TECHNICAL UNIVERSITY OF DENMARK

Theoretical and Computational Analysis of Dynamics on Complex Networks

MASTER THESIS FOR THE MSC. PHYSICS AND NANOTECHNOLOGY

Author:

Jana Sanne HUISMAN (**s142918**)

Supervisors:

Dr. Olivia WOOLLEY-MEZA (**ETH**)

Prof. Sune LEHMANN (**DTU**)

Prof. Dirk HELBING (**ETH**)

June 25, 2016

Abstract

Big data has allowed to study the communication and mobility patterns of humans with ever greater resolution. However, it is not yet clear how information from online social networks relates to offline face-to-face interactions. Knowledge of this directionality will help us to harness the increasing wealth of online information to improve predictions of for example offline disease spread. This study aimed to distill relevant information from online social networks to predict meaningful face-to-face contact behaviours. To this end data from the Copenhagen Network Study on the Facebook and face-to-face interactions of 850 students was used. First, a network of offline interactions was predicted using binary link prediction on features distilled from the Facebook interaction data. The network predictions of this model were then validated, using simulations of disease spread and comparison against the Erdös-Renyi random graph and configuration model network. It was found that stringent variables of offline contact, such as meeting during off-hours or meeting more than 5 times per week, could be predicted with 69% accuracy, which was 19% better than the Majority Vote Classifier. The target variable of meeting at least once a week could be predicted with 78% accuracy. The predicted network showed disease simulations that closely resembled those on the actual network, and performed significantly better than simulations on the Erdös-Renyi random graph. To the knowledge of the author, this is the first study to validate the quality of the network structure resulting from link prediction using disease simulations. It was shown that online network information can be used to predict offline contact networks which are useful for the investigation of the spread of disease. This study paves the way for future verification of disease models and development of intervention strategies using primarily online network information.

Acronyms

A legend to clarify the acronyms that are used often in this Thesis.

- **ACC**: Accuracy.
- **CNN**: The number of shared Facebook friends (Feature).
- **CNN_int**: The number of shared interaction friends (Feature).
- **Disc_int/comment/message_to/liked_story/message_tag/tagged_story**: The total number of exchanged interactions, 'comment', 'message_to', 'liked_story', 'message_tag', and 'tagged_story', while accounting for the activity of the users involved (Feature).
- **First**: The date of the first observed interaction between a pair (Feature).
- **FN**: False Negative.
- **FP**: False Positive.
- **Min/Max/Mean_Wait**: The minimal, maximal, and mean waiting time between two Facebook interactions (Feature).
- **RFC**: Random Forest Classifier.
- **Resp_rate**: The order of response (Feature).
- **Prev**: The total number of exchanged interactions in the previous month (Feature).
- **SIR Model**: Susceptible-Infected-Recovered Model.
- **TN**: True Negative.
- **Tot_int**: The total number of interactions (Feature).
- **TP**: True Positive.
- **TPR**: True Positive Rate, also called Recall.

Contents

1 Introduction and Literature Review

The spread of ideas, opinions, innovation, as well as behaviour, information, and disease is mediated by the multi-scale patterns of human interaction [1, 2]. Knowledge of human mobility networks at both global and urban scales has greatly improved predictions of the dynamics of disease and identification of the geographic origin of emergent diseases [3, 4]. However, this spread is further shaped by the social and communication networks humans are embedded in [5], which are much harder to quantify.

Measuring or modelling the exact contact network - given multiple spatial and temporal resolutions - remains a challenge [6]. Face-to-face interaction can only be measured using specialised hardware (e.g. sociometric badges using RFID transmissions) [7, 8] or smartphones with specialised software[1] [10]. The recent explosion of online social and information platforms has allowed for the direct observation of social interaction patterns. However, it is not clear how similar patterns of online interaction and face-to-face interaction are, nor how the two worlds relate and feedback into each other. Mobility information has been used to infer social ties [10, 11, 12, 13, 14]; however, there exist only few studies that investigate individuals' co-presence based on knowledge of their online interactions [12]. This direction of inference is particularly important if we are to harness the power of the digital world, where traces are naturally and clearly recorded, to address phenomena driven by offline interaction. More specifically, we are interested to distill the relevant information from online social networks to predict the face-to-face contacts that transmit disease.

Recent ubiquity of mobile computing and social sensors, combined with the exponential increase in computational power over the past fifty years, has greatly increased the capacity to collect and analyse data about all aspects of the lives and interactions of human beings [15]. Until quite recently, sociological data came primarily from questionnaires, census data, and anthropological field studies [10, 16]. Nowadays additional data sources such as travel card data, call detail records, emails, smartphone applications, and online social networks, are being used to track the movement, communication, daily rhythms, behaviour, and opinions of people [15, 17, 18, 19]. The availability of such data is predicted to increase further with developments on the Internet of Things, smart city initiatives, and increased digital access and participation [6, 20]. Under the name of computational social science [15], scientists are using this big data wealth to repose and revisit old questions in social science, as well as to take up entirely new ones. These data sources reduce the cost and organisational complexity needed to study very large cohorts of people, taking us closer to measuring micro-level behaviour and interactions on a societal scale. Furthermore, these measurements can be made at very high temporal resolutions and reduce the reliance on self-reported information, which is demanding to collect and can be prone to biases [11, 16].

Many of these sources provide overlapping but distinct information concerning the dynamics of the underlying networks of social relationships [11]. Each source or network reveals complimentary information about the underlying relationships between their users, and a more complete picture of human behaviour can be captured when we know how these different sources act together [16]. The value of interactions within the social network differs with the effort needed to initiate contact, and the culture of the network [21, 22]. In the language of physics, the relevance of network interactions is most naturally captured by different measures - e.g. duration, entropy, location - in different networks. If we are to compare or combine information from online and offline networks, we must understand how to map one network to the other.

This master thesis investigates (i) how the social network of Facebook maps onto the network of face-to-face social interactions, and (ii) how predictions based on online network information can be used to model the spread of disease through face-to-face networks. The analysis is based on data from the Copenhagen Networks Study, in which all interactions of approximately 850 students were recorded over the period of September 2013 to January 2016 [16]. To investigate these questions, supervised

[1] Recently, some work was undertaken to use Wifi signals to infer physical interactions and social ties [9].

link prediction based on binary classification [12, 23] is used to 'predict' offline interaction patterns, using features extracted from online interactions. The face-to-face network predicted through this procedure is then used to study the spread of disease in the offline population.

1.1 Social Interaction and Ties

When studying the dynamics of social networks, it is not the mere presence of a tie that matters, but rather how strong this tie is: not all relationships are created equal [22, 24, 25]. Weak ties are important in connecting distant nodes, and thus for disseminating information to far regions of the social network [26], whereas strong ties are important in collaborative problem solving [27]. Since tie strength affects both the speed and threshold at which connected individuals take up information, it is fundamental when modelling the dynamics of diffusion processes on networks [1].

There are many studies which attempt to predict the real-world strength of ties based on features from Facebook interaction between the pair of individuals. Gilbert and Karahalios defined 74 different interaction variables based upon sociology literature concerning tie strength, which they used to classify ties as strong or weak with 85 % accuracy using a linear regression model [24]. Jones et al. found that a logistic regression model using only the sum of all interactions between a pair already determines a user's closest friends with 82 % accuracy [28]. Both of these studies compared against a ground truth of self-reported friendship ties, which were collected using specially developed Facebook applications.

When dealing with anonymous large scale communication data - such as Call Detail Records [2] (CDR's) or online social networks - the subjective perception of a relationship is not directly observable [24]. The existence of an interaction in the observed communication network can be taken as evidence for the presence of a social tie. However, the strength of this tie then has to be inferred from e.g. the total amount, frequency, duration, or timing of interactions [10]. Variables that sociological studies have found to correlate with friendship can be used to assign an interpretation or characterise the relative importance of a tie [24]. A very important notion, which is often used, is that the similarity of individuals plays a strong role in the creation of social ties [12]. More similar individuals - in terms of demographic characteristics such as race, age, religion, education, occupation and gender, as well as in terms of their attitudes and interests, or simple geographic propinquity - are more likely to bond, and their ties dissolve less fast than between non-similar individuals. This effect is called homophily [30].

Furthermore, social networks are strongly influenced by spatial proximity [30, 31]. Using CDR data, Eagle et al. found that spatial and temporal context - specifically off-campus proximity in the evening and on weekends (which they coined as the 'extra-role factor') - is an important indicator of friendship [11]. However, Baratchi et al. argued that the 'on/off campus' variable is specific to social ties in one affiliation and does not translate to people with different spatial domains [14]. To identify different classes of social ties between people (n=5), they constructed two distinct variables: one to capture the amount of time two individuals spend in the same place, and one to capture the amount of places they spend time in together. The idea that the number of social contexts two nodes share - McPherson et al. called it the 'multiplexity' of a contact [30] - carries fundamental information about the importance or strength of a connection is also found in larger studies. Cho et al. found that users who visit similar places are more likely to be friends in online location-based social networks [13]. Furthermore, Wang et al. found that similarity between mobile phone users' movements (in terms of mobile homophily measures) correlates with their position in the social network (as measured by topological network connectivity measures[3]) and the intensity of their interaction (number of calls) [12].

[2] Call Detail Records are a telecom providers' records of phone calls and text messages, as well as information about the closest cell phone tower. This can be used as a proxy for geographical location and social interaction [17, 29].

[3] Specifically they used the Common Neighbors, Adamic-Adar, Jaccard Coefficient, and Katz measure.

These studies use social networks and mobility information to infer friendships. However, there exist few studies that attempt the inverse: to gain more insights about individuals whereabouts - e.g. predicting their location or colocation - based on knowledge of their online interactions and social ties. Although it has been shown that human trajectories show a high degree of temporal and spatial regularity [17] and are highly predictable on a society-wide scale, they are still erratic at the scale of the individual [29]. Predictability at an individual level increases when including information on the movement of friends or social circles [13, 32]. However, the most relevant predictor for offline interactions has not been identified so far. Therefore, the logical next step is to investigate exactly which information is most useful to predict offline interactions, which is the question that this thesis addresses.

