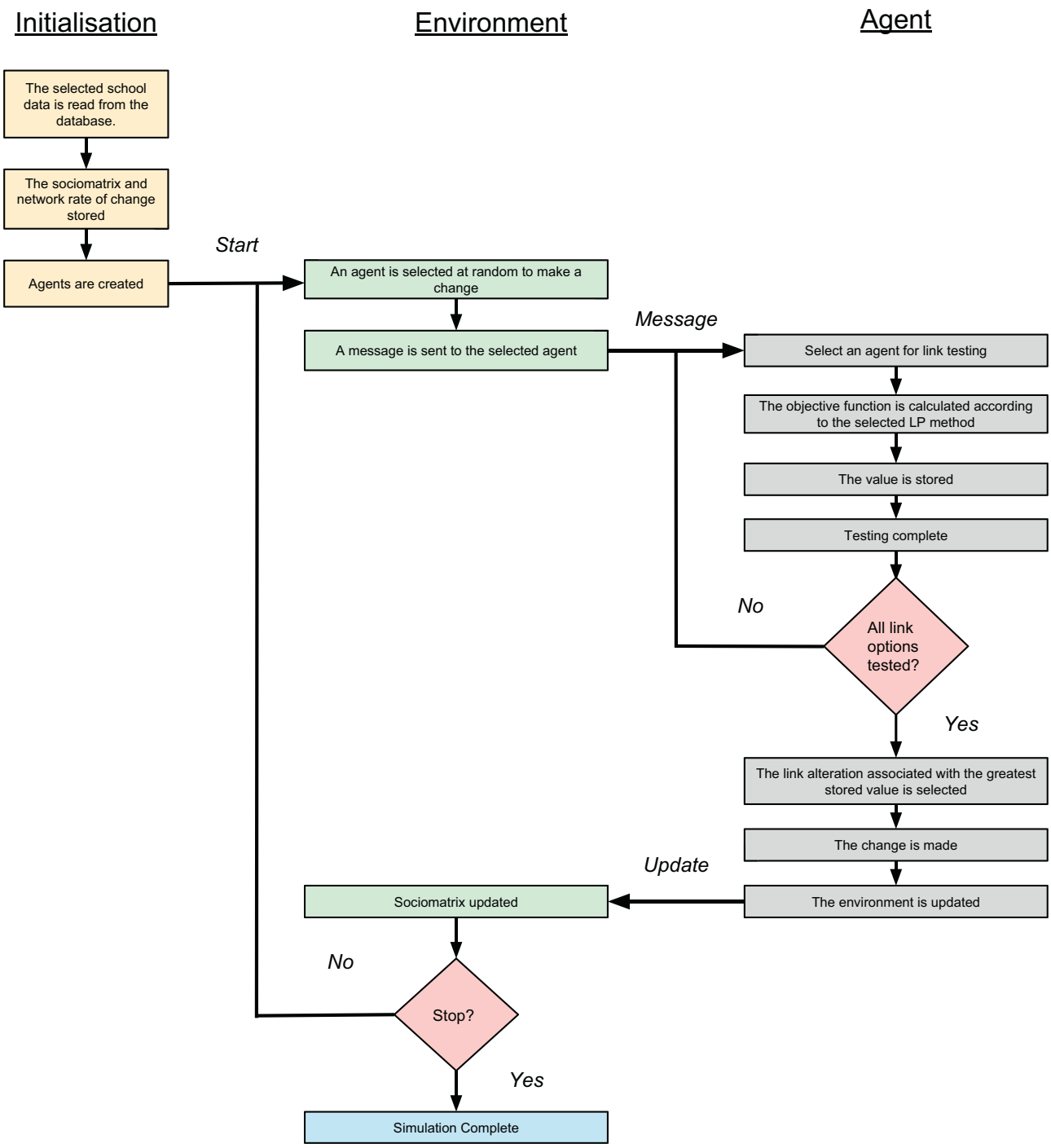

Fig. 4. Simulation logic describing the timing and agent-based decisions.

changes in the network without prior knowledge of its evolution; a diagram of the PR-Max logic is present in Fig. 5.

The PR of a webpage decides the ordering in which it is displayed on Google (following a search query). Users are said to be able to manipulate their webpage's PR by making educated link choices (Malaga, 2008), with the PR-Max method aiming to demonstrate this in the context of social relations. Researchers have attempted link prediction through the use of a 'Personalised PageRank' (Chen, 2012; Yung, 2012), which orders pages differently depending on what a specific user may find more relevant. In terms of link prediction, this means that the PR is calculated differently depending upon the specific searching agent seeking to make a new connection; this calculation process does not consider optimising an agent's own PR, which we consider in PR-Max.

