Social networks and health

Edward Hill, Frances Griffiths[a], Thomas House[b,c]

[a] *Warwick Medical School, University of Warwick, Coventry, UK*
[b] *Centre for Complexity Science, University of Warwick, Coventry, UK*
[c] *Mathematics Institute, University of Warwick, Coventry, UK*

Abstract

Human populations can be conceptualised as being arranged in social networks that influence the spread of diseases and behaviours. A common worldwide illness that may possess the ability to spread between social contacts is depression. Using data from the National Longitudinal Study of Adolescent Health, we explore the impact an adolescent school friendship network has on the emotional state of the individuals contained within it, in particular whether they report depressive symptoms or not. Our analysis approaches this problem in two distinct ways. First, for an individual in the social network, we find whether their friends emotional states and a number of their network properties are significant in attempting to predict their current emotional state. Second, assuming you know the current emotional state of an individual, how their friends emotional states can be used to predict if they are at risk of changing emotional state within a year. Our results show when attempting to predict the current emotional state of an individual, their out-degree has more significance than their in-degree, and their number of not depressed friends has more significance than their number of depressed friends. We show how the probability of an individual undergoing a change in emotional state in the future is influenced by the number of not depressed friends they have, while the number of depressed friends has no causal effect.

1. Introduction

The interaction of individuals with their social networks is thought to influence their human behaviour. Two main aspects of social networks have been studied extensively. The first aspect is their structure. The generative processes that give rise to widespread patterns in different types of social networks have been explored, including friendship networks [1] and sexual networks [2]. The second aspect is the affect of social networks on social dynamics. In recent times, social networks have been studied to determine the patterns of infectious disease spread [3]. Other studies explore whether certain behaviours spread from person to person, such as smoking [4] and whether the weight gain of one person is associated with weight gain in their social contacts [5, 6].

A common worldwide illness that might spread between social contacts is depression[7, 8]. The World Health Organisation estimates that, globally, there are currently more than 350 million people affected by depression [9]. At its worst, depression can lead to suicide. Suicide results in an estimated $1,000,000$ deaths worldwide every year [9]. The Office of National Statistics Child and Adolescent Mental Health Survey in 2004 found 1.4% of aged 11-16 year-olds (approximately 62,000 people) in the UK were seriously depressed [10]. In the United States, the Substance Abuse and Mental Health Services Administration examined the national prevalence of depression each year through the National Survey on Drug Use and Health (NSDUH). In the years 2004-2008, NSDUH found the prevalence of depression among the 12 to 17 years old age group in the United States in the years to be 7.9-9.0% [11]. There is evidence that social support is important for the mental well-being of adolescents [12] and evidence suggestive that befriending can have a positive effect on mental health [13]. Understanding the impact of social networks on depression would allow the development of novel interventions. These may be targeted at specific groups such as school attenders, or

more generally in society or through social media. In this study, when we refer to the "emotional state" of an individual we mean whether they are either not depressed or have depressive symptoms.

A recent area of interest has been modelling the spread of positive and negative emotions using theoretical tools from epidemiology. This follows models from epidemiology being used to explore the spread of other non-microbial infections, such as computer viruses [14]. The Framingham Heart Study dataset has been used to evaluate the spread of long-term emotional states across a social network. Hill et al. [15] use a form of the Susceptible-Infected-Susceptible disease model, with spontaneous infection included, and provide formal evidence that positive and negative emotional states behave like infectious diseases spreading over social networks over long periods of time.

In this study we investigate the spread of emotional states in a social network of adolescents that participated in the National Longitudinal Study of Adolescent Health (Add Health). We initially describe the observed depressive state behaviour in the two waves by considering two general questions: what network node properties, in particular the out-degree and in-degree, are significantly different when comparing respondents who are depressed and not depressed? Which of these network properties is the most significant if attempting to predict the current emotional state of an individual, and is this influenced by the emotional state of their school friends? We use statistical tests and generalised linear models on the dataset to address these problems. We go on to explore the level of causal effect, if any, certain variables have on emotional state by posing the following question: If you know the current emotional state of an individual, can you predict if they are at risk of changing emotional state (in the near future) based on the number of school friends they say they have, and is this influenced by the emotional state of their school friends? This is analysed using generalised linear models and Bayesian inference.

2. Methods

2.1. Add Health

The data used was drawn from the wave I and II in-home survey components of Add Health, a longitudinal study that explores health-related behaviour of a nationally representative sample of United States adolescents in grades 7 through 12[1] [16]. Wave I of the survey took place during the 1994-1995 academic year, with wave II taking place in 1996. These adolescents were surveyed from 132 schools, of which 80 were high schools, that were selected to ensure representation with respect to region of country, urbanicity, school size and type, and ethnicity. The 80 participating high schools helped to identify feeder schools that included a 7th grade and sent at least five graduates to that high school. From among the feeder schools, one was selected with a probability proportional to the number of students it contributed to the high school. The selected feeder school is referred to as the sister school. If the high school contained all grades from 7 to 12, then it did not need to be allocated a sister school.

The in-home survey contained over 2000 questions that covered many aspects of adolescent behaviours and attitudes. It was designed to be the largest, most comprehensive survey of adolescents ever undertaken to study their health-related behaviours. The wave I sample contained 20745 individuals and the wave II sample contained 14738 individuals. The wave II in-home interview sample was the same as the wave I in-home interview sample, the exception being the majority of 12th-grade respondents from wave I being removed from the wave II sample, as they exceeded the grade eligibility requirement in 1996. This was a major cause of the discrepancy in respondent numbers between the two waves. As with any survey, respondents may have answered questions dishonestly or with error. However, survey administrators took a number of steps to ensure data security and to minimize the potential for interviewer or parental influence. For example, respondents were not provided with printed questionnaires. Instead, all responses were recorded on laptop computers.

To enable analysis of social networks, all enrolled students in 16 schools were selected for in-home interviews. These schools were referred to as saturated schools. The following number of students from saturated schools completed an in-home survey; 3702 in wave I, 2777 in wave II. With each student completing the

[1] Students in the wave I and II sample were aged 12-19.

in-home survey, and details of their friendship networks, health and health behaviours being collected, this provided a "complete" social network for the saturated schools.

2.2. Friendship data

For both waves I and II, a section of the in-home questionnaire asked respondents to identify close friends. In wave I, 7106 respondents were asked to nominate up to 5 male and 5 female friends, including 3099 respondents who attended a saturated school (out of a possible 3702). The remaining 13640 respondents were asked to nominate one male and one female friend. In wave II, only those respondents who attended a saturated school were asked to identify as many as five male and five female friends. This group consisted of 2729 individuals. The remaining 12009 wave II respondents, who attended schools from which only a sample were selected for in-home interview, were asked to identify only one male and one female friend[2]. Each respondent was given a roster of names for their own school and their sister school. If a name was chosen from either roster, the corresponding AID number (i.e. identification number) for the nominated individual was recorded. Otherwise, there were three types of generic AIDs assigned to nominated friends that were not on the two rosters. If the friend attended the respondent's school, but the respondent could not find their name on the school roster, he or she was assigned the generic AID of 99999999. If the friend attended the respondent's sister school, but the respondent could not find their name on the school roster, he or she was assigned the generic AID of 88888888. If the friend did not attend either the respondent's school or the sister school, he or she was assigned the generic AID of 77777777.

2.3. CES-D scale

One measure of depression is a self-assessment based upon the Centre for Epidemiologic Studies Depression Scale (CES-D), developed by Radloff [17]. This is a 20-item measure that asks the respondent to rate how often over the past week they experienced symptoms associated with depression, such as feeling lonely or being too tired to do things. Response options ranged from 0 to 3 for each item (0 = rarely or none of the time, 1 = some or little of the time, 2 = moderately or much of the time, 3 = most or almost all the time)[3]. The Add Health survey includes 18 of the 20 questions that constitute the CES-D scale. The coded responses were summed to generate a score between 0 and 54. For the original 20 item CES-D scale, Roberts et al. [18] used a Receiver Operating Characteristic (ROC) analysis, a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied, to show a threshold of 24 for females and 22 for males provided the best agreement with clinical assessments for depression. Consequently, for this study using the rescaled CES-D involving 18 items, thresholds of 22 for females and 20 for males were used to create a binary indicator of emotional state (depression status), where respondents had a classification of being "not depressive" or having "depressive symptoms". It should be noted that the measures of depression used in this study do not come without limitations. Most importantly, these variables indicate depressive symptoms and do not represent medical diagnoses. In addition, while the CES-D asks the respondent to evaluate feelings in the last week, it is known to be a reliable measure of long-term emotions as opposed to short-lived moods. Studies have shown the CES-D scale scores to be stable for up to 12 months [17, 19].

2.4. Our sample datasets

For a respondent to be included in our study, they had to be from a saturated school, allowed to list up to 5 male and 5 female friends and provide answers to all the CES-D related questions in at least one of the wave I or II in-home interview surveys. Using students from saturated schools avoided results being influenced by a potentially biased method of choosing students from a school in which only a sample of students was taken. Students who were restricted to giving only one male and one female friend were not

[2]Note, 48 respondents in this group attended a saturated school, but were only told to give one male and one female friend.

[3]Four items assessed positive symptoms and were reversed before calculating the scores. These positive symptoms include how often the respondents (i) felt "happy", (ii) felt "that you were just as good as other people", (iii) felt "hopeful about the future", and (iv) "enjoyed life".

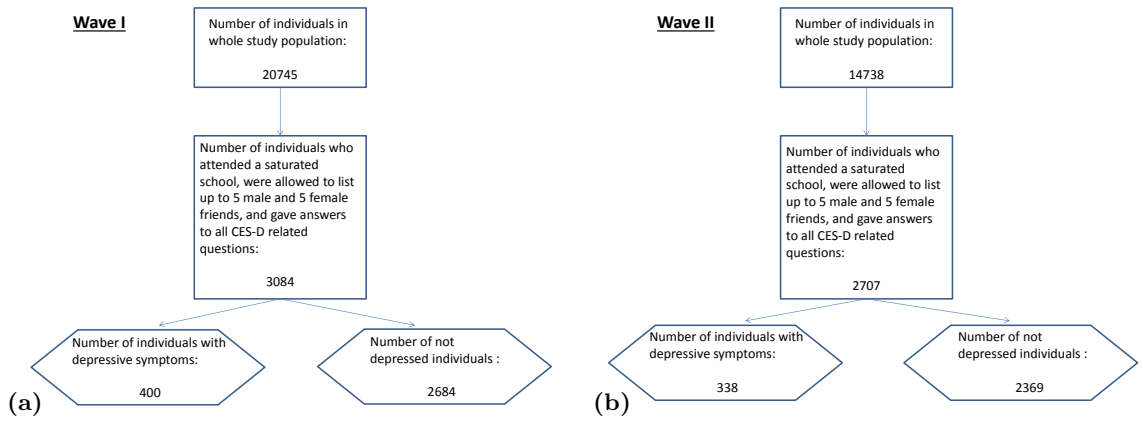

Fig. 1. Sampling flow diagram for our first data sample. At the top is the whole study population for (a) Wave I, and (b) Wave II , with the hexagon boxes giving the sample from the whole population that we used.

included due to the influence this would have on nodal properties, such as out-degree. Including a number of students with this restriction would have led to a biased result.