1.2 Complex Networks

Thusfar, this literature review primarily considered studies which investigate ties between two people at a time, i.e. dyadic relationships[4]. However, if these pair-wise relationships are all taken into account at the same time, a network structure of social relationships emerges. Networks - also called graphs - are an often used way to describe many systems: ranging from the Internet, organisational networks, food webs, and distribution networks to social networks [33, 34].

In each of these cases, one defines a network as an ordered pair $G = (V, E)$, which consists of a set of items V, called vertices or nodes, and the set E of connections between them, called edges [35]. The degree k_v of a node v is defined as the number of edges that are incident to this node. Since this thesis concerns itself only with simple graphs, i.e. undirected graphs without multiple edges or loops, the node-degree k_v is equal to the number of other nodes v is connected to [36]. In social networks, the nodes represent people embedded in a social context, and edges represent e.g. interaction, collaboration, or influence between those people [37].

Interestingly enough, real-world networks all exhibit properties that deviate strongly from the simplest random model one would expect. Such a network would be the Erdös-Renyi random graph: in this undirected graph with n nodes, each of the $\frac{1}{2}n(n-1)$ possible edges are placed at random between the nodes with some probability p, and the node-degrees are distributed according to a binomial distribution [33]. However, human-made networks, such as information, transport, and city infrastructure networks show non-trivial topological features that lead us to call them complex networks [36, 38, 39]. Many of these topological features have been studied in recent time. Examples include strong heterogeneity in the degree distribution (e.g. scale-free networks, where the degree distribution follows a power law [40]), small-world properties (i.e. high clustering and a small diameter [41]) and interesting fine-grained structure such as communities [42] and motifs [43].

Typically, the social networks we study are static graphs that have been constructed through aggregation of all interactions in a communication system over a time window Δt [44]. However, the social networks that people are embedded in are not static, rather they grow and change as new interactions in the underlying social structure arise or desist [37]. The networks of who interacts with whom, i.e. contact networks, are even more dynamic and on shorter timescales: who people interact with changes many times during the day. Only recently it has become possible to observe the contact networks in real-time - using e.g. GPS to track movement of people with very high temporal resolution [45].

As scientists increasingly investigate the time dynamics of complex networks in general, it is found that these new methods are relevant for the study of social networks in particular [46, 47]. Including the possibility that links may appear or disappear in a network (i.e. allowing for temporal changes of the network structure) removes transitivity: if A is connected with B at time t_{AB} and B is connected with C at time t_{BC}, information can only pass from A to C if $t_{AB} < t_{BC}$ [47]. Furthermore, there exists evidence that the timescale of aggregation has a strong effect on the observed dynamics

[4] In sociological literature, pairs of interacting individuals are called dyads.

of the social networks themselves [32, 44, 48]. By aggregating to a coarse time-binning, many fast-moving processes - such as the dynamics of social gatherings - are averaged away, and fundamental information about the system is lost [32, 49]. Stopczynski et al. have also shown that knowledge of the dynamics at an hourly timescale is instrumental for studying spreading processes in society [45]. They found that the modelling of spreading processes is strongly impacted by a reduction in the temporal fidelity close proximity interaction networks are measured at. Temporal subsampling, both due to restriction to snapshots in time, or due to limited coverage of all social interactions, results in less frequent and smaller outbreaks and drastically increases the time it takes for the spreading process to reach 50 % of the network [45].

1.3 Disease Spread on Networks

Modelling the dynamics of disease is one of the most successful applications of the study of complex networks. It has allowed researchers to study the effect of different intervention and disease prevention techniques, in particular their effect on the number of individuals affected by an epidemic and the speed of transmission. Furthermore, it has aided in locating 'patient zero' based on later snapshots of the network, and allows to investigate the epidemic threshold of a system [50].

There exists strong evidence for the importance of the network structure for the dynamics of spreading processes [50, 51, 52]. Early epidemiological models assumed a well-mixed population, where each individual has a non-zero probability of contact with everyone else [53]. Within this population, individuals are compartmentalised into groups, reflecting the different stages of disease development [50]: e.g. Susceptible, Infected, and Recovered individuals (S, I, R respectively). Infected persons transmit the disease to susceptible individuals with the infection rate β, and recover with the recovery rate γ [54]. For such a well-mixed SIR-model, the spread of the disease is governed by three ordinary differential equations (for the population numbers of S, I, and R). The dynamics of the disease are critically determine by the basic reproduction number R_0 and will always support an epidemic if $\frac{\beta}{\gamma} = R_0 > 1$ [50, 53]. In reality however, individuals have only a finite group of contacts that they can spread a disease to (the 'mixing network'). This restriction to the contact network slows down the spread of infection [50], and shifts the epidemic threshold.

Spreading phenomena have been studied on a wide range of idealised networks, which has led to many insights in the difference between standard random mixing and disease spread through networks [50]. After assumptions about the micro-level mechanisms[5] of the infections are made, knowledge of the network structure allows one to simulate epidemic dynamics at the population scale [50]. The quality of the spreading simulation will be notably influenced by the quality of the network approximation.

1.4 Link Prediction

In this thesis, we formulate the problem of mapping one network to another as the task of predicting a link on the face-to-face contact network given the existence of interaction in the online network. Link prediction is the task of estimating the likelihood of the existence of a currently unobserved link between two nodes, based on the observed links and the attributes of the nodes [59]. This can be a static problem, where one e.g. attempts to estimate the links missing from an incompletely observed

[5]These micro-level mechanisms for the transmission of disease can be simple probabilistic models of transmission, given a contact between an infectious and susceptible individual. However, they can also involve threshold rules [55] or other non-linear relationships between the number of times an individual is exposed and the probability they become 'infected'. These complex contagion mechanisms may further vary depending on the item that is spreading (e.g. different information content or behaviour), and the host medium that the item spreads through [56]. Furthermore, critical behaviour can occur when two different disease are allowed to interact (co-infection). Recent analysis of a co-infection model [57] has revealed sudden avalanches when there is cooperative behaviour between pathogens [58], which may be theoretically similar to the abrupt dynamics of innovation.

biological network. Or it can be used for a prediction in time, where one predicts which links are most likely to form in the next instance of the network [37].

Link prediction techniques are widely used in machine learning and have found application especially in recommender systems [60], but also e.g. for the detection of anomalous email traffic or terrorist cells [61]. Furthermore, link prediction can be used to study which rules underlie the formation of new ties between entities, i.e. which processes drive network formation [37, 59, 60, 62].

To solve the link prediction problem, the majority of the literature relies on similarity measures to perform a binary classification over the set of all potential links [37, 59, 63]. The node-pairs can be classified in a supervised or unsupervised manner [59]. In their seminal paper on the link prediction problem for social networks, Liben-Nowell and Kleinberg proposed the unsupervised method [37]. In this case, the set of all potential links is ranked according to similarity between the nodes constituting the pair, computed using one of the available network properties. Then the k top-ranked potential links - where k is the expected number of new links - are classified as new links [12, 37]. However, more recently, supervised methods are gaining traction [12]. In this case, one learns a classifier (e.g. a decision tree) on a training set of new and missing links, and then classifies each further pair from the test set as a new or missing link according to the learned classifier [63]. Wang et al. found that the precision of the supervised classifier is about double of their unsupervised counterpart [12]. Lichtenwalter et al. also argue for the use of supervised learning, since unsupervised methods are domain specific, and it is hard to predict the performance on a different network instance [64]. Further advantages of supervised methods are that their variance can be reduced by placing them in an ensemble framework (see section 2.1.1), and they can deal with class imbalance (which unsupervised methods can not) [23, 64]. The main disadvantage is that these methods require labelled training data, which is often not given [23].

The similarity measures, which rank the links according to existence likelihood, can be based on both the similarity of the nodes or their proximity in the network. In case of node similarity, nodes are considered similar if they have many common features. However, in real-world networks these attributes are often hidden, so structural similarity is used: nodes are considered similar according to some network similarity index [59]. Liben-Nowell and Kleinberg and Zhou et al. systematically compared a number of similarity indices (local and global, node-dependent or path-dependent) on real networks and found that the Common Neighbours (CN) index performs very well [37, 65]. Here the CN similarity measure is defined as: $s_{xy} = (A^2)_{xy} =$\{the number of different paths with length 2 connecting x and y\}. The best choice of similarity index also depends on which aspects of the network structure are deemed most important for the tie formation [59]. Furthermore, when the nodes have very different types of attributes, there exist two ways to combine the similarity information. Either one constructs different similarity measures per feature, which are then combined using a form of linear regression. Or one constructs different networks - based upon a subset of features, or featuring edge weights that reflect the property under investigation - and uses a 'normal' similarity measure to determine similarity between links [66].

1.5 Evaluating the Quality of Link Prediction

Evaluating the quality of the link predictions is challenging, in particular because false prediction of the network structure can have extreme consequences for the dynamic processes on these networks. The proficiency of the machine learner at predicting the new links is typically assessed using receiver operating statistic (ROC) curves, and its area under the curve (AUC) [12, 59, 62]. However, when predicting links on social networks, there exists an imbalance between the amount of links that could potentially form with respect to those that do, and as a result the precision of the prediction is often quite low [37, 63]. When predicting new contacts based upon CDR data, Wang et al. used progressive sampling of missing links and a restriction of the link prediction to links with common neighbours to select a subset of potential links and make the prediction more manageable [12]. Their precision was only 30% when considering 51 million missing links, however this increased to

73.5% when restricting the prediction to nodes that shared common neighbours and had very high Adamic-Adar and spatial correlation scores [12]. Rather than changing the training method, Liben-Nowell and Kleinberg changed the baseline for representing their predictor quality when predicting ties in a co-authorship network [37]. They compare against random prediction, i.e. a classifier which randomly predicts k links from the set of all possible new links (where the mean accuracy is naturally the number of possible correct predictions divided by the total number of possible predictions), and found that almost every predictor (using some topological similarity measure) performed better than random [37].