While the careful selection of outward links is said to be important, removal of specific links has also been shown to have an effect on PR (de Kerchove, Ninove, & van Dooren, 2008); this gives the PR-Max method a sensitivity to link disconnection. The AA, Katz and basic PR implementations do not demonstrate such explicit consideration of link disconnection, their focus being predominantly upon the prediction of new connections. Although the SAB method does account for disconnection, this is subject to the model generated prior to simulation. Therefore, the PR-Max method may be able to capture elements of network evolution

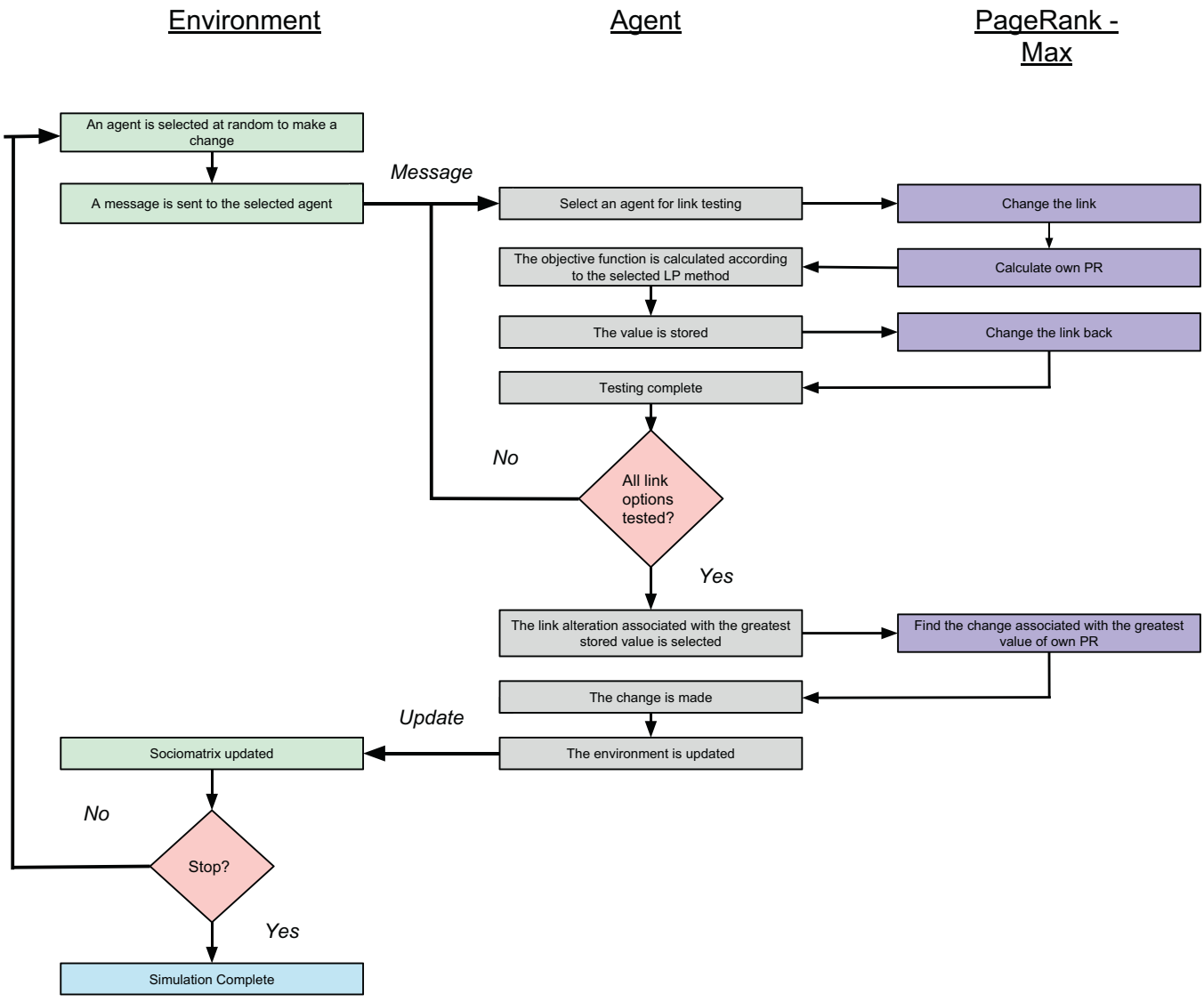

Fig. 5. Updated simulation logic describing the process of the PR-Max method.

more naturally. The performance of PR-Max is compared to the other link prediction methods in Section 8.

7. Validation

To gain confidence in the output of the simulation, validation and verification procedures have been conducted. As the simulation is attempting to validate social theories around how adolescents connect, the output of the simulation is in itself an evaluation of its validity. This is made evident by attempting to evaluate the accuracy of the link prediction approaches, against the empirical social network data – discussed further in Section 8. The following sections additonal elements of the validation process, prior to assessing the accuracy of the results: verification (Section 7.1), distributions and random sampling (Section 7.2), warm-up period (Section 7.3), number of runs (Section 7.4) and experimentation specification (Section 7.5).

7.1. Verification

Verification is described as a micro-check of the model, where a test of each individual element is performed. During the creation process, regular checks of the code were carried out - attempting to ensure the proper implementation of the designated logic. For each of the LP method implementations, the associated calculation of the objective function was performed to ensure calculations matched. Network visualisation was also used to verify consistency with the predictions made.