Our first data sample was used to analyse emotional state behaviour at a fixed time point (see Section 3.1), with groups of depressed and not depressed individuals who satisfied the inclusion criteria within each wave being compared (Fig. 1). The complete friendship networks[4] were used when nodal properties, such as out-degree, were calculated. However, generic AID values and AID values that did not correspond to individuals who took an in-home survey were ignored when calculating a respondents out-degree (to see the analysis that counts the generic AIDs for a respondent having a friend in either their own school or sister school, but not being able to find their friends name on the roster towards the respondents out-degree, see Appendix B).

We used a second data sample to explore whether there was a relationship between the number of friends an individual believed they had in wave I and a depressed individual in wave I recovering (no longer being depressed) by wave II, or a not depressed individual in wave I becoming depressed by wave II. As waves I and II took place a year apart, these results could be interpreted as seeing if there was a relationship between the number of friends an individual believed they currently had and the individual recovering within a year if initially depressed, or the individual becoming depressed within a year if initially not depressed (see Section 3.2). This used respondents who were part of both wave I and II in-home interview samples, and satisfied our inclusion criteria in each wave. These respondents could be split into 4 groupings, covering all combinations of emotional state across the two waves (Fig. 2).

2.5. Statistical tests

We first used statistical tests to investigate if: (i) there were relationships between certain network properties for an individual and their emotional state, (ii) there was a significant difference in the number of depressed friends and number of not depressed friends an individual had, based on their emotional state, (iii) there was a significant difference in the out-degree, number of not depressed friends and number of depressed friends between not depressed individuals who either became depressed or stayed not depressed, and between depressed individuals who either recovered or stayed depressed. The statistical tests used in each case were Mann-Whitney U tests, which test the null hypothesis that there is no significant difference between two samples, in particular their median value. For these tests, we used a significance level of 0.05 on the p-values.

[4]In addition to the individuals being studied, this included other respondents in the wave I/II sample who had at least one friendship tie of any type. These networks consisted of 9354 individuals in wave I and 5550 individuals in wave II.

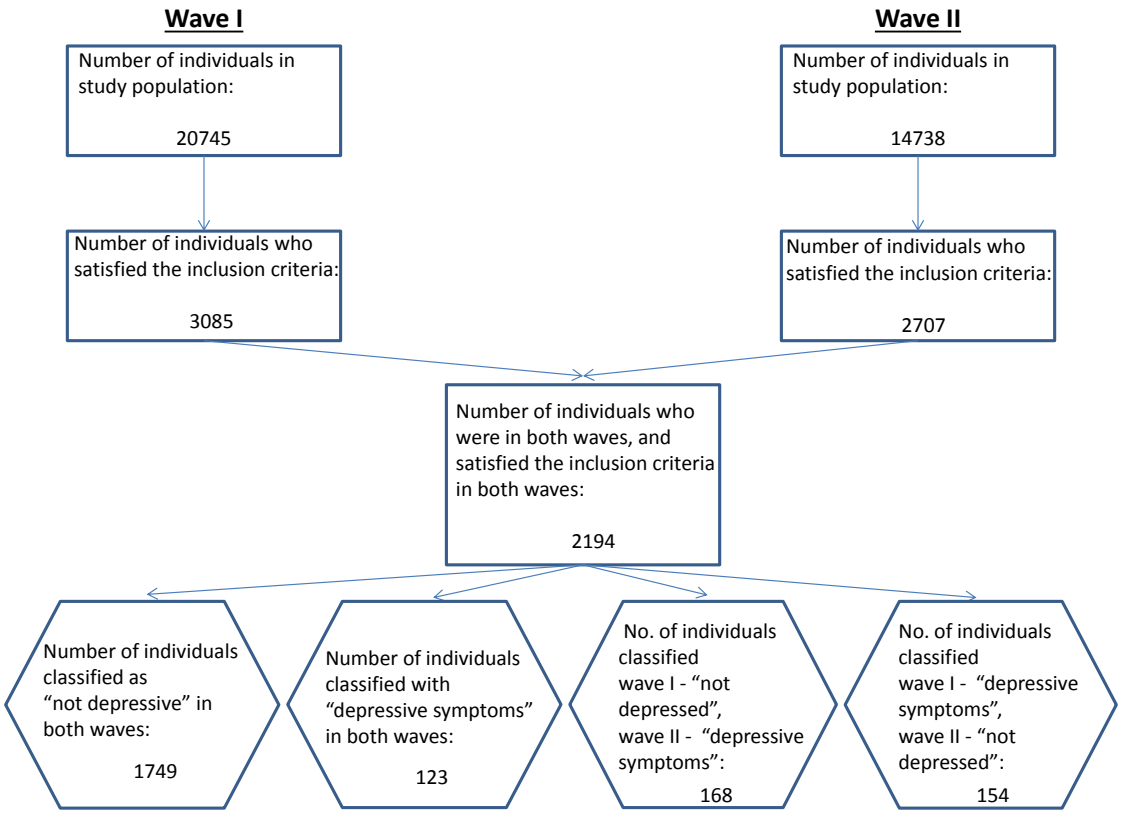

Fig. 2. Sampling flow diagram for our second data sample. At the top is the whole study population for Waves I and II, with the hexagon boxes giving the sample from the whole population that we used.

The network properties used were: (i) out-degree, (ii) in-degree, (iii) number of reciprocated ties (see Appendix A for betweenness centrality and PageRank centrality). In the context of friendship, out-degree can be interpreted as a measure of how sociable an individual is, while in-degree is often interpreted as a measure of popularity. Using the friendship data, we could construct an adjacency matrix $\mathbf{Y}$, with $Y_{ij} = 1$ when respondent i stated respondent j was a friend, and 0 otherwise. The out-degree, in-degree and number of reciprocated ties for an individual i could be calculated from $\mathbf{Y}$ as follows:

$$\begin{aligned} \text{Out-degree}: &\quad \sum_j Y_{ij} \\ \text{In-degree}: &\quad \sum_j Y_{ji} \\ \text{Reciprocated ties}: &\quad \sum_j Y_{ij} Y_{ji} \end{aligned}$$

2.6. Generalised linear models

To build on our statistical test findings, we investigated whether for a given individual: (i) their in-degree and out-degree were significant in predicting whether that individual had depressive symptoms, (ii) the number of not depressed friends and the number of friends that had depressive symptoms were significant in predicting whether that individual had depressive symptoms, (iii) the number of not depressed friends and the number of friends that had depressive symptoms were significant in predicting whether an individual would become depressed (referred to as the "switch to depression model") or recover from depressive symptoms (referred to as the "recovery model") within a year. To tackle these questions we used

generalised linear models (GLMs), a standard technique in biostatistics. In a GLM, an arbitrary function of the response variable varies linearly with the predicted values (referred to as the linear predictor). The linear predictor for individual i, η_i, is a linear combination of unknown coefficients of the covariates:

$$\eta_i = \sum_{j=1} \beta_j \omega_{ij} + \alpha$$

where β_j are the linear effects of the covariates ω_{ij}, and α is an intercept that was included. The relationship between the response variable mean, $\boldsymbol{\mu}$, and the linear predictor, η, is provided by the link function, h. It has the effect of transforming between the $(-\infty, \infty)$ range of the linear predictor and the range of the response variable. The mean of the distribution depends on the covariates as follows:

$$\mathbb{E}(\mathbf{Z}) = \boldsymbol{\mu} = h^{-1}(\eta).$$

For our three cases the response variable was a binary classification modelled using a binomial distribution. In case (i) and (ii), our binary classification was respondents who reported depressive symptoms coded with 1, with respondents who reported no depressive symptoms coded with 0. For the switch to depression model, respondents gaining depressive symptoms were coded with 1 and respondents remaining not depressed were coded with 0. For the recovery model, respondents who recovered from having depressive symptoms were coded with 1, and respondents who did not recover were coded with 0. For a response variable with a binomial distribution, the appropriate GLM to use is logistic regression. The link function was a logit function, defined as

$$\text{logit}(\text{p}_\text{i}) = \log\left(\frac{p_i}{1-p_i}\right).$$

Using logistic regression allowed us to model the probability of interest via the following expression:

$$\Pr(Z_i = 1) = \pi_i = \text{logit}^{-1}(\alpha + (\beta_1 \omega_{i1}) + (\beta_2 \omega_{i2}) + \ldots) \tag{1}$$

where ω_{ij} were the values of covariate j for individual i, and β_j were the unknown coefficients of covariate j. The α and β coefficients were estimated using the statistical package R, which used iteratively reweighted least squares to find the maximum likelihood estimates for the unknown parameters.

To determine what explanatory variables should be included in our logistic regression models, we used the Akaike information criterion (AIC). The AIC gave a measure of the relative quality of a statistical model, for a given set of data. The model that gave the smallest AIC value was the preferred choice.

2.7. Gaussian processes

We found using GLMs had its limitations. The known data for the recovery and switch to depression models suggested there were nonlinear relationships present (see Fig. 9), while a logistic regression model did not appear to provide a great fit (see Fig. 10). This led us to exploring other functional forms, motivating a Gaussian process framework.

We used Gaussian process regression (GPR) as a form of Bayesian inference of regression, which allowed us to learn a function f with confidence intervals from our data $\mathcal{D}$. Of particular interest was the conditional probability $p(f|\mathcal{D})$. To compute this we used a Gaussian process (GP), giving a prior over functions $p(f)$ that expressed beliefs about the underlying function being modelled. These are utilised in Bayesian regression, where:

$$p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})}$$

This process is completely specified by its mean function $m(x)$ and covariance function $k(x, x')$.