A further downside of expressing the quality of the link prediction in terms of false positive and false negative predictions only is that this does not reflect the position of the falsely predicted node or edge within the network. As demonstrated in Fig. 1, the position of a mis-classified tie greatly influences the network properties (such as connectedness) of the resulting predicted network. Unlike conventional applications of classification - e.g. finding terrorists or spam [59] - we are interested in the structure of the resulting network rather than the prediction of a single tie. Thus, new methods are needed to quantify prediction performance, i.e. the severity of the classification errors for the network prediction and possible systematic misclassifications based on network properties.

We take a novel approach and verify the quality of our machine learner by studying the suitability of its output - the predicted contact network - as input for our specific goal, namely the modelling of spreading dynamics on networks. This investigation paves the way for future verification of disease models and development of intervention strategies using primarily online network information.

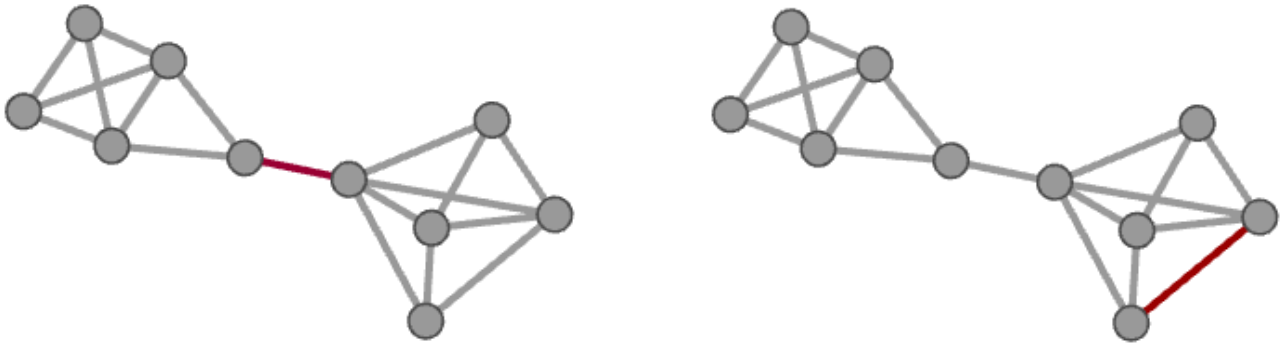


Figure 1: The structure-dependent impact of a false positive classification in a network. On the left, the two nodes share a falsely predicted tie (coloured red), which links two remote regions of the network. However, on the right the falsely predicted tie lies in a densely connected part of the network. Both predicted networks will exhibit entirely different spreading behaviour. The illustration was made using Gephi [67].

1.6 Scope of this Thesis

To summarise, this thesis investigates the relation between online interactions and offline meetings and studies how well we can use online networks, where data is easily available, to predict the outcome of offline spreading processes. In particular, we investigate the possibility to detect characteristic Facebook interaction behaviour that qualitatively separates user-pairs with different offline interaction behaviours. We aim to determine which characteristics of Facebook interaction are informative about the tie, and which aspects of the interaction do not matter for further link predictions. Secondly, we study the usability of such predictions for the investigation of disease spreading.

The structure is as follows: after the relevant methods have been introduced in chapter 2, chapter 3 introduces the dataset, and some descriptive analysis performed on this data. Moving from this, chapter 4 presents the work to relate features of the Facebook interaction data to offline behaviours. This chapter starts with a brief introduction, followed by section 4.1 which introduces the features

that were extracted from the Facebook data, and section 4.2 which presents the results of classifying offline behaviour based on these online interactions. These results are discussed in section 4.3. In chapter 5 we then use the resulting network of predictions as input for dynamic simulations of disease spread. The simulations on the predicted and actual offline network are compared against each other and two null models in section 5.2. This section also discussed the use of the predicted network for modelling the spread of disease in the offline population, as well as the most important network characteristics that drive the disease simulations. Lastly, the overall implications of the work are discussed in chapter 6, which also concludes this thesis with an outlook for further work.

2 Methods

This chapter covers the different methods used in this thesis. First, section 2.1 introduces Classification, and in particular section 2.1.1 focusses on the Random Forest Classifier. Next, section 2.2 discusses the process of training a classifier in section 2.2.1, and different metrics to quantify classifier performance in section 2.2.2. Building on this, 2.2.3 discusses the influence class imbalance has on these results, and section 2.2.4 looks at model comparison. The sections on methods are concluded with section 2.3 which introduces the SIR disease model and its simulation with the Gillespie algorithm.

2.1 Classification

Classification is a form of supervised machine learning, which requires training data that has been assigned a discrete class label [68]. The data consists of observations, in this case user-pairs, with a number of features, also called variables or attributes, that span a multi-dimensional 'feature space'. Each observation is assumed to belong to one of a number of discrete classes. In particular the investigations here are restricted to binary classification, i.e. there are only two possible class-labels: 0 or 1 [23]. Later, the class assignment will also be called the 'target variable', since it is the target the classifier is trained to predict. Given the training data, the classifier learns a function - which depends on the data features - to assign the correct class labels to observations it has not seen before. The learned classifier function is then applied to data from a labelled test set, and the quality of the classifier is determined by comparing these predictions against the known labels. This step is essential, because the key objective of the learning function is a good generalisation to previously unknown records. The resulting classifier function can be used both as a tool to distinguish between objects of different classes, or to predict class membership.

2.1.1 Random Forest Classifier

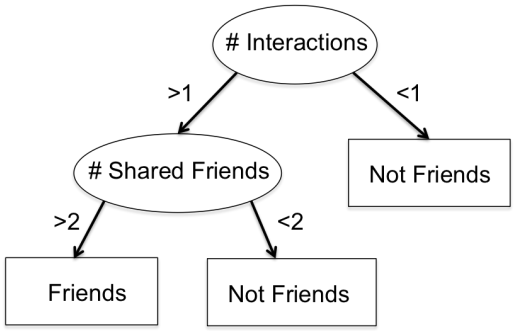


Figure 2: A decision tree. At each node a feature is chosen to split the data into successively purer subsets. This process is continued until the leaves contain observations of only one class. The illustrated tree classifies whether a pair are friends or not, based upon the number of interactions and the number of shared friends respectively.

Many classifier methods exist, and the optimal choice is often problem specific - depending a.o. on the number of observations, features, and noise in the data [68]. A simple yet effective classifier method is Decision Trees. This method works somewhat like the game 'Twenty Questions': through a series of questions about features of the data, it splits the dataset up into progressively purer subsets until it can uniquely assign a label to every set (see Figure 2) [23]. Selecting the optimal series of questions is non-trivial. Often Hunt's algorithm is used, which makes a series of local optimum decisions: at every split it chooses the feature which maximises the gain function, until the

remaining subset has only one class label. This gain function Δ is given by [23]:

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j) \tag{2.1}$$

where $I(\cdot)$ is the impurity measure of a given node. N is the total number of observations at the parent node, k is the number of attribute values, and $N(v_j)$ is the number of observations associated with the child node v_j. The impurity measure, also called split criterion, that we use here is the Gini criterion[6]:

$$I(t) = \text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \tag{2.2}$$

where $p(i|t)$ denotes the fraction of data points belonging to class i at node t [23]. In this basic form, feature selection based on impurity reduction is biased towards variables with more categories. Decision tree algorithms such as Classification And Regression Trees (CART) mitigate this effect by restricting the test condition to binary splits only [23]. The scikit-learn package that was used for machine learning algorithms in this thesis, uses an optimised CART [69]. The main advantage of Decision Trees is that it is a very fast algorithm, both in training and classifying new test records, which can be made more accurate by placing it into an ensemble framework.

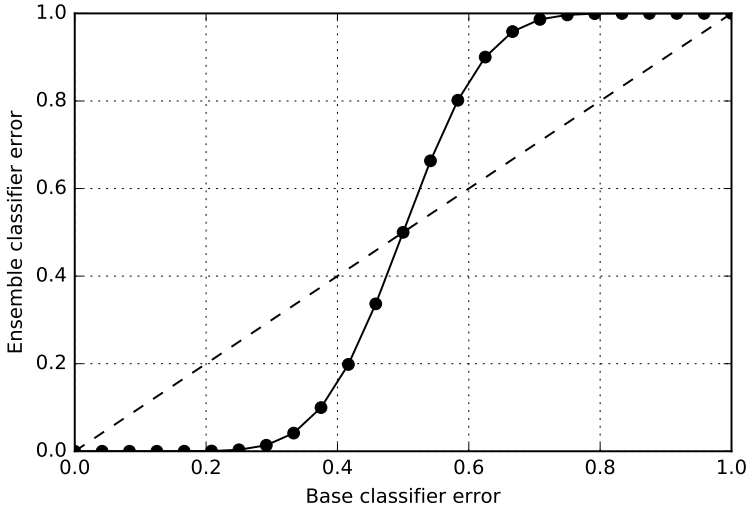


Figure 3: The ensemble error as a function of the base classifiers error, for 25 completely correlated (dashed line) or uncorrelated base classifiers (solid line). Combining perfectly uncorrelated classifiers greatly improves the ensemble's classification error, as long as the error of the base classifiers is below 0.5. Figure adapted from [23].

Ensemble methods build multiple base classifiers using the training data, and then aggregate the predictions of these base classifiers to the final classification [70]. This aggregation is done through simple majority voting. The prediction will be incorrect if more than half of the base classifiers make a wrong prediction. If the base classifiers are independent, and all have the same error rate ϵ_{base} we can write the error rate of the ensemble as:

$$\epsilon_{ensemble} = \sum_{i=\lceil N/2 \rceil}^{N} \binom{N}{i} \epsilon_{base}^{i} (1 - \epsilon_{base})^{N-i} \tag{2.3}$$

Figure 3 shows the ensemble error as a function of the base classifiers error, for 25 completely correlated or uncorrelated base classifiers. For correlated classifiers the ensemble result will not deteriorate with respect to that of the base classifier, however combining perfectly uncorrelated classifiers greatly improves the results as long as the error of the base classifiers is below 0.5 [23].