7.2. Distributions and random sampling

Statistical distributions and random sampling are used throughout the simulation, the values being derived from AnyLogic's own built in engine. Sampling of random numbers uses AnyLogic's default random number generator, which is an instance of the 'Random' Java class; this being a Linear Congruential Generator (AnyLogic, 2002). During the verification process, a number of runs were performed to assess the average number of changes in a selected school network; the confidence interval was calculated, and as the actual number of changes from the data fell within the bounds of the confidence interval, the distribution was said to be acting appropriately. During all testing and result generations common random numbers are implemented between scenarios.

Table 1
Model runtimes (minutes) by link prediction method.

	Random	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
Time (minutes)	5.1	6.7	10.9	13.8	15.4	492.3

7.3. Warm-up period

The starting conditions of the simulation (for a selected school at a given time point) are provided by the initial sociomatrix, which is read during the initialisation procedure. As such, a warm-up period is not required, as the agents begin with the required set up of connections.

7.4. Replications

As the simulation has various elements which include variability, multiple simulation runs are required. This work makes use of the confidence interval approach (Robinson, 2004) to select the number of replications, based on outcome-based precision criteria. Using the CI method, the required number of runs (η) is calculated as:

$$\eta = \left(\frac{100 \cdot S \cdot t_{(n-1,\alpha/2)}}{\widehat{d} \cdot \bar{x}} \right) \tag{9}$$

where $\bar{x}$ and S are the sample mean and standard deviation (respectively), $\widehat{d}$ the desired percentage deviation of confidence about the mean, and $t_{n-1,\alpha/2}$ from the standard t-distribution with $n-1$ degrees of freedom and significance level α (Robinson, 2004).

A selection if network measures were used to assess the variability in the outcome based metrics. With a significance level $\alpha = 0.05$, the greatest number of runs required was 9.49. Additionally, as a 'rule of thumb', (Law & Kelton, 1999) suggest a minimum of around 3–5 replications are required; should too many replications be selected, this wastes valuable running time and computing resources. Given that the identified maximum, 10 replications have been selected. This is greater than the rule of thumb, but does not appear excessive.

7.5. Experimentation specification

The investigation was conducted across eight intel i3 2120 dual core machines, with 8 gigabytes RAM. Each set up of the simulation was conducted on an individual machine, parallelised to make use of the dual cores. An example set up on a machine would be: School 12, PageRank-Max, $T_1 - T_2$. The approximate runtime for each link prediction method is shown in Table 1.

8. Results

The previous sections have described the creation of an ABS to predict social network evolution implementing five separate link prediction methods: Adamic/Adar (AA), Katz, Stochastic Table 2 Actor Based (SAB) Models, PageRank (PR) and PageRank-Max (PR-Max). This section discusses the results produced from evaluating each of these methods, across the breadth of the ASSIST network school data, for four different key network statistics: transivity, average degree, reciprocity and Average Path Length (APL).

For each of the control schools, a prediction is made from T_1 to T_2 and T_2 to T_3. The predicted networks at T_2 and T_3 shall be compared with the real data to evaluate their accuracy. The presentation of results is structured as follows: the *precision* of each algorithm in predicting the correct links is discussed in Section 8.1 and the individual *network structures* are presented in Section 8.2.

Table 3
Ranked average precision values.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Correct	3	4	2	5	1
	Missed	2	4	3	5	1
T_3	Correct	3	4	2	5	1
	Missed	3	4	2	5	1

8.1. Precision analysis

The first method to evaluate the T_2 and T_3 predictions is that of *precision*. The precision metric was first proposed by Cleverdon (1972) and has been used in the context link prediction methods (Lü & Zhou, 2011). Precision evaluates the number of *correct* predictions, y_c, relative to the number of predictions made, y_p, such that the precision is $\frac{y_c}{y_p}$. To benchmark performance, we also generate a network based upon link predictions made at random (the random method). The precision is expressed as a percentage improvement over predictions made at random; positive values indicate an improvement in correct predictions, while negative values indicate a reduction. Ten runs of the random method for each school network are performed to generate the random predictions.

Also of interest is the number of *missed* predictions, which examines the number of friendship changes not made in the predicted networks of T_2 and T_3, when a friendship change has actually occurred in the real data. The missed predictions are also expressed in terms of an increase compared to the random method, negative values indicating fewer predictions missed. Therefore, two metrics are calculated for each predicted network: the percentage increase of *correct* and *missed* link predictions over the random method.

Table 2 displays the average precision classified by method at each timestep. Each method is then ranked in terms of their precision performance; ranks are displayed in Table 3. Values that are significantly different from random at the 95% level, following an independent samples t-test for parametric data or a Mann–Whitney test for non-parametric data, are highlighted and starred.