A set of inputs $\mathbf{X} = \{x_1, \ldots, x_n\}$, and a corresponding set of random function variables $\mathbf{f} = \{f_1, \ldots, f_n\}$ constitute a GP if every subset of the function variables $\{f_i\}_{i=1}^n$, has a multivariate Gaussian distribution:

$$p(\mathbf{f}|\mathbf{X}) \sim \mathcal{N}(\mu, \Sigma)$$

where, for $p(f(x), f(x'))$:

$$\mu = \begin{pmatrix} m(x) \\ m(x') \end{pmatrix} \quad \Sigma = \begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix}$$

Similar results are obtained for $p(f(x_1), \ldots, f(x_n))$, with μ a $n \times 1$ vector and Σ a $n \times n$ matrix.

It is often assumed that the mean of a GP is zero everywhere. In these cases, what relates one observation to another is the covariance function (i.e. the covariance function completely defines the process' behaviour). In our models we used squared exponential covariances and the Matérn class of covariances.

For the squared exponential covariance (Eq. 2), function variables a short distance apart in input space were highly correlated, whilst those far away from each other were uncorrelated:

$$k_{\mathrm{SE}}(x, x') = \sigma_f^2 \exp\left[\frac{1}{2}\left(\frac{x - x'}{l}\right)^2\right] \tag{2}$$

where l and σ_f are lengthscale and amplitude hyperparameters respectively. These hyperparameters were used to control the general properties of the covariance. The lengthscale parameter l controls the rate of decay of the correlation between observations as distance increases, with large values of l indicating that sites that are relatively far from one another are moderately correlated. The amplitude parameter σ_f controls the smoothness of the underlying function.

The Matérn class of covariance functions has the general form:

$$k_{\mathrm{Matern}}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|x - x'|}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}|x - x'|}{l}\right) \tag{3}$$

with positive hyperparameters l and ν, where K_ν is a modified Bessel function. As in the squared exponential covariance, the lengthscale parameter l controls the rate of decay of the correlation between observations as distance increases. The smoothness parameter ν controls the behaviour of the covariance function for observations separated by small distances (see Fig. 3a). This class of covariance is very flexible. The scaling in the Matérn covariance class is chosen so it includes the exponential covariance function when $\nu = 1/2$, giving very rough sample functions (Fig. 3b), while for $\nu \to \infty$ the squared exponential covariance function is obtained. Being able to strike a balance between these two extremes makes it well suited for a variety of applications [20, 21]. In particular, Eq. 3 can be simplified when $\nu = p + 1/2$, where p is a non-negative integer. The covariance function becomes a product of an exponential and a polynomial of order p [20]:

$$k_{\nu=p+\frac{1}{2}}(x, x') = \exp\left(-\frac{\sqrt{2\nu}|x - x'|}{l}\right)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{\sqrt{8\nu}|x - x'|}{l}\right)^{p-i} \tag{4}$$

With the process becoming very rough for $\nu = 1/2$, and it being hard from finite noisy training examples to distinguish between values of $\nu \geq 7/2$, the cases of interest to us were $\nu = 3/2$ and $\nu = 5/2$. For these two values Eq. 4 simplifies to:

$$\begin{aligned} k_{\nu=\frac{3}{2}}(x, x') &= \left(1 + \frac{\sqrt{3}|x - x'|}{l}\right)\exp\left(-\frac{\sqrt{3}|x - x'|}{l}\right) \\ k_{\nu=\frac{5}{2}}(x, x') &= \left(1 + \frac{\sqrt{5}|x - x'|}{l} + \frac{5|x - x'|^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}|x - x'|}{l}\right) \end{aligned}$$

A key assumption in GP modelling is that the dataset of interest can be represented as a sample from a multivariate Gaussian distribution. As the n observations in an arbitrary data set, $\mathbf{X} = \{x_1, \ldots, x_n\}$, can be thought of as a single point sampled form some n-variate Gaussian distribution, by working backwards the dataset can be partnered with a GP.

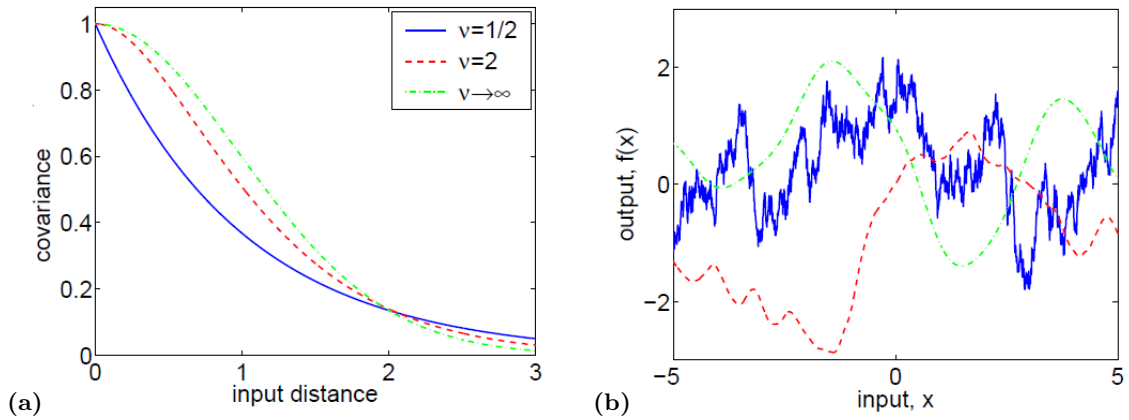

Fig. 3. (a) Covariance functions against input distance $|x - x'|$, and (b) random functions drawn from Gaussian processes with Matérn covariance functions (Eq. 3) for different values of ν, with $l = 1$. The sample functions in (b) used a discretisation of the x-axis of 2000 equally-spaced points. Reproduced from Rasmussen and Williams [20].

2.7.1. Gaussian process regression

GPR starts with a GP as a Bayesian prior distribution over functions. This was linked with the observed data we had via a Gaussian noise model. For this type of model, the observations y_i can be thought of as being generated with Gaussian white noise ϵ around an underlying function f:

$$\begin{aligned} y_i &= f(x_i) + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_n^2) \end{aligned}$$

Equivalently, the likelihood is:

$$p(\mathbf{y}|\mathbf{f}) \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I)$$

In our case, the prior on the noisy observations was mean zero and integrating over the function variables gave the marginal likelihood [20]:

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f})\, d\mathbf{f} \\ &\sim \mathcal{N}(0, \Sigma + \sigma_n^2 I) \end{aligned}$$

Note the noise term has been incorporated into the covariance. Our known data points $(\mathbf{x}, \mathbf{y})$ were treated as our training data set. Our interest was to infer $\mathbf{y}_*$[5] given the observed data vector $\mathbf{y}$ at new spatial locations $\mathbf{x}_*$ (i.e. our test data set), by combining the Gaussian prior with a Gaussian likelihood function for each of the observed values. Using the assumption that our data could be represented as a sample from a multivariate Gaussian distribution, the joint training and test marginal likelihood $p(\mathbf{y}, \mathbf{y}_*)$ was [20]:

$$p(\mathbf{y}, \mathbf{y}_*) \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

where if there were n training data points and n_* test point locations, then $K(X, X_*)$ denotes the $n \times n_*$ matrix of the covariance evaluated at all pairs of training and test location points, and similarly for the other entries $K(X, X), K(X_*, X)$ and $K(X_*, K_*)$.

[5]The objective was to predict $\mathbf{y}_*$, not the actual underlying function $\mathbf{f}_*$ Their expected values are identical, but their variances differ owing to the observational noise process.

With the prior distribution and joint density $p(\mathbf{y}, \mathbf{y}_*)$ being Gaussian, the resulting posterior distribution [22],

$$p(\mathbf{y}_*|\mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{y}_*)}{p(\mathbf{y})},$$

is also Gaussian. The mean and covariance of the posterior distribution can be computed from the observed values, their variance, and the covariance matrix derived from the prior [20]:

$$\mathbf{y}_*|\mathbf{y} \sim \mathcal{N}(\overline{\mathbf{y}_*}, \mathrm{cov}(\mathbf{y}_*))$$

where:

$$\begin{aligned} \overline{\mathbf{y}_*} &= K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y} \\ \mathrm{cov}(\mathbf{y}_*) &= K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_*) \end{aligned}$$

The covariance consists of two terms; the first term $K(X_*, X_*)$ giving the prior covariance, the subtracted (positive) term represents the information the observations give us about the function. Often only the marginal variances are required, making it sufficient to consider a single test input x^*:

$$\overline{y_*} = K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y} \quad (5)$$

$$\mathrm{var}(y_*) = k(x_*, x_*) - K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, x_*) \quad (6)$$

Eq. 5 gives the best estimate for y_*, while the uncertainty in this estimate is captured in its variance (Eq. 6).

The reliability of our regression was dependent on how well we selected the covariance function, and its hyperparameters θ. An advantage of this probabilistic GP framework was the ability to choose hyperparameters and covariances directly from the training data. The maximum a posteriori estimate of θ occurs when $p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$ is at its greatest. By Bayes' theorem, this corresponded to maximizing the marginal likelihood $\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ with respect to the hyperparameters $\boldsymbol{\theta}$. The marginal likelihood is given by [20]:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{N}(0, K + \sigma_n^2 I)$$

where its log is:

$$\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi$$

This takes into account the uncertainty in the function variables $\mathbf{f}$.

The calculations for GPR were carried out in Matlab using the GPML package. We used a composite mean function, adding a linear and a constant term. The initial hyperparameters for the mean and covariances functions were set to 0. When the covariance function was of the Matérn form we set $\nu = 3/2$. The likelihood functions were specified to be Gaussian in all our models, with varying initial values for the standard deviation of the noise. In each simulation, all the hyperparameters were learnt by optimizing the marginal likelihood, with a maximum of 100 function evaluations being performed.

3. Results

3.1. Factors that influence emotional state and network properties at a fixed time point

Using the first data sample (see Section 2.4), we explored the relationships between the emotional state of an individual and a number of different network properties measured at fixed time points (Section 3.1.1), before focusing on what significance the out-degree and in-degree of an individual had on their probability of having depressive symptoms at that time (Section 3.1.2). Finally, we investigated whether the emotional state of the friends an individual had differed based on the emotional state of the individual (Section 3.1.3).