Nonetheless, it is often hard to obtain perfectly uncorrelated base classifiers. By definition robust methods will converge to a similar prediction with the same input data. Therefore, the ensemble of base classifiers is generated using one or several kinds of randomisation: by changing the training set, the features that are considered, or the algorithm itself. The training set is varied by resampling the original training data according to some sampling distribution: when sampled with replacement from an invariant sampling distribution this is called bootstrap aggregating (bagging), whereas algorithms that adapt the sampling distribution to observations that are hard to classify are called boosting[7] [70].

The classification error is typically a combination of three effects: Bias + Variance + Noise [23]. This is called the bias-variance decomposition. In case of bagging, the ensemble classifiers typically have smaller variance than the constituent classifiers, i.e. they reduce overfitting on noisy data. In case of boosting, the ensemble typically has a good effect on the bias of the classifier. It can sometimes be more accurate than bagging, but also tends to overfit.

Random Forest is a class of ensemble methods, based on Decision Trees as base classifier [72]. The method was developed by Leo Breiman and Adele Cutler in 1996. It uses bootstrap aggregating and the random subspace method to grow less correlated Decision Trees. Per split of the tree, the random subspace method - also called attribute bagging - restricts the choice of an optimal feature for splitting to a subspace of the original feature space. This randomisation helps to reduce the correlation among the decision trees so that the generalisation error of the ensemble can be improved. Compared to individual Decision Trees, Random Forests have a much lower variance (i.e. they are less prone to overfitting). Using the Random Forest Classifier, we can also analyse which features carry the most weight in learning the correct classification function. Feature importance can be computed as the (normalised) reduction of the split criterion brought by that feature [69], or - similarly - by assigning higher importance to those features that were used in earlier splits of the trees making up the Random Forest.

2.2 Measuring Classifier Performance

2.2.1 Cross-Validation

No statistical model or machine learning method is useful without a way to measure it's performance, i.e. to quantify the estimation or prediction error, and to compare the results against other models. Therefore, it is important to get an estimate of the classification error when using the classifier on previously unseen data, which is also called the generalisation error.

This can be done by, before training the model, splitting the available data into two independent parts, to allow for both training and testing. To improve the estimation of the generalisation error, a particularly well-established technique is to use k-fold cross-validation[8] [59, 23]: here the available data is split into k independent test sets. For each cross-validation fold, one of the test sets is used, and the other $k-1$ sets are used as training set. The generalisation error is calculated for each of the test sets, and then averaged [23].

Furthermore, cross-validation can be used for model selection. If the machine learning method contains parameters that have to be tuned for an optimal performance, the training set should be split into training and validation sets. First an inner cross-validation is performed, whereby the method is trained simultaneously for several different parameters in each fold and test errors are calculated on the validation sets. Hereupon the best model is selected based on the validation error, and the outer cross-validation is used to find the generalisation error for this model (see Figure 4).

[7] The most well known and widely used example of a boosting algorithm is called AdaBoost and was developed by Freund and Schapire in 1995 [71].

[8] The importance of cross-validation in data mining and machine learning is so big that the statistics and machine learning section of the popular site stack overflow has been renamed 'cross-validated' recently.

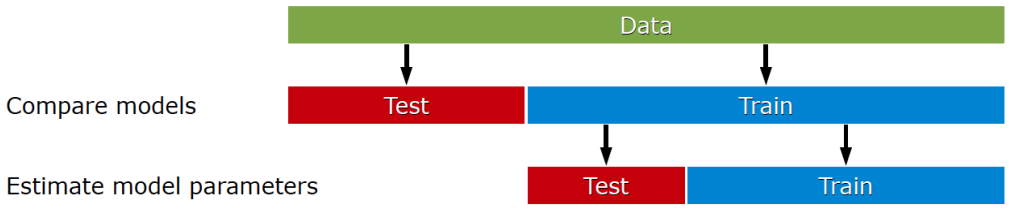


Figure 4: Test, training, and validation set. To illustrate the way the available data is split first into a test and training set, and the training set is then further split into training and validation sets. Image from the course '02450 Introduction to machine learning and data mining' at DTU, Fall 2015.

2.2.2 Performance Metrics

For a binary classifier, we have two types of errors: misclassifying the 0- and 1-class respectively. Since 0 is also associated with a negative, and 1 with a positive outcome, these are called False Positives (FP) and False Negatives (FN) respectively, and can be related to Type I and Type II errors from statistical hypothesis testing. An overview of these error types can be found in Table 1.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 1: The possible outcomes of binary classification.

The quality of the binary classifier, or most other machine learning and statistical methods, can be evaluated using different metrics, depending on which class is of interest and whether type I or type II errors are deemed more important. Each of these metrics focuses on a slightly different aspect of the classification. They also allow one to compare the performance of models that were trained with a different number of features, different parameters, or different classifier methods, and thus decide which model is most useful for the task at hand.

The accuracy (ACC) of a classifier denotes the fraction of total decisions N that were accurately predicted:

$$ACC = \frac{TP + TN}{N} \tag{2.4}$$

Precision, also called positive predictive value (PPV), is the fraction of observations labeled positive that the machine learning method classified correctly:

$$PPV = \frac{TP}{TP + FP} \tag{2.5}$$

However, this metric does not account for the sensitivity of the classifier to the positive class. That is why it is often combined with the Recall metric, also called the Sensitivity or True Positive Rate (TPR), which records the fraction of Positives that the method correctly classified as positives:

$$TPR = \frac{TP}{TP + FN} \tag{2.6}$$

Equivalently we can define the False Positive Rate (FPR), as the fraction of Negatives that the method incorrectly classified as positives:

$$FPR = \frac{FP}{TN + FP} \tag{2.7}$$

The relevance of the metric depends on the dataset: in a highly imbalanced dataset a classifier may achieve a very high accuracy by labelling all observations according to the majority class in the training sample. However, all minority class points then result in either False Negatives (if the 0-class has the majority) or False Positives (if the 1-class has the majority), which can be detected by looking at e.g. the FPR. Further, the Precision and Recall do not depend on the absolute number of samples in the negative class. Thus, it can be relevant to look at the Precision-Recall curve especially when we have a large number of true negatives. The precision and recall are summarised together into the $F1$-score, which is actually a weighted average of the two.

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2.8}$$

A useful measure that is not affected by class imbalance is the Matthew's correlation coefficient: it takes into account all four classes of predictions, and calculates a correlation coefficient value between the predicted and actual class labels (+1 denotes a perfect prediction, 0 an average random prediction and -1 an inverse prediction).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.9}$$

The final choice of metric will also depend on the application: In the scope of this thesis, where we use Facebook interactions to predict who meets offline, several different cost decisions could be made. If we wish to implement an effective vaccination scheme False Negatives carry the largest cost. As such, we wish to detect all people that meet offline, in exchange for detecting a few false positives. Then Recall is the metric of choice. However, when using the classifier to predict friendships in order to recommend specific events or venues, it may be more costly to have a high number of False Positives (assuming the act of recommending costs something). Then it is more important that we label pairs as 'meeting' exclusively when they do, and thus we would be more interested in the Precision of the classifier.

So far, we have assumed that the classifier makes a clear binary decision of the class label. However, typically a classifier will not assign a class, but rather the probability of belonging to the positive class. It then depends on the value of the threshold θ above which probability an observation is assigned this class. With the value of θ the number of true and false negatives and positives will change, and the performance metrics will vary accordingly. Which threshold to choose is thus another design parameter that can be varied to put more emphasis on correctly predicting the positive or negative class respectively.

A method often used to depict the tradeoffs between benefits (true positives) and costs (false positives) is the Receiver Operating Characteristic (ROC) curve [73]. The ROC plots the number of positives included in the sample as percentage of the total number of positives (TPR), against the number of negatives in the sample as proportion of the total number of negatives (FPR) [68]. This curve is insensitive to class imbalance.

When comparing classifiers, the ROC is often reduced to a single scalar value that summarises the expected performance over all thresholds. This value is typically the area under the ROC curve, abbreviated to AUC ROC. It has a few nice properties, foremost that it can be equated with the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [73]. A random classifier will follow a straight line $TPR(\theta) = FPR(\theta)$ as ROC curve, and thus the AUC ROC will be 0.5 for such a classifier.

The ROC curve and ROC AUC are often used, especially since they allow for evaluation of model performance even in highly imbalanced datasets, since they are not affected by the absolute number of positive or negative samples. Nonetheless, some criticism has been voiced in recent years. For example Fawcett warns against comparing ROC curves between different classifiers, as the curves reflect a relative ranking based on the internal workings of the classifier rather than true probabilities which could be compared at a common threshold [73].

2.2.3 Class Imbalance

Restrictive criteria for the positive class, such as meeting on the weekends, or belonging to a specific sociodemographic group, greatly reduce the relative size of this class with respect to the zero target. Thus, when considering more specific target variables, sample imbalance quickly becomes an increasingly pressing issue.

In a classification context, learning from imbalanced data means that the training set contains an imbalance in the number of samples from each class. The strength of the imbalance has a direct effect on the error rate of the minority class classification, even for minor imbalances [74]. The class imbalance affects the result of training a classifier in several ways.

First of all, there are problems both with relative and absolute rarity of the minority class. A classifier trained on highly imbalanced data often does not have enough information on minority class examples to draw the decision boundary correctly. This is only a problem if we have good reason to assume that in the true population the prevalence of the minority class is much higher. A widely-used method to deal with this imbalance is to oversample the minority class, or under-sample the majority to achieve a more balanced training set. However, the latter removes a lot of relevant information from the training set, and the former may cause the classifier to overfit for the minority class data points. Either way, the classifier is trained for an artificial distribution, based upon the assumption that all classes are equally common. This may introduce strong bias into the model, which is not reflected by new observations.

Secondly, some performance metrics are insufficient descriptors for imbalanced datasets. The prediction accuracy is regarded a very bad metric in a class imbalanced case, since a classifier that will always predict the majority class will have a very high accuracy. Thus, other metrics are needed to describe the quality of the classifier in an imbalanced case: typically the (AUC) ROC is used.