The boxplots shown in Figs. 6 and 7 display the correct prediction scores at T_2 and T_3, respectively. They demonstrate the higher proportion of correct predictions for the PR-Max method when compared with all other selected methods. Overall, the precision

Table 2
Average of all school networks at T_2 and T_3, displaying the percentage increase over random predictions. Highlighted and starred values are significantly different at the 95% level.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Correct	7.69	5.16	10.03	0.63	40.35*
	Missed	-7.23	-3.44	-6.12	-0.54	−23.95*
T_3	Correct	5.96	5.23	10.16	0.65	42.87*
	Missed	-4.23	-3.40	-5.64	-0.53	−26.63*

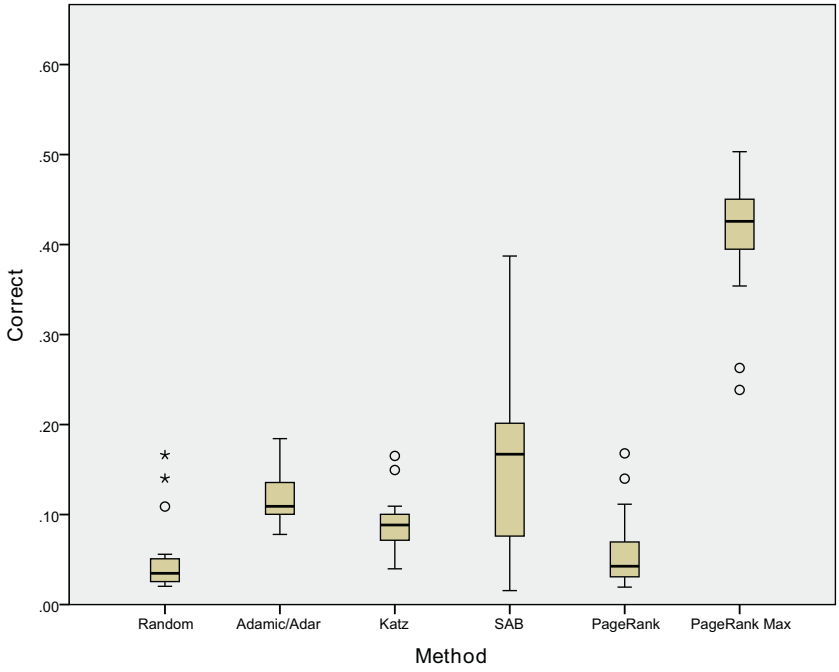

Fig. 6. Box plot of correct prediction proportions for each method at T_2. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers. Starred points indicate extreme outliers.

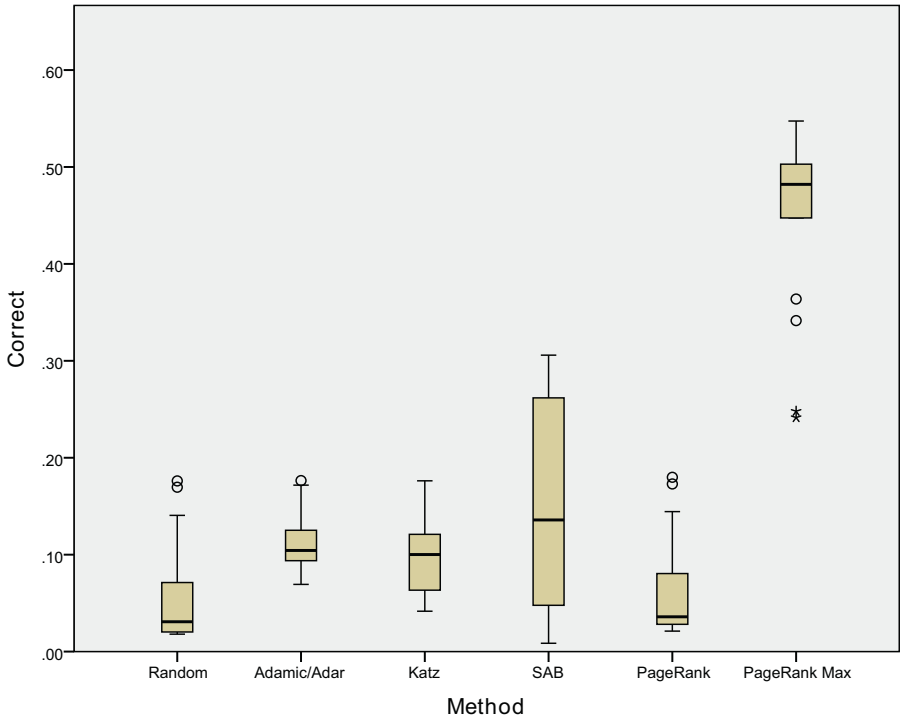

Fig. 7. Box plot of correct prediction proportions for each method at T_3. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers. Starred points indicate extreme outliers.

analysis has highlighted a number of key outcomes with regard to the link prediction methods tested, summarised as follows:

- PR-Max is the method which performs the best in terms of increasing correct predictions, and decreasing missed predictions.
- All methods experience variability in their performance, with certain methods capturing school-specific network evolution more accurately – potentially a result of the school's underlying friendship mechanisms.
- The PR-Max observes a significant increase in overall average precision at T_3 from T_2, adding further weight to the notion of time sensitivity in friendship evolution – the eigen-centrality of a student potentially becoming more important as they get older.