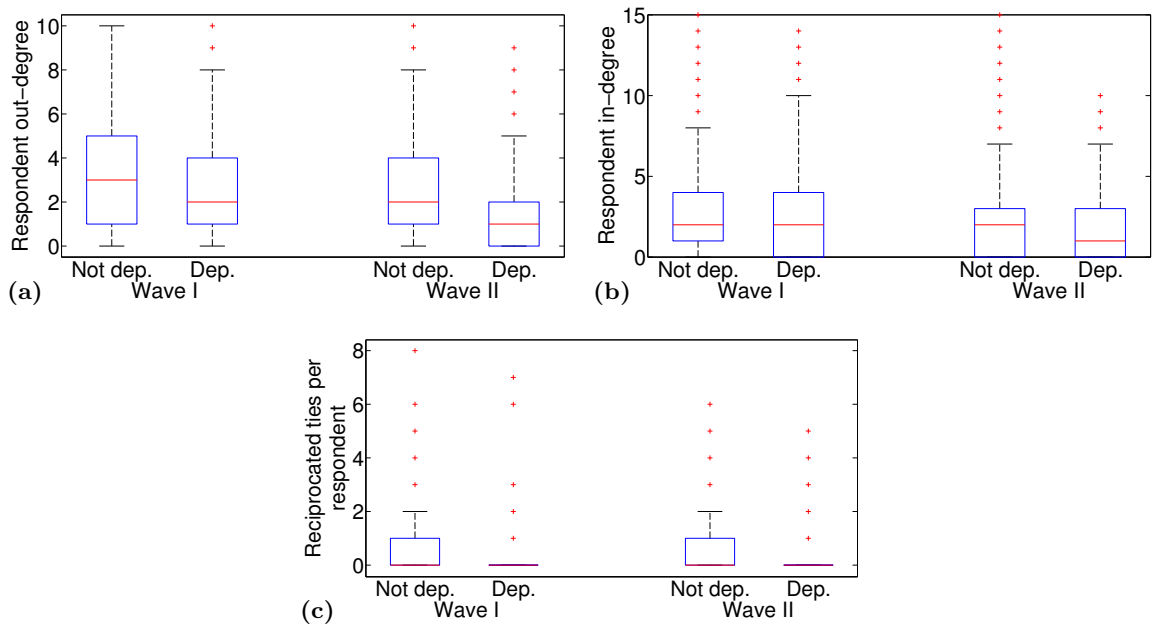

Fig. 4. Wave I and wave II box plots for the (a) out-degree, (b) in-degree, and (c) number of reciprocated ties for respondents who were not depressed and respondents with depressive symptoms. Not dep. corresponds to respondents classified as not depressive, while Dep. corresponds to respondents classified as having depressive symptoms.

3.1.1. Relationships between network properties and emotional state

Comparing the respondents out-degree, in-degree and number of reciprocated ties based on their emotional state, the respondents that were not depressed had (across both waves): higher median and upper quartile out-degree values (Fig. 4a), at least equal lower quartile, median and upper quartile in-degree values (Fig. 4b), non-zero values for reciprocated number of ties that were not considered outliers (Fig 4c). This suggested that the out-degree, in-degree and number of reciprocated ties were higher for a not depressed respondent compared to a respondent with depressive symptoms. We tested these hypotheses using one-sided Mann-Whitney U tests. This found, for both waves I and II, a not depressed respondent was more likely to have higher values than a respondent with depressive symptoms for out-degree (wave 1 p-value: $p < 0.0001$, wave 2 p-value: $p < 0.0001$), in-degree ($p = 0.0006$, $p = 0.0001$) and number of reciprocated ties ($p = 0.0008$, $p < 0.0001$).

3.1.2. Significance of out-degree and in-degree

Using logistic regression models, we investigated whether the out-degree and/or in-degree of a given individual had a significant impact on the probability of that individual having depressive symptoms. We obtained separate models for our wave I and II datasets, containing out-degree and in-degree as regression parameters. It was found both were an improvement over a reduced model that contained only an intercept term (wave I AIC values - complete model: 2367.9, reduced model: 2381.7; wave II AIC values - complete model: 2000.6, reduced model: 2040.4)

Letting π_i be the probability of individual i being depressed, the models obtained for waves I and II were (following the format of Eq. 1):

$$\text{Wave I:} \quad \pi_i = \text{logit}^{-1}(-1.61 + (-0.0784 \times \text{out-degree}_i) + (-0.0310 \times \text{in-degree}_i)) \tag{7}$$

$$\text{Wave II:} \quad \pi_i = \text{logit}^{-1}(-1.53 + (-0.190 \times \text{out-degree}_i) + (-0.0320 \times \text{in-degree}_i)) \tag{8}$$

We let β_1 and β_2 correspond to the model coefficients for out-degree and in-degree.

Table 1. The wave I logistic regression model parameters, with the coefficient maximum likelihood estimates (MLEs) and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value ($\Pr(>\|z\|)$)
Out-degree	-0.0784	0.0264	-2.97	0.00352
In-degree	-0.0310	0.0242	-1.28	0.201

Table 2. The wave II logistic regression model parameters, with the coefficient MLEs and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value ($\Pr(>\|z\|)$)
Out-degree	-0.190	0.0386	-4.91	9.05×10^{-7}
In-degree	-0.0320	0.0316	-1.01	0.312

We then performed significance tests on the regression parameters in the wave I model (Table 1) and wave II model (Table 2). We tested the null hypothesis $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ to see if the out-degree and/or in-degree respectively were significant. The p-values of these tests were calculated using z-values, which gave the number of standard deviations an observation was above the mean. Both models found that, when fixing the in-degree, the out-degree had a significant impact on the probability of the given individual being depressed (wave I: $p = 0.003$, wave II: $p < 0.0001$). When the out-degree was fixed, the in-degree was not significant in either model ($p = 0.201$, $p = 0.312$).

3.1.3. Relationship between current emotional state and friends emotional state

For each respondent, focusing only on the number of individuals they reported as being friends who were depressed (Fig. 5a) or not depressed (Fig. 5b) gave an insight into whether the emotional state of a respondents friends influenced the respondents own emotional state. The majority of respondents had either 0 depressed friends or 1 friend that was depressed, regardless of the respondents emotional state. In contrast, the majority of respondents had at least one not depressed friend. Overall, respondents with depressive symptoms appeared to have a lower number of friends that were not depressed (lower quartile of 0 and upper quartile 3 in wave I, lower quartile of 0 and upper quartile of 2 in wave II) compared to not depressed respondents (lower quartile of 1 and upper quartile 4 in wave I, lower quartile of 0 and upper quartile of 3 in wave II). Applying one-sided Mann-Whitney U tests, a respondent who did not have depressive symptoms was likely to have a greater number of not depressed friends ($p < 0.0001$, $p < 0.0001$) compared to a respondent who was classified as having depressive symptoms. Applying two-sided Mann-Whitney tests found there was no significant difference in the number of depressed friends based on emotional state ($p = 0.24$, $p = 0.17$).

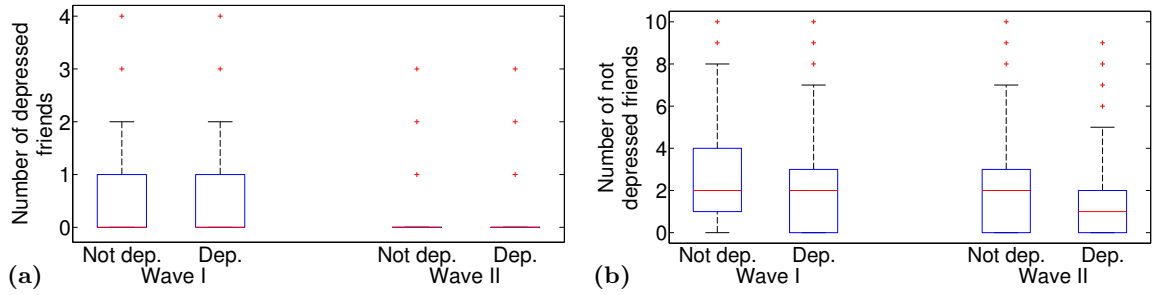

Fig. 5. Wave I and wave II box plots corresponding to the (a) number of depressed friends, and (b) number of not depressed friends for respondents who were not depressed and respondents with depressive symptoms.

Table 3. The wave I friends emotional state logistic regression model parameters, with the coefficient MLEs and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value ($\Pr(>\|z\|)$)
Not depressed friend	-0.132	0.0269	-4.92	8.5×10^{-7}
Depressed friend	0.204	0.0841	2.42	0.0154

Table 4. The wave II friends emotional state logistic regression model parameters, with the coefficient MLEs and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value ($\Pr(>\|z\|)$)
Not depressed friend	-0.213	0.0359	-5.93	3.03×10^{-9}
Depressed friend	-0.0975	0.130	-0.752	0.452

When we used a logistic regression model on each dataset, with number of not depressed friends and number of depressed friends as regression parameters, it was found both were an improvement over a reduced model that contained only an intercept term (wave I AIC values - complete model: 2357.4, reduced model: 2381.7; wave II AIC values - complete model: 2002.5, reduced model: 2040.4).

When performing significance tests on the regression parameters in the wave I model (Table 3) and wave II model (Table 4), where the p-values of these tests were calculated using z-values, both models found that when fixing the number of depressed friends, the number of not depressed friends had a significant impact in the probability of the respondent being depressed (wave 1: $p < 0.0001$, wave 2: $p < 0.0001$). As the coefficient for not depressed friends was negative in both models (-0.132 and -0.213 respectively), an individual increasing their number of not depressed friends resulted in a reduced probability of the individual themselves having depressive symptoms. When fixing the number of not depressed friends, the wave I model found the number of depressed friends was significant ($p = 0.0154$), while the wave II model found the number of depressed friends was not significant ($p = 0.452$).

3.2. Predicting those most at risk of having a change in emotional state

The results in this section were obtained using the second data sample (see Section 2.4).

Figure 6 shows the quartiles and range of values for the 4 groupings. For respondents who were not depressed in wave I, compared to those that became depressed in wave II, those that stayed not depressed in wave II had greater median (3 vs 2) and upper quartile values (5 vs 4). For respondents who were depressed in wave I, compared to those that remained depressed in wave II, those that became not depressed in wave II had greater upper quartile values (5 vs 3.75) and equal median and lower quartile values. We hypothesised that respondents who became depressed had slightly fewer friends initially compared to those who did not, and that the respondents who recovered from depression started out with more friends compared to the respondents who did not recover. We carried out Mann-Whitney U tests to evaluate this. For respondents who were not depressed in wave I, the out-degree was likely to be higher for a respondent that remained not depressed compared to a respondent that had become depressed by wave II ($p = 0.0002$). For respondents who were depressed in wave I, the out-degree for a respondent that was no longer depressed in wave II was likely to be greater than the out-degree for a respondent that remained depressed ($p = 0.009$).