2.2.4 Model Comparison

To make statements about the goodness of a classifier, and to guide progressive model building, it is important to compare the performance of different classifiers. We need statistical tests to tell whether the difference in performance metrics is significant. In particular we can compare models by testing if their error rates were drawn from the same distribution. If they are, we can not claim that one model is better than the other.

Firstly, when comparing the performance of two models that are based on the same input data and target assignment, we can use the paired t-test. In each cross-validation fold, the performance of both models on the test data is saved as the value x_i or y_i respectively. Together the results on all k folds represent a sample of the underlying distribution of the test error of this model on the data. As such the series of performance information over all folds, $\{x\}_k$ and $\{y\}_k$, can be compared with a paired t-test to see if one differs significantly from the other [68].

One particular use of this technique is used to tell whether a model is significantly better than random. In this case the classifier in question, such as the Random Forest Classifier, is compared against the Majority Vote Classifier (also called zero rate classifier), which will classify all observations according to the majority class in the training data. This is a particularly useful baseline when class imbalance skews the performance statistics, e.g. when a high true positive rate is mostly driven by the large number of samples of the positive class rather than the classifier's predictive power.

Secondly, when comparing models that have been trained on different data, we must use the unpaired t-test. An example of this case will be introduced later when we compare the results of training a classifier on all Facebook data against the classifiers for single months.

2.2.5 Confidence Bands for the Accuracy

It is possible to estimate a confidence interval for the accuracy of a classifier, or equivalently for its generalisation error, depending on the number of instances in the test set.

Suppose that p is the true accuracy of the classifier. For each sample in the test set we have a chance p of success, and $p-1$ of misclassification. Thus the observation of $X = TP + TN$ successes in a test set of size N is a Binomial random variable with mean Np and variance $Np(1-p)$. Then the observed accuracy $f = X/N$ will also be a Binomial random variable, with mean p and variance $\frac{p(1-p)}{N}$. We can use this to estimate confidence bands for the true accuracy [23, 68]:

$$p = \frac{N \cdot \left(f + \frac{z^2}{2N} \pm z \cdot \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}} \right)}{N + z^2} \tag{2.10}$$

where z is the confidence limit, i.e. the chance that X lies more than $z = 1.65$ standard deviations above the mean corresponds to 90% confidence bands.

However, since we use cross-validation to get a better estimate of our generalisation error, we should combine these confidence estimates for the k folds. We can use the property of the binomial distribution, that if $X \sim B(n,p)$ and $Y \sim B(n,p)$ are two independent binomial variables with the same probability p, their sum $Z = X + Y$ will also be binomially distributed, with $Z \sim B(n+m,p)$ [75]. In this case, if we look at all successes over all $k = 10$ test sets, i.e. $Z = \sum_{i=1}^{10} X_i$, we can define an accuracy for the combined predictions $f_{tot} = \frac{Z}{N_{tot}}$. In this formulation equation 2.10 can be used again to calculate the confidence bands for the total accuracy.

2.3 Modelling the Spread of Disease

2.3.1 SIR Model

Disease models are typically based on a compartmentalisation of a population of individuals into different groups, reflecting the stages of disease development [50]. Here we investigate the SIR model, which contains three compartments: Susceptible, Infected, and Recovered individuals, with population numbers denoted with $S(t)$, $I(t)$, $R(t)$ respectively. The total population is fixed to $N = S(t) + I(t) + R(t)$. Infected individuals can transmit the disease with the infection rate β, and recover with the recovery rate γ [54]. These are transmission rates per node. Under the assumption that an individual is equally likely to transmit the disease to all others it is connected to, the transmission rate per link becomes $\beta_l = \beta/k$, where k is the degree of the node.

When the number of contacts in a well-mixed population is assumed independent of the population size (i.e. in the frequency dependent case), the force of infection λ - the per capita rate of infection - is $\lambda = \beta \cdot I(t)/N$. The rate at which susceptible individuals get infected is then $dS/dt = (\beta \cdot I(t) \cdot S(t))/N$. Thus, the disease evolves according to the following differential equations:

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \tag{2.11}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I \tag{2.12}$$

$$\frac{dR}{dt} = \gamma I \tag{2.13}$$

These rates change when looking at networks. In a completely random network, we can make a mean-field approximation, which yields equations very similar to those of a well-mixed population (equations 2.11, 2.12, 2.13): only the transmission rate has to be replaced by a transmission per link

instead of per node. Let $\langle k \rangle$ be the average degree of a network node. Under the assumption that the degree distribution is not too skewed the average degree is a good approximation of the actual degree of a node, and we can replace β by $\beta_l = \langle k \rangle \cdot \beta$ in the equations above.

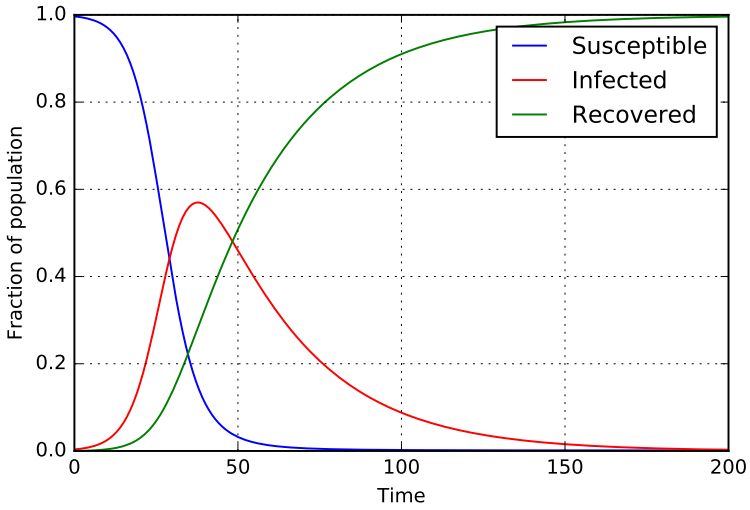


Figure 5: The infection curve for the mean-field approximation. Here $\gamma = 0.035$ and $\beta = 0.05$.

The SIR model will always support an epidemic if $\frac{\beta}{\gamma} = R_0 > 1$. The constant R_0 is called the basic reproduction number, and depends on the type of disease and the host population. In the case of a well-mixed system an epidemic always results in an infection of all connected individuals. However, on networks it becomes interesting to study how quickly the epidemic spreads to a significant fraction of the largest connected component[9], and which fraction of the connected individuals the infection reaches before being trapped by a boundary of recovered individuals. This model excludes demographic processes such as birth and death, which are negligible given the timescales of our study. If one were to allow waning dynamics and re-infection, e.g. in an SIRS model, it becomes interesting to determine the endemic level of infection [54].

2.3.2 Gillespie Algorithm

An efficient and stochastically exact method for simulating the SIR model is using the Gillespie algorithm [54]. This method was originally developed as the stochastic simulation algorithm (SSA) to simulate the chemical master equation (CME) [76] [10]. A master equation describes the time evolution of a system with discrete states, which switches between these states probabilistically. In particular the equation governs the time evolution of the probabilities of being in a given state. The probability of switching between states, e.g. the rate of a chemical reaction, changes depending on the state vector of the system.

In SSA, instead of developing the entire probability density over time, a numerical realisation of the system trajectory is constructed. By averaging over many runs of the algorithm, the probability density can then be approximated.

The total reaction rate here is:

$$a_0(\mathbf{x}) = \sum_{j=1}^{M} a_j(\mathbf{x}) \tag{2.14}$$

[9] A connected component is a subsection of the graph in which any two nodes A and B are connected by a path, i.e. one can reach B from A by moving from node to node using only the edges that connect them, and which is connected to no other nodes in the graph [35].

[10] The Gillespie algorithm is also equated with Kinetic or Dynamic Monte Carlo.

where:

$$a_1 = \beta_l \cdot |\{\text{SI-pairs}\}| \tag{2.15}$$

$$a_2 = \gamma \cdot |\{\text{Infected individuals}\}| \tag{2.16}$$

The algorithm consists of the following steps [76]:

0. Initialise the system at time t_0 and state x_0.
   This consists of initialising the network to run the simulation on (either by generating an idealised network, or loading the data of the predicted or actual network), and randomly selecting an initial set of infected individuals (to a total of $I(0)$ individuals) [11]. The number of recovered individuals is assumed $R(0) = 0$, so $S(0) = N - I(0)$. The set $\{\text{SI-pairs}\}$ is initialised corresponding to the selection of infected individuals.

1. Given the system in state **x** at time t, evaluate all reaction rates $a_j(\mathbf{x})$.
   The infection and recovery rate depend on the number of SI-pairs and the number of infected individuals as stated in eq. 2.15 and 2.16.

2. Generate the values for the next time-step τ and reaction type j.
   The time and reaction instances of the next reaction are generated using:

$$\tau = \frac{1}{a_0(\mathbf{x})} \ln\left(\frac{1}{r_1}\right) \tag{2.17}$$

$$j = \min_i \text{ where } \sum_{j=1}^{i} a_j(\mathbf{x}) > r_2 a_0(\mathbf{x}) \tag{2.18}$$

   where $r_{1/2}$ are random numbers drawn from the uniform distribution on the unit interval. Alternatively, τ can also be drawn from an exponential distribution with rate $a_0(\mathbf{x})$ directly.

3. Update the state vector and time according to that reaction (from 2).
   In case of infection (i.e. $j = 1$), a random susceptible neighbour of an infected person is also infected. They are removed from the set of susceptible individuals and SI-pairs, and added to the set of infected individuals. Their susceptible neighbours are added to the SI-pairs.
   In case of recovery (i.e. $j = 2$), a random infected individual recovers. They are removed from the set of infected individuals and added to the set of recovered. Their susceptible neighbours which no longer have a connection to an infected individual are removed from the set of SI-pairs.

4. Record $(\mathbf{x}, t)$. Return to 1, or end the simulation.

This simulation of the SIR model is then repeated a number of times (n_{runs}). Since the resulting time vectors t will feature different time-steps, they are binned to unit intervals. For each interval, **x** assumes the last state we recorded before this point in time.