Table 4
AES for each link prediction method; highlighted values are significantly different at the 95% level.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Transitivity	11.80	2.65	30.43	53.29	2.32
	Average Degree	22.22*	13.04*	11.11	21.01	23.60*
	Reciprocity	7.28	16.11	12.27	33.56	1.59
	Average Path Length	13.20*	3.68*	6.33*	4.64*	29.45*
T_3	Transitivity	12.67	4.62	36.68	54.37	4.34
	Average Degree	29.25*	20.01*	15.94	23.73	13.02*
	Reciprocity	7.75	19.85	16.20	32.85	1.46
	Average Path Length	139.01*	99.05*	26.30*	24.74*	12.78*

Table 5
AES ranks for each link predicition method.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Transitivity	3	2	4	5	1
	Average out-degree	4	2	1	3	5
	Reciprocity	2	4	3	5	1
	Average Path Length	4	1	3	2	5
T_3	Transitivity	3	2	4	5	1
	Average out-degree	5	3	2	4	1
	Reciprocity	2	4	3	5	1
	Average Path Length	5	4	3	2	1

8.2. Network structure analysis

To analyse the predicted network structures, the output of the 10 simulation runs (for each school, at each timestep, for every link prediction method) are compared with the structural values from the data. The metrics selected for the analysis (transitivity, average degree, reciprocity, APL) are not on the same scale as each other; as such, meaningful comparisons between metrics is not intuitive. To rectify this issue, a new approach to network comparison is proposed making use of 'effect size'.

The effect size is a measure that represents the magnitude of a relationship, quantifying the difference between two groups; it is the central component of a meta-analysis, which attempts to summarise the finding of multiple investigations (Hedges & Olkin, 1985). The effect size used for this analysis is Glass' Δ, calculated as:

$$\Delta = \frac{\bar{x_d} - \bar{x_p}}{s_p} \tag{10}$$

where $\bar{x_d}$ and $\bar{x_p}$ are the mean values of a metric from the data and predicted networks respectively, and s_p is the associated predicted network standard deviation. $\bar{x_p}$ and s_p are calculated from the 10 simulation runs, while $\bar{x_d}$ is taken directly from the data. To evaluate each method's performance, a rank for each method is produced. This is calculated by taking the average absolute effect size (AES) across schools, for each structural metric. This analysis is only concerned with the magnitude of effect size, the directionality (overestimation or underestimation) being irrelevant; as such, the absolute effect size is taken in the calculation of AES.

Table 4 displays the AES for each method and measure. To compare AES differences between time steps, paired sample t-tests (parametric) or paired sample Wilcoxon signed-rank (non-parametric) tests are performed – the values significantly different at the 95% level are highlighted and starred. Each method is then ranked by structural measure, values with the lowest AES achieving the highest ranks –Table 5.

The differences in AES between time steps is apparent from Table 4. The APL is predicted significantly differently across all methods, with predictions being worse for T_3 in AA (139.01), Katz (99.05), SAB (26.30) and PR (24.74) methods than T_2; however, AES is reduced for PR-Max at T_3 (12.78), this indicating a significant improvement in predictions. AES for average degree is also significantly different between T_2 and T_3, with AA (29.25) and Katz (20.01) increasing; once again, PR-Max values improve at T_3, with the AES value decreasing significantly.

The AES values indicate an improvement in the PR-Max structural accuracy at T_3. This is further reinforced by the ranks of Table 5, which demonstrate a movement of out-degree and APL predictions from last place (5) at T_2, to first place at T_3 (1). When the harmonic mean of the individual rankings is taken for each method, PR-Max is placed first across both time steps (T_2: 1.7, T_3: 1.0), however, at T_2 this is very closely followed by the Katz method (1.8).

The precision analysis of Section 8.1, placed the Katz method as fourth overall at both T_2 and T_3. However, it would appear that the method performs well in terms of structure at T_2, ranking first in APL AES and second for transitivity and average out-degree. This suggests that, while the specific links in the predicted networks may not be accurate, the overall network structure generated is more representative than other link prediction methods - only being outperformed by PR-Max in terms of transitivity and reciprocity. The findings demonstrate the importance of considering the predicted network structure when discussing link prediction methods, potentially providing further insight than simply considering precision.

Overall, the method structural performance analysis has reinforced many of the conclusions from Section 8.1. There would appear to be differences in the performance of methods at T_2 and T_3, suggesting an underlying change in the friendship mechanisms of adolescents within the ASSIST data. Further evidence of the strength of the PR-Max method (in predicting network evolution) is also provided, the method performing particularly well at T_3.

9. Conclusions

9.1. Using simulation as a tool to explore theories around behaviours

This paper has outlined the development of a simulation based framework, incorporating link prediction algorithms, for applica-

tion upon adolescent social network data. The simulation employed four existing link prediction methods: Adamic/Adar, Katz, SAB models and PageRank, and developed a new method PR-Max based upon the optimisation of an agent's eigen-centrality.

The existing methods selected were chosen due to their success in a wealth of prior applications, with the PR-Max method being developed to provide an alternative perspective of status. A limitation of the study may be the selection of only five methods to explore in depth. However, given the rigorous selection process of the chosen it was felt that an appropriate representation of the most widely used methods was presented.