As the total number of friends appeared to influence which respondents had a change in emotional state by wave II (i.e. within a year), we investigated if the same result held if the number of depressed friends and number of not depressed friends at the wave I time point were used instead. For the respondents who were not depressed in wave I (Fig. 7), the number of depressed friends per respondent was likely to be no different for a respondent that remained not depressed by wave II compared to a respondent that became depressed by wave II (two-tailed Mann-Whitney U test: $p = 0.99$). The median number of not depressed friends was higher for respondents that had stayed not depressed by wave II compared to respondents that

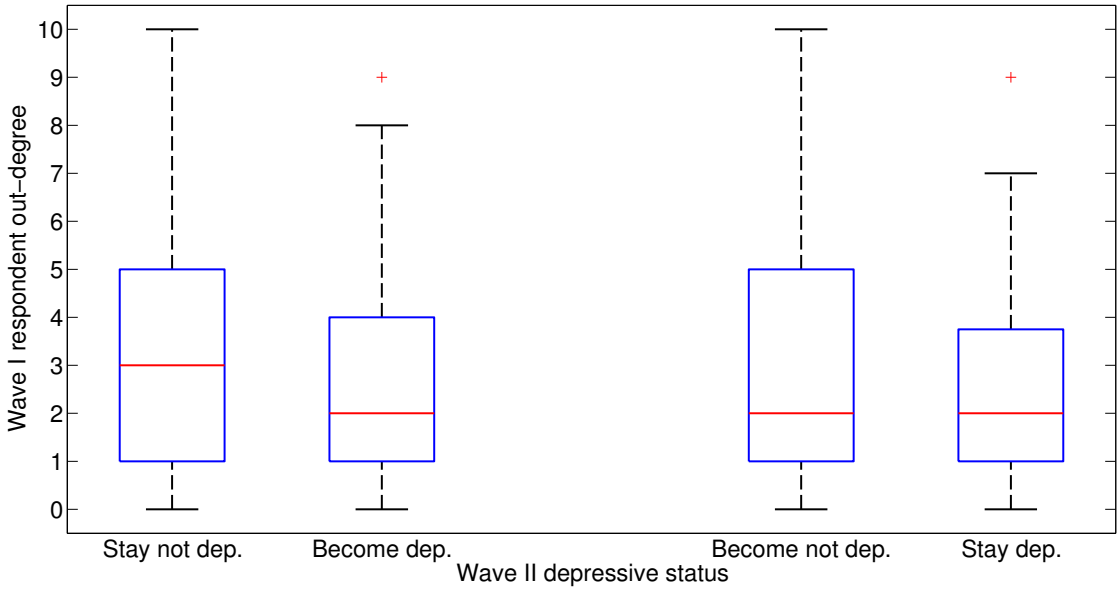

Fig. 6. Box plots of the out-degree in wave I for the four groups. The two left hand box plots contain all respondents who were not depressed in wave I, while the two right hand box plots contain all respondents who had depressive symptoms in wave I.

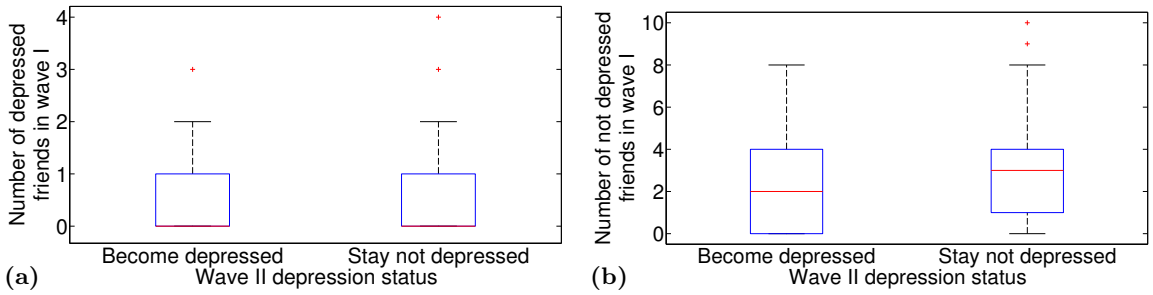

Fig. 7. Box plots of the (a) number of depressed friends in wave I, and (b) number of not depressed friends in wave I for the two groups of respondents who were not depressed in wave I and then either stayed not depressive or had depressive symptoms in wave II.

had become depressed by wave II (3 vs 2). A one-tailed Mann-Whitney U test showed the median to be significantly lower ($p = 0.0002$).

For the respondents who were depressed in wave I (Fig. 8), the group that became not depressed and the group that stayed depressed by wave II both had a median number of depressed friends of 0. However, the probability of a respondent who was no longer depressed having a higher number of depressed friends than a respondent who remained depressed was significant ($p = 0.04$). Furthermore, the median number of not depressed friends was greater for respondents that no longer had depressive symptoms in wave II compared to respondents that still had depressive symptoms (2 vs 1). In this case, the median was shown to be significantly lower ($p = 0.02$).

The recovery and switch to depression logistic regression models (see Section 2.6) were then used to further our analysis. We found both were an improvement over reduced models that contained only an intercept term (Switch to depression model AIC values - complete model: 1132.1, reduced model: 1140.8; Recovery model AIC values - complete model: 379.2, reduced model: 382.5).

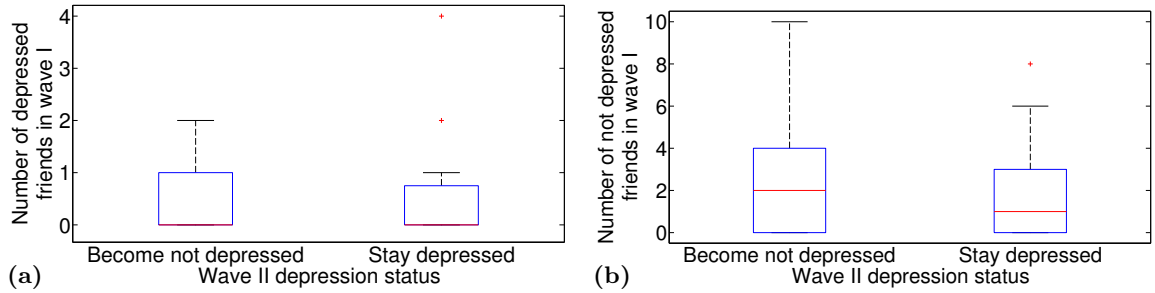

Fig. 8. Box plots of the (a) number of depressed friends in wave I, and (b) number of not depressed friends in wave I for the two groups of respondents who had depressive symptoms in wave I and then either became not depressive or still had depressive symptoms in wave II.

Table 5. The switch to depression logistic regression model parameters, with the coefficient MLEs and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value (Pr($>|z|$))
Not depressed friend	-0.135	0.0392	-3.44	0.000572
Depressed friend	0.0717	0.131	0.545	0.586

When performing significance tests on the regression parameters in the switch to depression model (Table 5), with the p-values of these tests calculated using z-values, we found that when fixing the number of depressed friends, the number of not depressed friends had a significant impact in the probability of a not depressed respondent becoming depressed ($p = 0.0005$). The coefficient for not depressed friends was negative (-0.135), thus an individual increasing their number of not depressed friends resulted in a reduced probability of a not depressed individual becoming depressed. When fixing the number of not depressed friends, the number of depressed friends was not significant ($p = 0.59$). For the recovery model (Table 6), the number of not depressed friends had a significant impact in the probability of a depressed respondent recovering ($p = 0.02$). The coefficient for the not depressed friends parameter was positive (0.138), thus a depressed individual increasing their number of not depressed friends resulted in a higher probability of them recovering. When fixing the number of not depressed friends, the number of depressed friends was again not significant ($p = 0.42$). However, we note that the coefficient for the depressed friends parameter was also positive (0.151), meaning having depressed friends still increased the probability of recovery.

3.3. Emotional States - do they show causal effects?

There was no significant relationship[6] between the number of depressed contacts in wave I, and the proportion of respondents with that number of depressed contacts who went from having depressive symptoms in wave I to not depressed in wave II (Fig. 9b, coeff. $= -0.14$, $p = 0.22$), or went from being not depressed in wave I to having depressive symptoms in wave II (Fig. 9d, coeff. $= -0.014$, $p = 0.37$).

For the relationship between the number of not depressed contacts in wave I, and the proportion of respondents with that number of not depressed contacts that went from having depressive symptoms in wave I to being not depressed in wave II, we found a significant positive correlation (Fig. 9a, coeff. $= 0.056$, $p = 0.00046$). A significant negative correlation was found between the number of not depressed contacts in wave I, and the proportion of respondents with that number of not depressed contacts that went from being not depressed in wave I to having depressive symptoms in wave II (Fig. 9c, coeff. $= -0.011$, $p = 0.00024$). However, a linear relationship did not appear to be a good fit in either case. As the coefficients obtained in

[6]T-tests were used to test the significance of the correlations, with a null hypothesis of the true correlation was equal to 0.

Table 6. The recovery logistic regression model parameters, with the coefficient MLEs and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value ($\Pr(>\|z\|)$)
Not depressed friend	0.138	0.0598	2.31	0.0210
Depressed friend	0.151	0.190	0.793	0.428