[11] In the current implementation, the initial condition (which person in the network is infected) is varied in every simulation. However, for other research questions, such as finding the most influential person in a network, this would be precisely the parameter of interest.

3 Descriptive Analysis of the Dataset

This thesis is based upon data from the Copenhagen Network Study, a cohort study which collected longitudinal data of approximately 850 individuals over the period of September 2013 to January 2016 [16]. The goal of the study is to investigate human interaction and social networks across multiple communication channels. The individuals in questions are a densely connected population of undergraduate students at the Technical University of Denmark (DTU). These students were given Android smartphones [12], which were used as sensors to collect information about their face-to-face interactions, telecommunication, and location. This data is further supplemented with information from Facebook, questionnaires - with questions concerning e.g. personality and health -, demographics and other background information [16].

The study was deployed in two rounds: a pilot which started in September 2012 with 200 phones, and the main study with 1000 phones, which commenced in 2013. The two deployments differed in the manner they attracted participants, as well as the time at which the phones were handed out. Here, we will concern ourselves only with data from September 2013 onwards. For this deployment the focus was to engage students as early as possible: 300 phones were handed out to freshmen before the start of the Fall semester and 200 in the first few weeks; after which undergraduates from older years were also invited to participate [16]. This culminates in a maximum number of users around February 2014 (867 unique Facebook accounts, and 661 participants with at least one Bluetooth observation). However, there was no point at which the 1000 phones corresponded to the same number of users, since many phones were lost or got broken [9].

The data is collected, bundled and managed using an open Personal Data System [16]. All interactions between participants (phones) and the data platform relies on the use of anonymised 'tokens', and in the final dataset participants are distinguished using pseudonym identifiers. The study was approved by the danish data protection agency, and complies with EU and local rules. Participants were given access to same API as used by the researchers, so that they could see which data was collected about them, and had the opportunity to change their privacy settings. Surpassing this, the study data has been used as a means to investigate privacy implications of multimodal data collection [77].

3.1 Types of Data Collected

3.1.1 Bluetooth Data

Bluetooth is a wireless technology used for short-range communication - 5 to 10 meters - between mobile devices, which can be used as proxy for face-to-face interactions when the devices of other participants are detected [16]. The use of Bluetooth as proxy for face-to-face interactions was pioneered by Eagle and Pentland in their Reality Mining experiment [10]. It has the advantage of highly time-resolved and fine-grained data collection, without reliance on user actions.

In the Copenhagen Network Study, the individual Bluetooth scans are saved in the form (i, j, t, σ) when device i has observed device j at time t with signal strength σ. Every five minutes - measured from the last time the phone was powered on, the participant's phones collected information on all devices found in the vicinity (all signals within a 10 m range) [16]. To account for the desynchronised scanning, Sekara and Lehmann bin the Bluetooth scans into fixed-length time windows [78]. Furthermore, since Bluetooth scans produce very few false positives, they assume the resulting adjacency matrix $W_{\Delta t}$ to be symmetric. Nonetheless, this method is not perfect and a better proxy for face-to-face interactions is obtained when thresholding the signal strength at a received signal

[12] LG Nexus 4 for the 2013 deployment.

strength indicator (RSSI) value greater than -80dB to obtain only interactions at < 1 meter distance [78] [13].

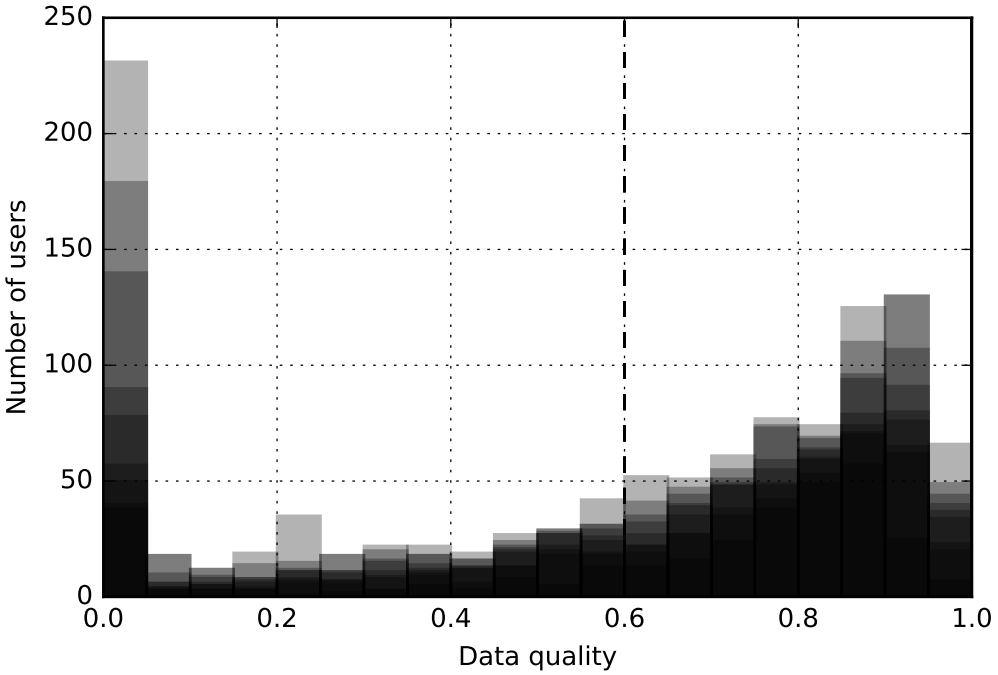


Figure 6: Bluetooth Data Quality. Overlaid histograms of the data quality per user, i.e. number of observations per month divided by the maximum number of observations, for all 9 considered months. The histogram is sufficiently two-lobed to suggest a thresholding at 60% (dashed line).

Although the software installed on the participant's phones was meant to counteract Android's automatic deactivation of the Bluetooth scanning, there are still a number of users with very few Bluetooth observations. In their papers on the spreading of disease on proximity networks, using the Bluetooth proximity data of the Copenhagen Network Study, Stopczynski et al. [80], and later Mones et al. [79], only consider users who exhibit a Bluetooth scan in at least 60% percent of the time bins in a given month.

Per month with n_{days} days, there are $b_{max} = n_{days} \cdot 24 \cdot 12$ possible 5-minute timebins. This is the maximum number of bins that each user should have observations for. An observation is recorded as long as the participant's phone was on and the Bluetooth is working, i.e. there are many observations with no other devices nearby and $RSSI = 0$. The data quality is then defined as the number of observations b_{user} a user has relative to b_{max}. Since the Bluetooth signals are deemed symmetric, we could further include bins where a user is observed by others in the vicinity. However, this will introduce a bias where users with low data-quality who are in a study-line with many other study participants are not excluded, whereas users with less contact with other study participants are.

In this thesis, users with less than 60% data quality (see section 4.2.1) were excluded, and thereafter only Bluetooth interactions between study participants at more than -80 dB are taken into consideration. Although Stopczynski et al., and Mones et al. find that this thresholding results in roughly 80% of the users in the month of February 2014 [80, 79], the percentage I find is much lower for this and the other months [14]. The discrepancy is most likely a result of the different thresholding of interactions, however my values are in exact correspondence with a more recent work by Stopczynski et al. where they select 476 participant in February 2014 [45]. Furthermore, the percentage of retained pairs scales roughly quadratically: if N is the number of users and $N \to 0.8 \cdot N$ then the

[13] Mones et al. even use a threshold of -75dB [79].

[14] Percentage of users kept. Sep: 41%, Oct: 57%, Nov: 58%, Dec: 74%, Jan: 72%, Feb: 67%, Mar: 70%, Apr: 66%, May: 68%

maximum number of pairs goes roughly $N(N-1) \rightarrow 0.64 \cdot N(N-1)$. Here the retention is lowest in September, where only 17.6 % of the pairs are kept, and reaches a maximum of 47.5 % in December, on average it is roughly 35 % [15]. This means that the number of observations that feature less common Facebook interactions, such as tagging each other in a message, falls very low in some months (most notably in January).

Since the users and all their interaction pairs are removed based only on the Bluetooth quality, which is independent of the Facebook interaction, this does not introduce bias into our sample. However, it does increase the reliability of our comparison between Facebook and Face-to-face interactions, since it decreases the amount of noise in the latter. In particular, it improves the classification results, since it removes user-pairs with potentially ambiguous or noisy classification. Otherwise user-pairs may be assigned the 0-class, i.e. non-interacting, because of lacking data rather than lacking interaction. In our dataset this greatly reduces the class imbalance.

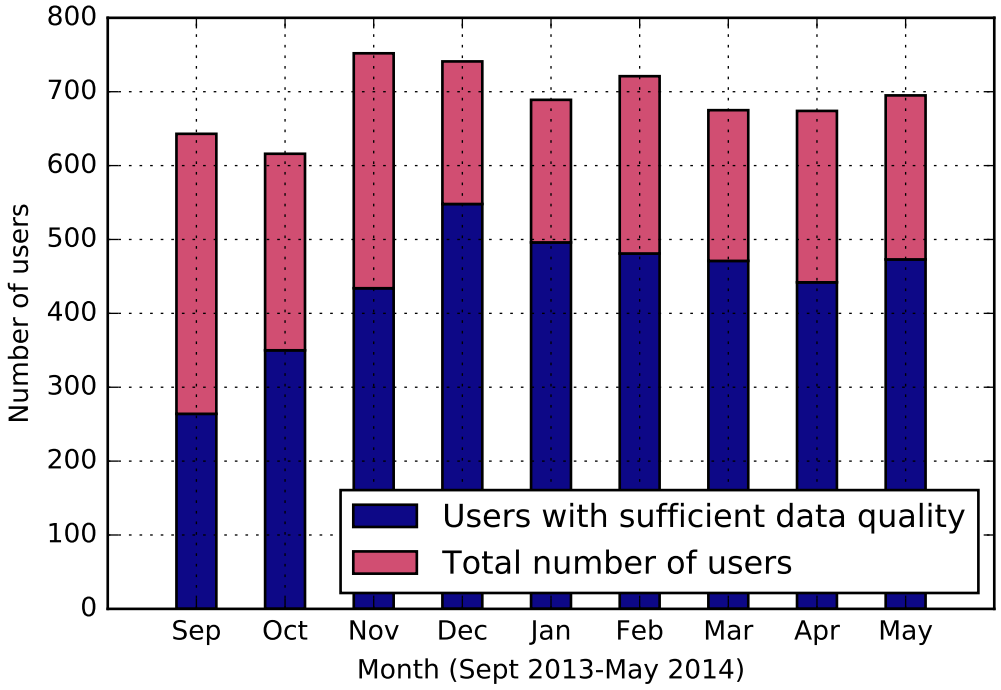


Figure 7: The number of users per month with sufficient Bluetooth data quality, as compared to the total number of users. The retention of users is lowest in September, and highest in December. After November the number of users with sufficient data quality (above 60%) remains roughly constant between 450 and 500 users.