The social network analysis offers novel contributions to both link prediction and simulation literature. Although the SAB method uses simulation as an underlying tool for the generation of statistical models, this work is seemingly the first study to structure the link prediction problem within an ABS framework. The development of the PR-Max method also provides a new approach to link prediction, whereby agents use eigen-centrality to actively improve their current social situation. Furthermore, this investigation expands the current literature relating to social applications of simulation, signalling a potential future direction for ABS research.

9.2. Pagerank Max is an effective predictor of future social structure, which suggests status is important in friendship selection

This analysis has concluded that the proposed PR-Max method was the most successful (of those tested) in predicting the evolution of adolescent friendships, in terms of both precision and network structure.

The PR-Max method highlighted that status may be a key factor in the evolution of adolescent social networks, especially as the individuals mature. This identifies status (an interpretation of eigen-centrality) as a key focus for future investigations of adolescent social networks. This suggests that a salient part of the adolescent friendship making process is the befriending of an individual who will likely increase ones own status. This contributes to literature describing adolescent social connection and may impact future adolescent peer diffusion studies.

A further relevant feature of the PR-Max is the process by which links were broken, with agent's removing connections that negatively impacted upon their eigen centrality. This suggests friendship degradation is an important factor in social connections, with adolescent social networks continually evolving over time. The results demonstrated the abilities of a simulation based link prediction structure in gaining insights unobtainable by conventional social network analysis.

9.3. Implications for policy makers and public health managers

Smoking is a major global health challenge, with 6 million deaths from tobacco use worldwide per year. Secondary schools are the common point at which people start smoking, so it is vital to intervene at this age given the addictive nature of tobacco and the longer-term health effects. Our conceptual approach and contribution to the problem is in providing a proof-of-concept for targeted interventions driven by social network analysis. We demonstrate the utility of using emerging sources of social network data for public heath interventions.

The ASSIST programme was shown to provide a cost-effective method for reducing adolescent smoking rates. The ASSIST programme resulted in a 2.1% reduction in smoking prevalence at 2 years, and the incremental cost per student not smoking was 1500. The intervention also affected students beliefs about longer term smoking behaviour, with a lower proportion of students in the intervention schools believing that they would be a smoker at age 16 years. The ASSIST findings, if extrapolated to all 12-year-old students in the UK, would cost £38m but would result in 20,400 fewer adolescent smokers at age 14 years. Placing these results in a broader context, NHS expenditure on treating lung cancer in 2010 was £261 million in England alone (Hollingworth et al., 2012).

The research has demonstrated, through the use of ABS and link prediction methods, the potential importance of eigen-centrality in adolescent friendships selection. As such, these learnings may be fed back into future peer-led interventions to aid in the selection of appropriate peer supporters. Furthermore, it provides encouraging results in demonstrating the ability to predict forward and identify highly connected individuals in a social network, informing policy-makers who to target to diffuse positive public health messages.

Acknowledgements

The authors gratefully acknowledge funding for this research from EPSRC as part of the LANCS Initiative [EP/F033613/1]. We would also like to take this opportunity to express our thanks to Professor Lawrence Moore and Dr. Jo Holliday (Centre for the Development and Evaluation of Complex Interventions for Public Health Improvement, DECIPHer), for granting access to the relevant ASSIST data and for their general assistance with queries relating to the ASSIST programme. Furthermore, we would like to thank the anonymous reviewers for their helpful remarks to improve the paper, and to Professor Brian Denton, University of Michigan, for the interest he has shown in this research and for his helpful suggestions.

References

Action on Smoking and Health (ASH) (2016). Smoking statistics, http://www.ash.org.uk/files/documents/ASH_93.pdf. Last accessed: November 2016.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks, 25*(3), 211–230. doi:10.1016/S0378-8733(03)00009-1.

An, L. (2012). Modeling human decisions in coupled human and natural systems: Review of agent-based models. *Ecological Modelling, 229*, 25–36. doi:10.1016/j.ecolmodel.2011.07.010.

AnyLogic (2002). http://www.xjtek.com/, Last Accessed: Jan 2014.

Audrey, S., Cordall, K., Moore, L., Cohen, D., & Campbell, R. (2004). The development and implementation of a peer-led intervention to prevent smoking among secondary school students using their established social networks. *Health Education Journal, 63*(3), 266–284. doi:10.1177/001789690406300307.

Barabási, A. L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: The topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications, 281*(1–4), 69–77. doi:10.1016/S0378-4371(00)00018-2.

Bernstein, G., & O'Brien, K. (2013). Stochastic agent-based simulations of social networks. In *Proceedings of the 46th annual simulation symposium*. Article no: 5

Bollobas, B. (2013). *Modern graph theory*. London: Springer-Verlag.

Brailsford, S., Harper, P. R., Patel, B., & Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation, 3*(3), 130–140.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1), 107–117.

Bryan, K. (2006). The $25,000,000,000 eigenvector: The linear algebra behind Google. *SIAM Review, 48*(3), 569–581.