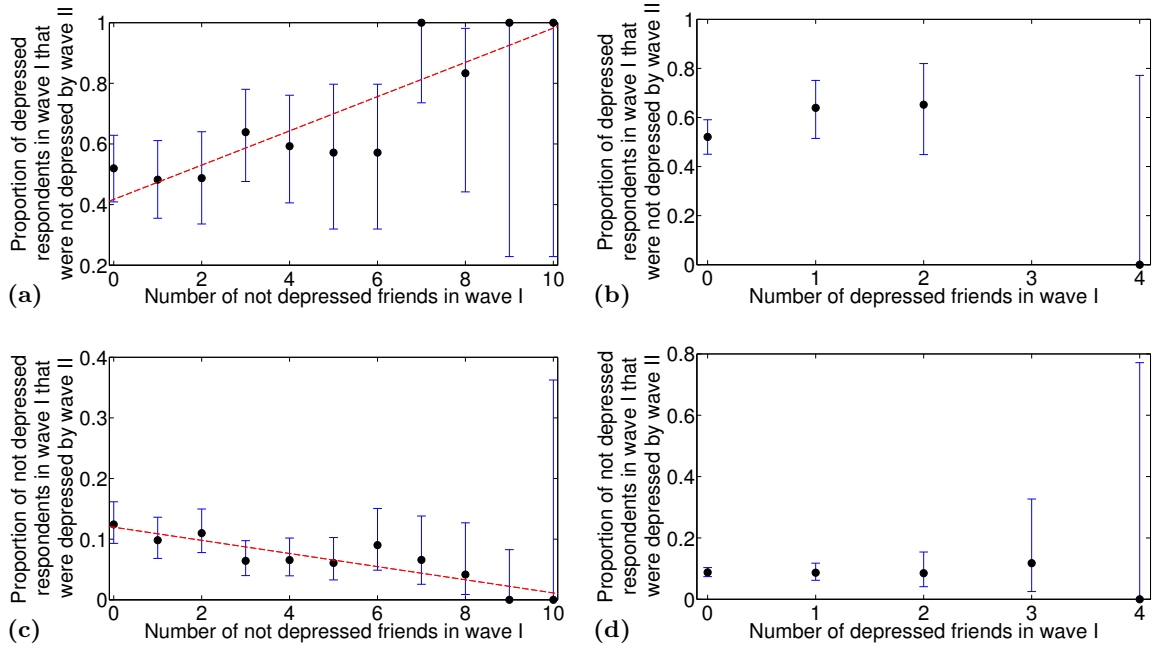

Fig. 9. (a) Recovering from depression: the probability of transitioning from depressed to not depressed increases with the number of not depressed contacts (slope = 0.056, $p = 0.00046$, intercept = 0.42). (b) Recovering from depression: the probability of transitioning from depressed to not depressed does not depend on the number of depressed contacts (slope = -0.14, $p = 0.22$, intercept = 0.70). (c) Becoming depressed: the probability of transitioning from not depressed to depressed decreases with the number of not depressed contacts (slope = -0.011, $p = 0.00024$, intercept = 0.12). (d) Becoming depressed: the probability of transitioning from not depressed to depressed does not depend on the number of depressed contacts (slope = -0.014, $p = 0.37$, intercept = 0.10). The error bars on the observed proportions (black dots) are the 95% binomial proportion confidence intervals, calculated using the Jeffreys interval (see Appendix C).

our switch to depression and recovery logistic regression models also did not truly fit the data (Figs. 10a-10b), we wanted a principled way of saying that the relationship between number of contacts and probability of change in emotional state was non-linear. This was explored using Gaussian process regression analysis, which suggested the relationship between number of not depressed friends and a change in emotional state was non-linear (Figs 11-12).

For the recovery model, using a squared exponential function gave a predicted function that was choppy and touched the known data points, suggesting it was over-fitted (Fig. 11a). In comparison, using the Matérn class of covariances gave a more realistic fit in the form of a monotonically increasing predicted function, with steeper probability increases occurring between 2 to 3 and 6 to 7 not depressed friends (Fig. 11b).

For the switch to depression model, using either a squared exponential covariance (Fig. 12a) or the Matérn class of covariances (Fig. 12b) gave similar predicted functions, though the Matérn class of covariances did give a tighter 95% confidence interval. The probability decreased as the number of not depressed friends

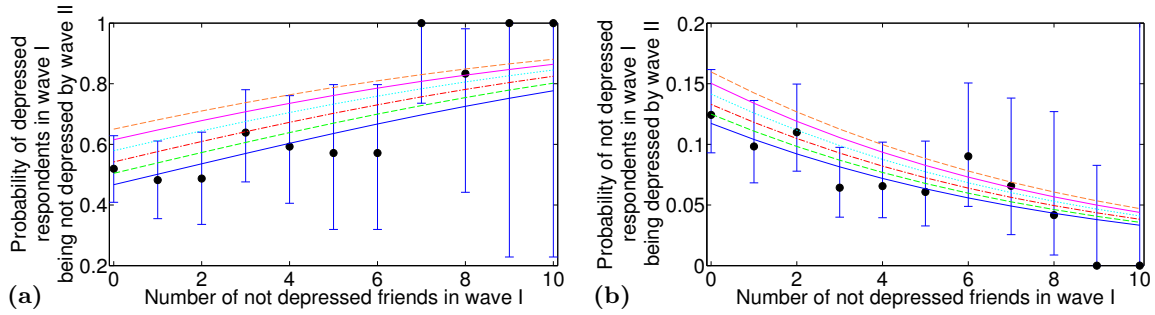

Fig. 10. Logit curves for the logistic regression (a) recovery model, and (b) switch to depression model for a fixed number of depressed friends. Each curve corresponds to the following fixed number of depressed friends: dark blue solid curve - 0, green dashed curve - 1, red dot-dash curve - 2, cyan dotted curve - 3, magenta solid curve - 4, orange dashed curve - 5. The error bars on the observed proportions (black dots) are the 95% binomial proportion confidence intervals, calculated using the Jeffreys interval.

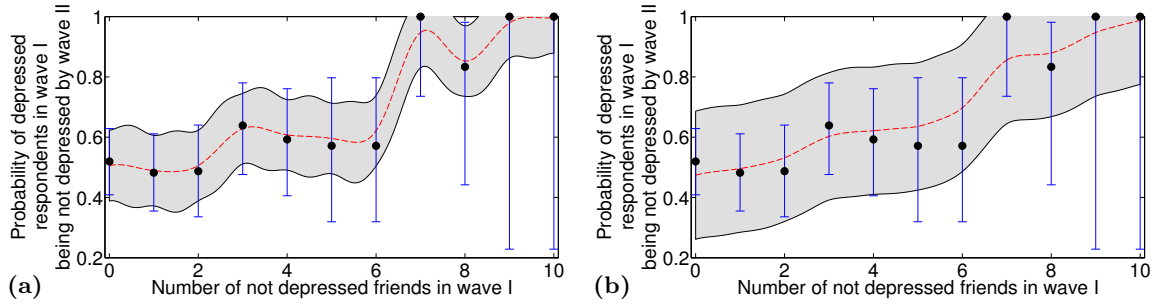

Fig. 11. Gaussian process regression for the recovery model using a composite mean function and a (a) squared exponential covariance function , and (b) Matérn class of covariance functions (with $\nu = 3/2$). In both cases, the likelihood function was Gaussian, with the initial hyperparameter for the standard deviation of the noise set at 0.05. A maximum of 100 function evaluations were performed. The red dotted line corresponds to the inferred predictions. The solid black lines give the boundaries of the 95% confidence interval. The grey shaded region gives the values within the 95% confidence interval. The error bars on the observed proportions (black dots) are the 95% binomial proportion confidence intervals, calculated using the Jeffreys interval.

increased from 0 to 4, then plateaus between 4 and 6 not depressed friends before decreasing again and reaching a probability of 0 for 10 not depressed friends.

4. Discussion

4.1. Summary of findings

As the number of depressed friends an individual had did not statistically effect the change in emotional state (Figs. 9b and 9d), this suggests the number of depressed friends has no causal effect on the emotional state of the individual. Furthermore, our findings using GPR suggest that the relationship between number of not depressed friends and a change in emotional state is non-linear (Figs 11-12). For the recovery model, using a Matérn class of covariances gave a monotonically increasing function, with steeper probability increases occurring between having 2 and 3 not depressed friends, and 6 and 7 not depressed friends (Fig. 11b). This suggests there may be threshold values of not depressed friends, that when surpassed could result in a sudden increase in the probability of a depressed individual recovering from depressive symptoms. This advanced our findings using a logistic regression model and statistical tests, which found the number of not depressed

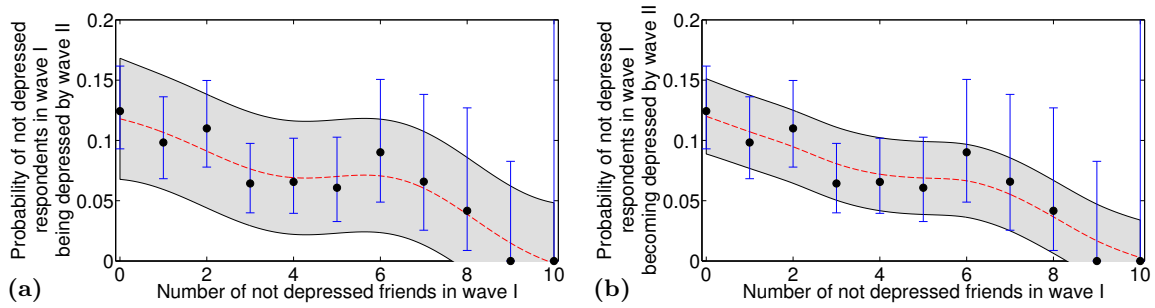

Fig. 12. Gaussian process regression for the switch to depression model using a composite mean function and a (a) squared exponential covariance function , and (b) Matérn class of covariance functions (with $\nu = 3/2$). In both cases, the likelihood function was Gaussian, with the initial hyperparameter for the standard deviation of the noise set at (a) 1, and (b) 0.01. A maximum of 100 function evaluations were performed. The red dotted line corresponds to the inferred predictions. The solid black lines give the boundaries of the 95% confidence interval. The grey shaded region gives the values within the 95% confidence interval. The error bars on the observed proportions (black dots) are the 95% binomial proportion confidence intervals, calculated using the Jeffreys interval.

friends was the influential factor in which not depressed individuals were more likely to became depressed between waves I and II, rather than the number of depressed friends (see Table 5 and Fig. 7). The predicted functions for the switch to depression model had regions of steady probability decrease and regions where the probability plateaued (Fig. 12). Though the general trend was that the more not depressed friends a not depressed individual had, the less risk they had of becoming depressed (Fig. 9c), our predicted function suggests that gaining a single additional not depressed friend is not necessarily beneficial straight away in decreasing the probability of a not depressed individual becoming depressed. Once again, this improves our findings found with a logistic regression model, which told us the number of not depressed friends was a more influential factor than number of depressed friends in predicting which depressed individuals were more likely to recover between waves I and II (see Table 6).

On the other hand, in the recovery model the coefficient for the depressed friends parameter was positive (0.151), which means having depressed friends still increases the probability of recovery. Thus, to try and predict if an individual would recover from being depressed, looking at the number of friends the individual has, regardless of the depressive status of each friend, appears to be a suitable strategy. In addition, our statistical tests found that not depressed respondents who became depressed had slightly fewer friends initially compared to those who remained not depressed, and that the respondents who recovered from depression started out with more friends compared to the respondents who did not recover (Fig. 6). Thus, despite the number of not depressed friends appearing to be more influential, having depressed friends in addition to not depressed friends that boost the total number of friends an individual has could result in their likelihood of changing emotional state being altered.