3.1.2 Facebook Data

Study participants could opt-in to give the researchers access to their Facebook profiles, which a large majority of users did [16]. The study's Facebook data was collected using the Facebook Graph API, version 1.0 [16]. During the study, user profiles were queried every 24 hours. This included socio-demographic information (hometown, location, interests, and work), the friends list, and platform-related actions (feed, likes, statuses) mostly saved in the form of pairwise interactions. This data was made available for this thesis in two different forms: the first included monthly friendship graphs, and the other included all Facebook interactions between two users. These Facebook interactions are saved as directed tuples $(i, j, \mu, t, \theta_1, \theta_2)$ which correspond to statements of the form

[15] Percentage of pairs kept. Sep: 18%, Oct: 24%, Nov: 31%, Dec: 48%, Jan: 34%, Feb: 35%, Mar: 36%, Apr: 32%, May: 45%

[16] This API has been changed many times since, and version 1.0 is now deprecated [81].

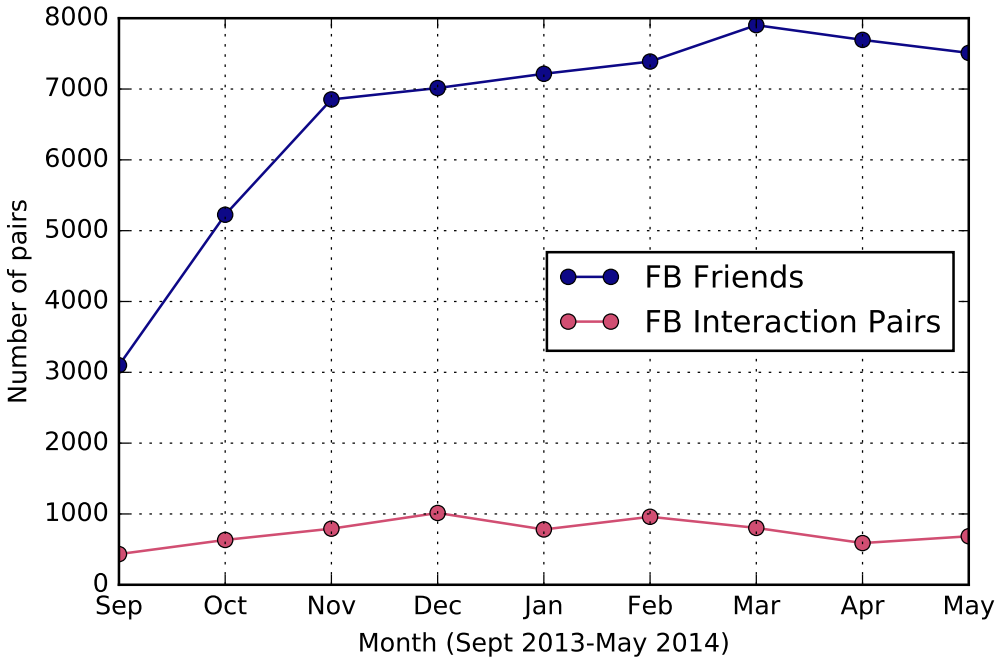


Figure 8: Number of Facebook pairs. The number of within-study pairs that are Facebook Friends, and those that interact in a given month. We clearly see that the number of friend-pairs plateaus around 7500, whereas the interaction has a development similar to the Bluetooth interaction, which suggests a seasonal variation in the amount of interaction between the study participants. Furthermore, the number of interactions pairs falls almost by half between February and April 2014, which is due to insufficient renewal of participant's Facebook access tokens.

'at time t user i interacted with user j in a message of type (θ_1, θ_2) with message ID μ'. Type θ_1 consists of five categories: 'comment', 'liked_story', 'message_tag', 'message_to', and 'tagged_story'. Whereas type θ_2 consists of sixteen categories: 'added_photos', 'added_video', 'app_created_story', 'approved_friend', 'checkin', 'link', 'mobile_status_update', 'offer', 'photo', 'question', 'shared_story', 'status', 'swf', 'tagged_in_photo', 'video', 'wall_post'.

The interaction data was extracted from the platform in two different dumps: one containing only the months Sep. 2013 - Feb. 2014, and a second one containing Mar.-May 2014. In its raw form, the second batch of data contains a high amount of duplicate rows. According to Radu Gatej, data manager of the Copenhagen Network Study, this could be due to a difference in the way the first and second dataset were queried: the second probably included all available data at every 24 hour sample (i.e. also data from the past every time). Therefore, I simply removed the duplicate rows before conducting the analysis in this thesis. Furthermore, the queried data contains user-pairs that do share an interaction on Facebook, yet are not Facebook friends. This is likely due to interactions via friends of friends. Since it is not a priori preferable to exclude this type of online interactions, these user-pairs are included in the analysis. Lastly, as shown in Figure 8, the number of interacting user-pairs start to decline slowly from March 2014 onwards. This is also mentioned by Sapiezynski et al. [9], and is due to insufficient renewal of participant's Facebook access tokens.

Since we characterise Facebook pairs based upon data about their interactions, it is relevant to know how many pairs are included in this analysis. In Figure 10 the number of pairs that use each type of interaction in a given month are displayed. There is a large variation in the amount of times certain types of interaction are engaged in, with 'Message_Tag' being used on average by only 100 pairs. An initial investigation further shows that the presence of any kind of Facebook interaction strongly increases the chance a pair will meet face-to-face when compared to a mere Facebook friendship tie (see Fig. 11). This correspond to earlier findings by Viswanath et al. and Sapiezynski et al. [9, 22]. The relative importance of the different interactions will be analysed further in section 4.2.

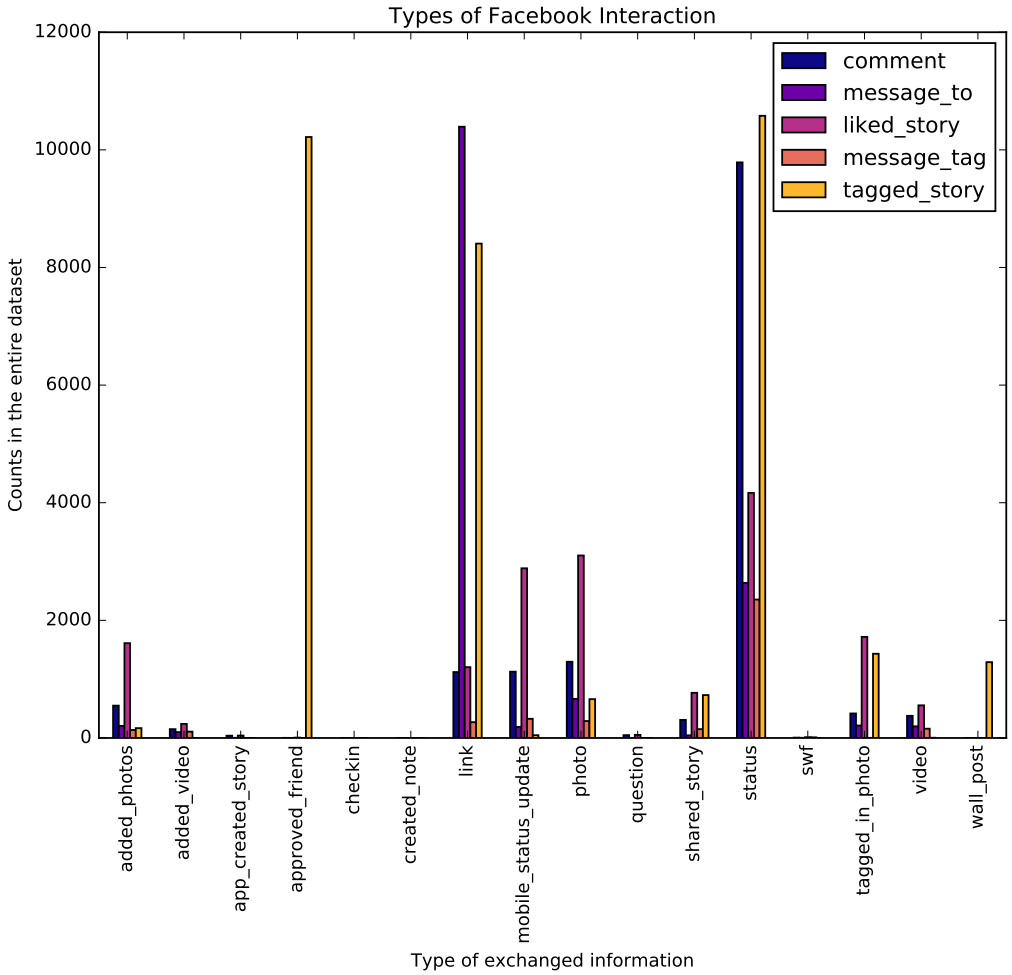


Figure 9: The different types of Facebook interaction in the dataset.