Campbell, R., Starkey, F., Holliday, J., Audrey, S., Bloor, M., Parry-Langdon, N., et al. (2008). An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): A cluster randomised trial. *Lancet, 371*(9624), 1595–1602. doi:10.1016/S0140-6736(08)60692-3.

Carley, K. (1991). A theory of group stability. *American Sociological Review, 56*(3), 331–354.

Carrington, P. (2005). *Models and methods in social network analysis*. New York: Cambridge University Press.

Centers for Disease Control and Prevention (2014). 2014 Surgeon general's report—The health consequences of smoking—50 Years of progress. http://www.cdc.gov/tobacco/data_statistics/sgr/50th-anniversary/index.htm. Last accessed: November 2016.

Centers for Disease Control and Prevention (2016). Quitting smoking. www.cdc.gov/tobacco/data_statistics/fact_sheets/cessation/quitting/. Last accessed: November 2016.

Chen, E. (2012). Edge prediction in a social graph: My solution to Facebook's user recommendation contest on Kaggle, http://blog.echen.me/2012/07/31/edge-prediction-in-a-social-graph-my-solution-to-facebooks-user-recommendation-contest-on-kaggle, Last Accessed: Jan 2014.

Cleverdon, C. W. (1972). On the inverse relationship of recall and precision. *Journal of Documentation, 28*(3), 195–201. doi:10.1108/eb026538.

Domingos, P., & Richardson, M. (2007). Markov logic: A unifying framework for statistical relational learning. In *Proceedings of the ICML-2004 workshop on statistical relational learning and its connections to other fields.: 339* (pp. 49–54).

Farrell, S., Campbell, C., & Myagmar, S. (2005). Relescope: An experiment in accelerating relationships. In *Proceedings of conference on human factors in computing systems* (pp. 2–7).

Feld, S. L. (1981). The focused organization of social ties. *American Journal of Sociology, 86*(5), 1015–1035.

Festinger, L. (1949). The analysis of sociograms using matrix algebra. *Human Relations, 2*, 153–158. doi:10.1177/001872674900200205.

Festinger, L., Back, K. W., & Schachter, S. (1950). *Social pressures in informal groups: A study of human factors in housing*: 3. Stanford University Press.

Fetta, A., Harper, P., Knight, V., & Williams, J. (2017). Simulation software for predicting adolescent social networks to stop smoking in secondary schools, 10.5281/zenodo.823817.

Franco, L., & Hämäläinen, R. (2017). Engaging with behavioral operational research: On methods, actors and praxis. *Behavioral operational research: Theory, methodology and practice* (pp. 3–25.). Palgrave Macmillan.

Fronczak, A., Fronczak, P., & Holyst, J. (2004). Average path length in random networks. *Physical Review E, 70*(5), 056110. doi:10.1103/PhysRevE.70.056110.

Getoor, L. (2003). Link mining: A new data mining challenge. *ACM SIGKDD Explorations Newsletter, 5*(1), 84–89.

Getoor, L., & Diehl, C. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter, 7*(2), 3–12.

Goldenberg, A., Kubica, J., & Komarek, P. (2003). A comparison of statistical and machine learning algorithms on the task of link completion. In *Proceedings of KDD workshop on link analysis for detecting complex behavior* (p. 8).

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The odd protocol: A review and first update. *Ecological Modelling, 221*(23), 2760–2768. http://dx.doi.org/10.1016/j.ecolmodel.2010.08.019.

Hämäläinen, R. P., Luoma, J., & Saarinen, E. (2013). On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. *European Journal of Operational Research, 228*(3), 623–634. http://dx.doi.org/10.1016/j.ejor.2013.02.001.

Han, X., Zhao, Z., Hadzibeganovic, T., & Wang, B. (2014). Epidemic spreading on hierarchical geographical networks with mobile agents. *Communications in Nonlinear Science and Numerical Simulation, 19*(5), 1301–1312. doi:10.1016/j.cnsns.2013.09.002.

Harary, F. (1994). *Graph theory*. Boulder: Westview Press.

Hasan, M. A., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. In *Proceedings of of SDM 06 workshop on link analysis, counterterrorism and security*.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Holliday, J. (2006). *Identifying and using influential young people for informal peer-led health promotion*. Cardiff University, Ph.D. thesis.

Hollingworth, W., Cohen, D., Hawkins, J., Hughes, R. A., Moore, L. A. R., Holliday, J. C., et al. (2012). Reducing smoking in adolescents: Cost-effectiveness results from the cluster randomized ASSIST (A stop smoking in schools trial). *Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco, 14*(2), 161–168. doi:10.1093/ntr/ntr155.

Holme, P. (2005). Network reachability of real-world contact sequences. *Physical Review E, 71*(4), 046119. doi:10.1103/PhysRevE.71.046119.

Huang, Z., Li, X., & Chen, H. (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries* (pp. 141–142). New York. doi:10.1145/1065385.1065415.

Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W., & Bakker, P. J. (2012). Taxonomic classification of planning decisions in health care: A structured review of the state of the art in or/ms. *Health Systems, 1*(2), 129.