In our analysis of emotional state behaviour at a fixed time point, using a logistic regression model (Eqs. (7),(8)) we found the number of friends a person believed they had (their out-degree) more significant factor in predicting their current emotional state than the number of people that stated the given individual was their friend (their in-degree) (see Tables 1-2). Furthermore, our logistic regression model with number of not depressed friends and depressed friends as regression parameters (see Tables 3-4), found the probability of an individual being depressed was more influenced by their number of not depressed friends, rather than their number of depressed friends. Having more not depressed friends reduces the probability of an individual being depressed at that particular time. This reinforces our statistical test results, with not depressed individuals more likely to have a greater number of not depressed friends compared to depressed individuals, while there is no significant difference for an individual in their number of depressed friends based on their emotional state (Fig. 5). This suggested people who are depressed do not cluster together.

4.2. Limitations of the study

Though the CES-D scale is a widely used tool for assessing symptoms in populations, our method of using it to give a binary classification of the emotional state of an individual has limitations. As stated previously, the CES-D can only indicate depressive symptoms and does not represent a medical diagnoses, though we did use the score cut-off that has been shown to correlate with a clinical diagnosis of depression.

Furthermore, there were issues with missing data. The number of friends the respondents were allowed to list was still limited to at most 5 per gender. Any additional friends they had were not reported. For individuals within saturated schools, the in-home survey response rates were not 100% in either wave. Information was also lost through a respondent not finding a friend on the school roster who they knew went to that school, resulting a generic AID code being used instead. Consequently, these combination of factors resulted in an incomplete friendship network being formed for the saturated schools, which in turn impacts the reliability of our findings.

A limitation of GPR was by using approximate Bayesian inference we did not target to do exact inference. An alternative technique that would address this issue is Markov chain Monte Carlo methods, as the results get increasingly more accurate as the number of samples tends to infinity.

4.3. Findings in the context of other research findings

One of our findings is that depressed individuals may have fewer depressed and non-depressed friends compared to non-depressed individuals. This agrees with a result found by Schaefer et al. [23], who used the Add Health data and a dynamic network model to find that depression homophily could emerge due to the social withdrawal that accompanies depression. This withdrawal mechanism reduces the number of friends, but it does not directly affect the distribution of depression among one's friends (i.e. depressed individuals may have fewer depressed and non-depressed friends).

Using datasets such as Add Health, Christakis and Fowler [24] propose that human social networks may exhibit a "three degrees of influence" property with respect to phenomena as diverse as obesity, smoking, and happiness. Our findings that the not depressed emotional state appears to have a causal effect at the friend level can be built upon to investigate if it extends to friends of friends (two degrees of separation) and friends of friends of friends (three degrees of separation).

4.4. Further Work

A number of extensions to this work can be pursued. Our analysis can be extended by looking for evidence of confounding variables, where the number of school friends an individual has may just a proxy for another factor, such as their socio-economic status. There could also be differences between gender. Rueger et al. [12] examined the importance of examining gender differences in the social experience of adolescents, so this should be considered in further research. As the social network in our dataset was restricted to friends from the same school or sister school, a further extension would be to perform a similar analysis on a more realistic social network for an individual containing all their friends and family, with no age group restriction.

Acknowledgement

Thanks to Add Health for allowing the use of their dataset, and to Robert Goudie for providing R code and advice on how to utilise the data.

References

[1] S. M. Goodreau, J. A. Kitts, and M. Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103–125, 2009.

[2] P. S. Bearman, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110:44–91, 2004.

[3] M. J. Keeling and K. T. D. Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2:295–307, 2005.

[4] N. A. Christakis and J. H. Fowler. The collective dynamics of smoking in a large social network. *The New England Journal of Medicine*, 358(21):2249–2258, 2008.
[5] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, 357(4):370–379, 2007.
[6] M. M. Ali, A. Amialchuk, S. Gao, and F. Heiland. Adolescent Weight Gain and Social Networks: Is There a Contagion Effect? *Applied Economics*, 44(23):2969–2983, 2011.
[7] T. E. Joiner and J. Katz. Contagion of Depressive Symptoms and Mood: Meta-analytic Review and Explanations From Cognitive, Behavioral, and Interpersonal Viewpoints. *Clinical Psychology: Science and Practice*, 6(2):149–164, 1999.
[8] K. Ueno. The effects of friendship networks on adolescent depressive symptoms. *Social Science Research*, 34(3):484–510, 2005.
[9] World Health Organization. *Depression - Fact Sheet No.369* . Available at: `http://www.who.int/mediacentre/factsheets/fs369/en/index.html`, 2012. [Accessed: 19 July 2013].
[10] H. Green, Áine McGinnity, H. Meltzer, T. Ford, and R. Goodman. *Mental health of children and young people in Great Britain 2004*. London:Palgrave, 2005.
[11] National Institute of Mental Health. *Major Depressive Disorder in Children* . Available at: `http://www.nimh.nih.gov/statistics/1mdd_child.shtml`, 2013. [Accessed: 19 July 2013].
[12] S. Y. Rueger, C. K. Malecki, and M. K. Demaray. Relationship Between Multiple Sources of Perceived Social Support and Psychological and Academic Adjustment in Early Adolescence: Comparisons Across Gender. *Journal of Youth and Adolescence*, 39(1):47–61, 2010.
[13] N. Mead, H. Lester, C. Chew-Graham, L. Gask, and P. Bower. Effects of befriending on depressive symptoms and distress: systematic review and meta-analysis. *The British Journal of Psychiatry*, 196(2):96–101, 2010.
[14] R. Pastor-Satorras and A. Vespignani. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, 86:3200–3203, 2001.
[15] A. L. Hill, D. G. Rand, M. A. Nowak, and N. A. Christakis. Emotions as infectious diseases in a large social network: the SISa model. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701):3827–3835, 2010.
[16] K. Harris, C. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J. Udry. *The National Longitudinal Study of Adolescent Health: Research Design* . Available at: `http://www.cpc.unc.edu/projects/addhealth/design`, 2009. [Accessed: 02 July 2013].
[17] L. S. Radloff. The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401, 1977.
[18] R. E. Roberts, P. M. Lewinsohn, and J. R. Seeley. Screening for Adolescent Depression: A Comparison of Depression Scales. *The Journal of the American Academy of Child and Adolescent Psychiatry*, 30(1):58–66, 1991.
[19] R. E. Roberts. Reliability of the CES-D scale in different ethnic contexts. *Psychiatry Research*, 2(2):125–134, 1980.
[20] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
[21] J. A. Hoeting, R. A. Davis, A. A. Merton, and S. E. Thompson. Model Selection for Geostatistical Models. *Ecological Applications*, 16(1):87–98, 2006.
[22] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2002.
[23] D. R. Schaefer, O. Kornienko, and A. M. Fox. Misery does not love company: Network selection mechanisms and depression homophily. *American Sociological Review*, 76(5):764–785, 2011.
[24] N. A. Christakis and J. H. Fowler. Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior. *Statistics in Medicine*, 32(4):556–577, 2013.
[25] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.
[26] L. D. Brown, T. T. Cai, and A. Dasgupta. Interval Estimation for a Binomial Proportion. *Statistical Science*, 16:101–133, 2001.
[27] H. Jeffreys. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Appendices

Appendix A. Betweenness Centrality and PageRank Centrality Analysis

Betweenness centrality is a measure of the centrality of a node in a network. It is equal to the number of shortest paths from all vertices to all others that pass through that node. The betweenness centrality of a node v is defined as

$$B(v) = \sum_{s \neq \nu \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{A.1}$$

where σ_{st} is the total number of shortest paths from node s to t and $\sigma_{st}(v)$ is the number of those paths that pass through ν. Note that the betweenness centrality of a node scales with the number of pairs of nodes. We normalise the betweenness of a node according to

$$B_{\mathbf{norm}}(v) = \frac{2B(v)}{n(n-3)(n+2)},$$

where $B(v)$ is the raw betweenness (Eq. A.1), and n is the number of vertices in the graph.

In many circumstances the importance of a node in a network is increased by having connections to other nodes that are themselves important. This gave rise to the idea of eigenvector centrality, where each node has a score corresponding to importance that is a proportion of the sum of importance scores of its neighbours. The form of eigenvector centrality we used was the PageRank, as defined by Newman[25], where the centrality a node derived from its network neighbours was proportional to their neighbours centrality divided by their out-degree. The result was nodes that point to many others pass only a small amount of centrality on to each of those others, even if their own centrality was high. The PageRank centrality for node i, x_i, is given by

$$x_i = \alpha \sum_j Y_{ij} \frac{x_j}{k_j^{\text{out}}} + \beta, \tag{A.2}$$

where α and β are positive constants, x_j is the PageRank centrality of node j and k_j^{out} is the out degree of node j. Note, if any nodes in the network had an out-degree $k_i^{\text{out}} = 0$, then the first term in Eq. A.2 would be undefined. However, nodes with no out-going edges should contribute zero to the centrality of any other node. This effect was achieved by artificially setting $k_i^{\text{out}} = 1$ for all such nodes. The second term corresponds to every node being given a small amount of centrality, regardless of its position in the network or the centrality of its neighbours. Its inclusion meant that nodes with zero in-degree still got a centrality of β, and nodes they pointed to derived some advantage from being pointed to. Consequently, any node that was pointed to by many others had a high centrality, but nodes that were pointed to by others with a high centrality themselves still did better. In matrix notation, the vector of PageRank centralities for all individuals in the network, $\mathbf{x}$, satisfies

$$\mathbf{x} = \alpha \mathbf{Y} \mathbf{D}^{-\mathbf{1}} \mathbf{x} + \beta \mathbf{1},$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \max(k_i^{\text{out}}, 1)$. For the PageRank centrality expression to converge, α had to be set less than the inverse of the largest eigenvalue of $\mathbf{Y}\mathbf{D}^{-\mathbf{1}}$, which was 1 in our case [25]. We used $\alpha = 0.85$. For simplicity, β was set to be 1.

The mean normalised betweenness centrality and mean PageRank centrality for respondents with depressive symptoms were slightly lower than respondents with no depressive symptoms in both waves I and II (Fig. A.13). We tested if the means for the not depressive group were statistically significantly greater than the group with depressive symptoms using one-sided two-sample t-tests. This found that, for both waves I and II, the not depressive groups had significantly greater mean values for the PageRank centrality than the depressive symptom groups ($p < 0.0001$, $p < 0.0001$). However, the mean normalised betweenness centrality was not significantly greater for the not depressive group compared to the depressive symptom groups in either wave ($p = 0.27$, $p = 0.069$).