Ipsen, I. C. F., & Selee, T. M. (2008). PageRank computation, with special attention to dangling nodes. *SIAM Journal on Matrix Analysis and Applications, 29*(4), 1281–1296. doi:10.1137/060664331.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.

Kelly, J. A., St Lawrence, J. S., Stevenson, L. Y., Hauth, a. C., Kalichman, S. C., Diaz, Y. E., et al. (1992). Community AIDS/HIV risk reduction: The effects of endorsements by popular people in three cities. *American Journal of Public Health, 82*(11), 1483–1489.

de Kerchove, C., Ninove, L., & van Dooren, P. (2008). Maximizing PageRank via outlinks. *Linear Algebra and its Applications, 429*(5), 1254–1276. doi:10.1016/j.laa.2008.01.023.

Kirke, D. M. (1996). Collecting peer data and delineating peer networks in a complete network. *Social Networks, 18*(4), 333–346. http://dx.doi.org/10.1016/0378-8733(95)00280-4.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600). ACM.

Law, A., & Kelton, W. D. (1999). Simulation modeling and analysis. *Industrial engineering and management science series*. McGraw-Hill Science/Engineering/Math.

Liben-Nowell, D., & Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications, 390*(6), 1150–1170. doi:10.1016/j.physa.2010.11.027.

Luce, R. D., & Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika, 14*(1), 95–116.

Malaga, R. (2008). Worst practices in search engine optimization. *Communications of the ACM, 51*(12), 147–150.

Mao, L. (2014). Modeling triple-diffusions of infectious diseases, information, and preventive behaviors through a metropolitan social network – An agent-based simulation. *Applied Geography, 50*, 31–39. doi:10.1016/j.apgeog.2014.02.005.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444.

Mei, S., Sloot, P., Quax, R., Zhu, Y., & Wang, W. (2010). Complex agent networks explaining the HIV epidemic among homosexual men in Amsterdam. *Mathematics and Computers in Simulation, 80*(5), 1018–1030. doi:10.1016/j.matcom.2009.12.008.

Meyer, C. (2000). *Matrix analysis and applied linear algebra*. SIAM. ISBN: 0-89871-454-0.

Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2008). Growth of the flickr social network. In *Proceedings of the first workshop on online social networks* (pp. 25–30). New York, New York, USA: ACM Press. doi:10.1145/1397735.1397742.

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*(1), 016132. doi:10.1103/PhysRevE.64.016132.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167–256.

Newman, M. E. J., Forrest, S., & Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E, 66*(3), 035101. doi:10.1103/PhysRevE.66.035101.

Niazi, M., & Hussain, A. (2011). Agent-based computing from multi-agent systems to agent-based models: a visual survey. *Scientometrics, 89*(2), 479–499. doi:10.1007/s11192-011-0468-9.

Page, L., & Brin, S. (1999). The PageRank citation ranking: Bringing order to the web. *Technical Report, Stanford InfoLab*.

Parker, J. G., & Seal, J. (1996). Forming, losing, renewing, and replacing friendships: Applying temporal parameters to the assessment of children's friendship experiences. *Child Development, 67*(5), 2248–2268. doi:10.1111/j.1467-8624.1996.tb01855.x.

Pearson, M., & Michell, L. (2000). Smoke rings: Social network analysis of friendship groups, smoking and drug-taking. *Drugs-education Prevention and Policy, 7*(1), 21–37.

Pidd, M. (2004). *Computer simulation in management science*. Chichester: John Wiley & Sons.

Popescul, A., & Ungar, L. (2003). Statistical relational learning for link prediction. In *Proceedings of the workshop on learning statistical models from relational data*.

Pujol, J., Sanguesa, R., & Delgado, J. (2002). Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the first international joint conference on autonomous agents and multiagent systems: Part 1* (pp. 467–474).

Robinson, S. (2004). *Simulation: The practice of model development and use*. Chichester: John Wiley & Sons.

Salter-Townshend, M. (2012). Analysing my Facebook friends. *Significance, 9*(4), 40–42.

Schank, T., & Wagner, D. (2005). Approximating clustering coefficient and transitivity basic definitions. *Journal of Graph Algorithms and Applications, 9*(2), 265–275.

Snijders, T. A. B. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology, 21*(1–2), 149–172.

Soffer, S., & Vázquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E, 71*(5), 057101. doi:10.1103/PhysRevE.71.057101.

Taskar, B., Wong, M.-F., Abbeel, P., & Koller, D. (2004). Link prediction in relational data. *Advances in neural information processing systems* (pp. 659–666).

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442. doi:10.1038/30918.

World Health Organisation (2015). Tobacco facts, http://www.who.int/mediacentre/factsheets/fs339/en/, Last accessed: September 2016.

World Health Organisation (2016). http://www.who.int/tobacco/about/vision/en/.

Yung, D. (2012). Personalized Pagerank for link prediction. http://shom83.blogspot.co.uk/, Last Accessed: May 2014.

Zhu, J., Hong, J., & Hughes, J. G. (2004). PageCluster: Mining conceptual link hierarchies from Web log files for adaptive Web site navigation. *ACM Transactions on Internet Technology (TOIT), 4*(2), 185–208.