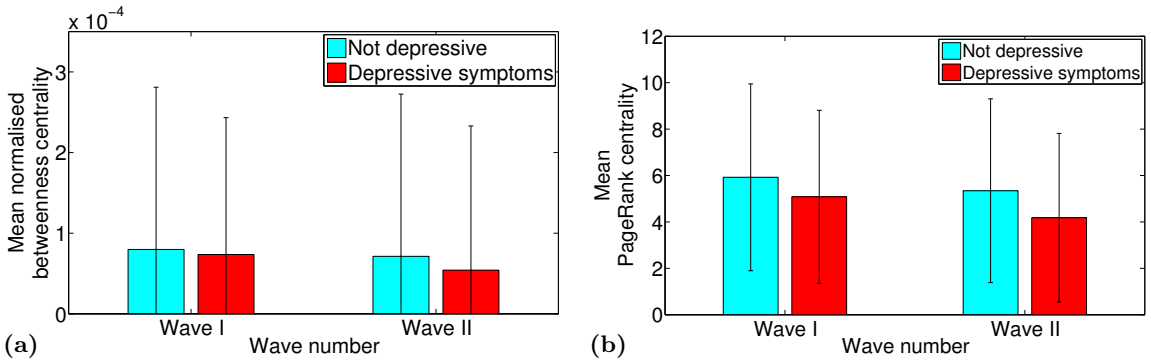

Fig. A.13. Wave I and wave II mean values for the (a) normalised betweenness centrality, and (b) PageRank centrality (with $\alpha = 0.85$) for respondents who were not depressed and respondents with depressive symptoms. The error bars give the range of values that were both within 1 standard deviation of the mean and greater than 0.

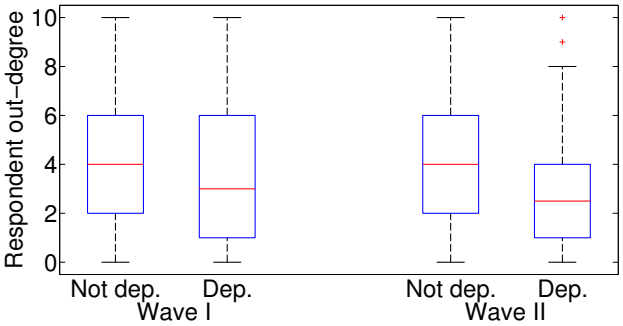

Fig. B.14. Wave I and II mean out-degrees for the no depression and depressive symptom respondents. Not dep. corresponds to respondents classified as not depressive, while Dep. corresponds to respondents classified as having depressive symptoms.

Appendix B. Full out-degree Analysis

Only blank entries and entries corresponding to a friend not being from the same school as the respondent (or their sister school), corresponding to the code 77777777, were removed when calculating the out-degree for each respondent in this case.

Comparing the out-degree values of not depressed and depressive symptom respondents, for both waves, showed the not depressed group had greater values for the lower quartiles and medians, with at least matching upper quartile values (Fig B.14). Applying a one-sided Mann-Whitney U test found the out-degree for a respondent with no depressive symptoms was more likely to be greater than the out-degree for a respondent with depressive symptoms in both waves ($p < 0.0001$, $p < 0.0001$).

We produced logistic regression models containing both out-degree and in-degree terms and compared them to reduced models that only contained an intercept term.

For wave I, the complete model had an AIC of 2366.2, while the reduced model had an AIC of 2381.7, indicating our model containing out-degree and in-degree terms was an improvement over a model with just an intercept term. Following the format of Equation 1, with the probability of individual i not being depressed given by π_i, the model for wave I was:

$$\pi_i = \text{logit}^{-1}(-1.56 + (-0.0701 \times \text{out-degree}_i) + (-0.0328 \times \text{in-degree}_i) \tag{B.1}$$

We let β_1 and β_2 correspond to the model coefficients for out-degree and in-degree.

Table B.7. The wave I logistic regression model parameters, with the coefficient maximum likelihood estimates and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value ($\Pr(>\|z\|)$)
Out-degree	-0.0701	0.0215	-3.26	0.00112
In-degree	-0.0328	0.0236	-1.39	0.165

Table B.8. The wave II logistic regression model parameters, with the coefficient MLEs and standard deviations, parameter z-values and p-values. All values are given to 3 significant figures.

Regression Parameter	Coefficient MLE	Standard deviation	z-value	p-value ($\Pr(>\|z\|)$)
Out-degree	-0.114	0.0246	-4.63	3.64×10^{-6}
In-degree	-0.0511	0.0313	-1.63	0.103×10^{-6}

We then performed tests on the individual regression parameters (Table B.7) in Eq. B.1. The p-value of these tests were calculated using z-values. To test $H_0 : \beta_1 = 0$ we used $z = -3.26$ (p-value= 0.00112). To test $H_0 : \beta_2 = 0$ we used $z = -1.39$ (p-value= 0.165). Thus, in the wave I logistic regression models, containing out-degree and in-degree as regression parameters, only the out-degree had a significant impact on the probability of the given individual being depressed (using a significance level of 0.05).

For wave II, the complete model had an AIC of 2004.0, while the reduced model had an AIC of 2040.4. Thus, the model for wave II was:

$$\pi_i = \text{logit}^{-1}(-1.45 + (-0.0114 \times \text{out-degree}_i) + (-0.0511 \times \text{in-degree}_i) \tag{B.2}$$

We then performed tests on the out-degree regression parameter (Table B.8) in Eq. B.2. To test $H_0 : \beta_1 = 0$ we used $z = -4.63$ (p-value= 3.64×10^{-6}). To test $H_0 : \beta_2 = 0$ we used $z = -1.63$ (p-value= 0.103) Thus, in the wave II logistic regression models, the out-degree again had a significant impact on the probability of the given individual being depressed (as in wave I). This result implied the higher the out-degree for a given individual, the lower the probability of that individual having depressive symptoms.

Figure B.15 shows the quartiles and range of values for the 4 groupings. For respondents who were not depressed in wave I, compared to those that became depressed in wave II, those that stayed not depressed in wave II had greater median (4 vs 3) and upper quartile values (7 vs 6). For respondents who were depressed in wave I, compared to those that remained depressed in wave II, those that became not depressed in wave II had greater upper quartile values (6 vs 5) and equal median and lower quartile values. We hypothesised that respondents who became depressed had slightly fewer friends initially compared to those who did not, and that the respondents who recovered from depression started out with more friends compared to the respondents who did not recover. For respondents who were not depressed in wave I, the out-degree was likely to be higher for a respondent that remained not depressed in wave II compared to a respondent that had become depressed by wave II ($p = 0.0005$). For respondents who were depressed in wave I, the out-degree for a respondent that was no longer depressed in wave II was likely to be greater than the out-degree for a respondent that remained depressed ($p = 0.02$). These results provided some backing to our claim.

Appendix C. Jeffreys Interval

We wanted to allow for sampling error in the proportions calculated from our statistical sample. This was achieved using a binomial proportion confidence interval. For the recovery model and switch to depression model, by assuming we had a fixed number of trials, that the trials were statistically independent and the probability of changing emotional state was the same for each trial, we were able to use Jeffreys interval for our binomial proportion confidence intervals.

Introduced by Brown et al. [26], the Jeffreys interval has a Bayesian motivation. It gives a Bayesian credible interval obtained when for the binomial proportion p a non-informative Jeffreys prior is used, where

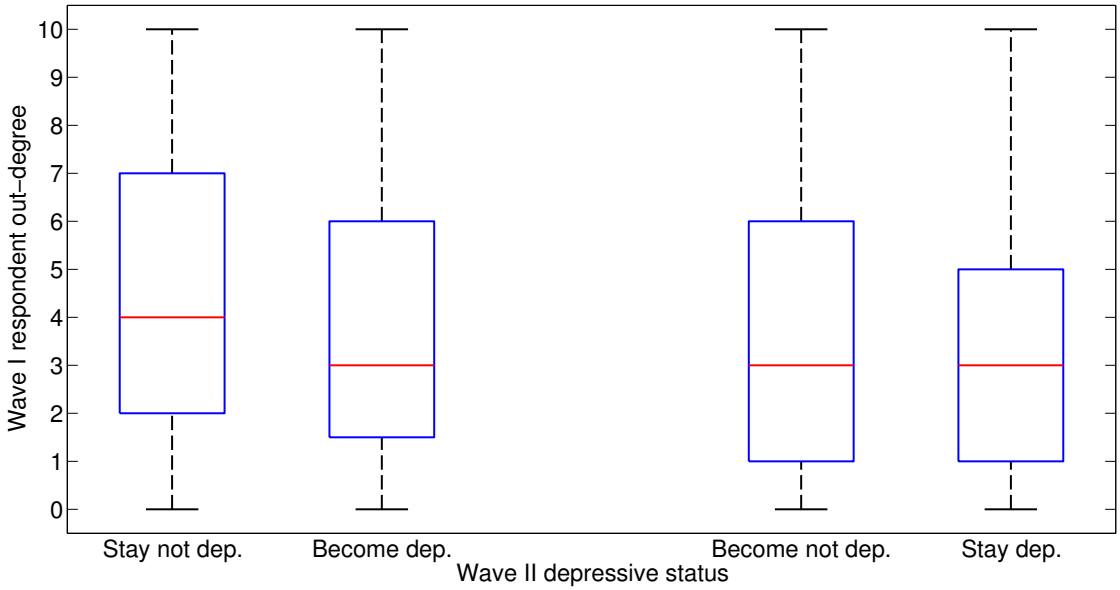

Fig. B.15. Box plot of the out-degree in wave I for the four groups. The two left hand box plots contain all respondents who were not depressed in wave I, while the two right hand box plots contain all respondents who had depressive symptoms in wave I.

the Jeffreys prior is a Beta distribution with parameters $(1/2, 1/2)$ [27]. After observing x successes in n trials, the posterior distribution for p is a Beta distribution with parameters $(x + 1/2, n - x + 1/2)$. In the case $x \neq 0$ and $x \neq n$, the Jeffreys interval is taken to be the $100(1 - \alpha)\%$ equal-tailed posterior probability interval as follows:

$$\begin{aligned} \text{Lower bound:} \quad L &= \mathcal{B}^{-1}\{\alpha/2, x + 1/2, n - x + 1/2\} \\ \text{Upper bound:} \quad U &= \mathcal{B}^{-1}\{1 - \alpha/2, x + 1/2, n - x + 1/2\} \end{aligned}$$

where $\mathcal{B}^{-1}\{x, s_1, s_2\}$ is the inverse cumulative distribution function of the beta distribution at the quantile x, with shape parameters s_1 and s_2.

Note that when $x = 0$ the upper limit was calculated as before but the lower limit was set to 0, and when $x = n$ the lower limit was calculated as before but the upper limit was set to 1.