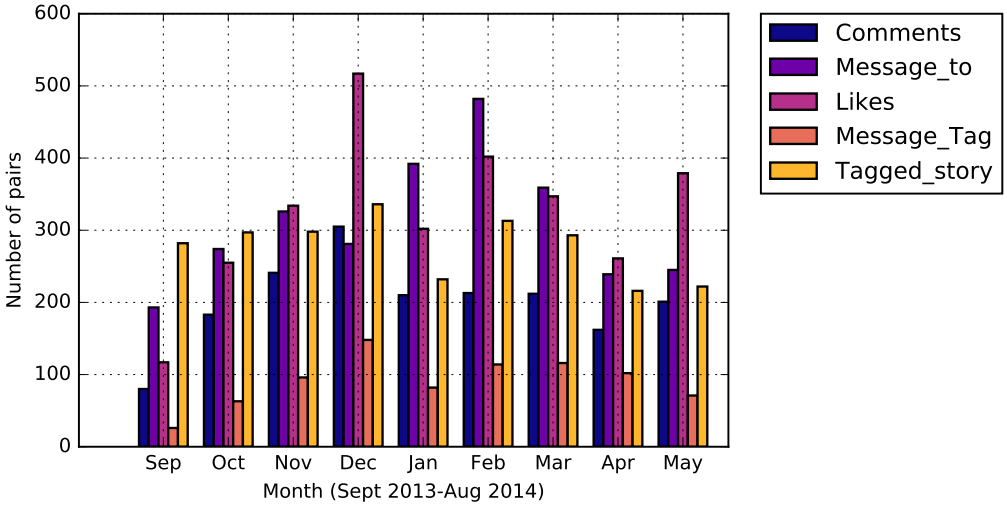


Figure 10: Pairs per type of Facebook interaction. The number of (within-study) pairs that engage in a certain type of Facebook interaction in a given month, per type of interaction (user pairs can appear twice). There is a large variation in the amount of times certain types of interaction are engaged in, with eg. 'likes' being exchanged by many more pairs than tagging in messages. Furthermore, we observe that some types of interaction are more variable than others. The relative prominence of 'Tagged_Story' in September and October is because the formation of a new friendship on Facebook is registered under this type.

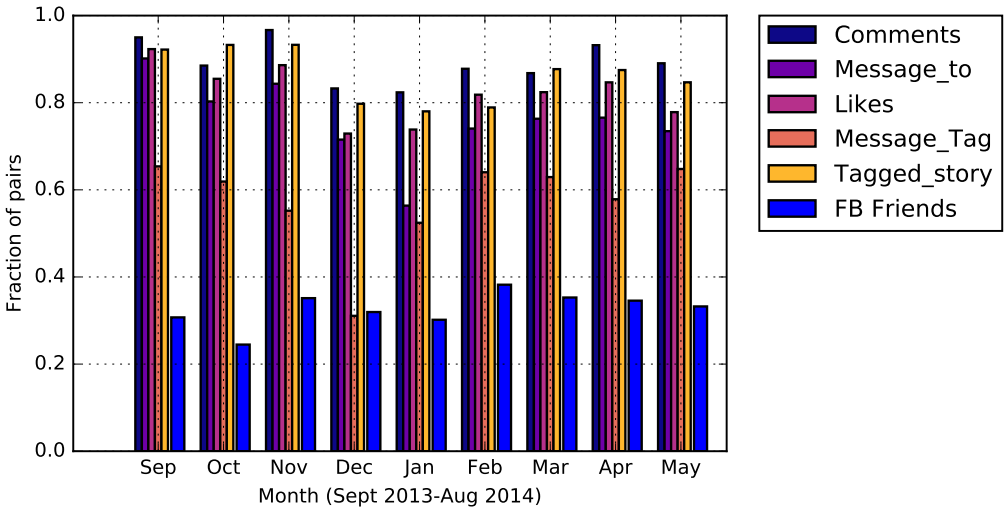
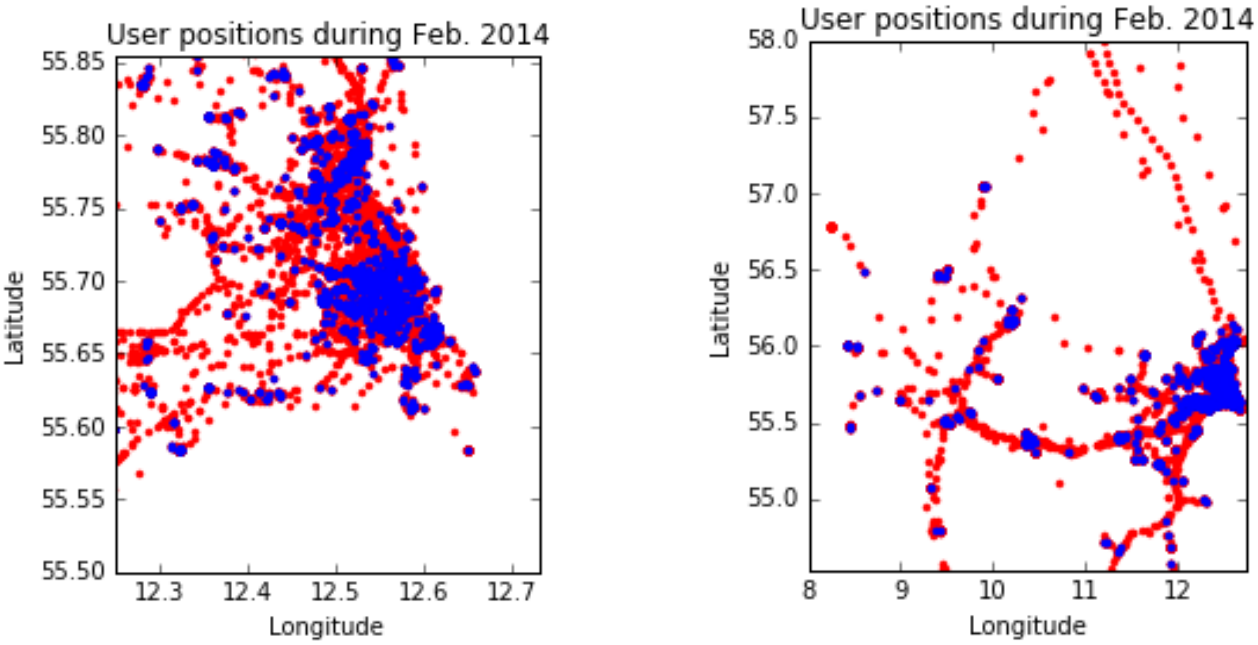


Figure 11: The fraction of (within-study) pairs that meet offline at least once in a given month, given their engagement in Facebook interaction of a given type (user pairs can appear twice). This can be regarded as a conditional probability: given a certain type of interaction on Facebook, what is the chance a user-pair will meet offline? Here we see that sending a message is a much stronger condition than merely being related on Facebook.

3.1.3 Location Data



(a) Region Hovedstaden (Copenhagen)

(b) Denmark.

Figure 12: The locations (red) and stop-locations (blue) of 100 users during February 2014. In Figure (a) one can clearly see the Copenhagen city centre, the location of the DTU campus, and the coasts of Zealand and Amager. With some goodwill, the transportation network making up the greater city area can also be discerned. Furthermore, when plotting all locations in a coordinate-frame corresponding to the boundaries of Denmark (Figure (b)), we can clearly see the stop locations in areas corresponding to the major cities, and location points along the traffic arteries (take special note of the island of Fyn, and its role as connection between Zealand and Jutland).

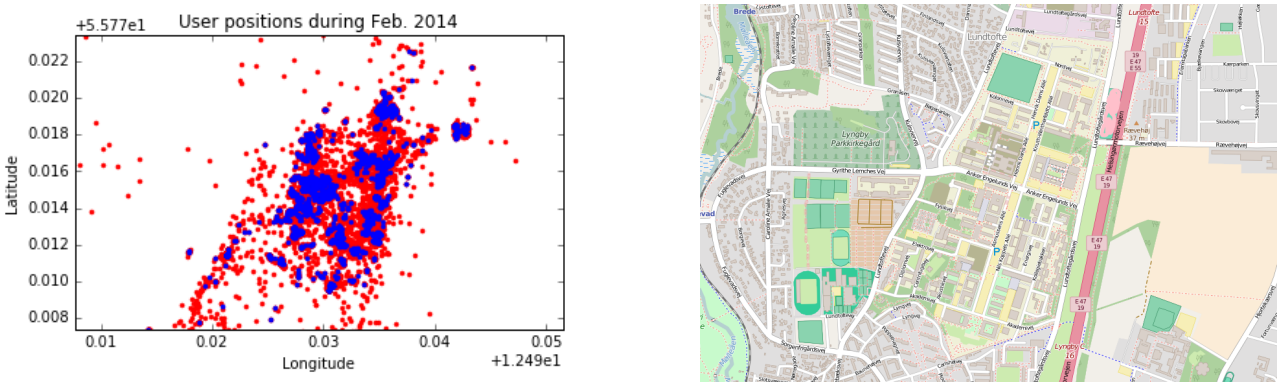
In the Copenhagen Network Study, user location was sampled opportunistically: whenever a smartphone application made a GPS request, this was recorded for study purposes [16]. Because of this opportunistic sampling, the time between observations is not constant. However, a good data quality is achieved when binning the GPS traces in 15 minute intervals, whereby the spatial accuracy is about 60 meters [82][17]. The location data is saved in the form (i, lat, lon, t) if a participant i was recorded at latitude lat and longitude lon at time t.

Previous work has shown that this data can be used to infer stop locations [82]. Cuttone et al. tested three different methods for the extraction of stop locations from the Copenhagen Network Study data: distance grouping, speed thresholding, and Gaussian Mixture Model clustering. However, they do not clearly state a preference or result regarding the comparison of these three. Distance grouping links later visits to earlier visits within a predefined radius, here set to 60 m. Speed thresholding bins locations and then determines the speed needed to go from one bin to the next. Whereas a Gaussian Mixture Model, with a predefined number of expected locations K, tries to find the combination of K normal distributions which maximises the likelihood of the observations (this is also used by Cho et al. to show that the locations 'home' and 'work' inform most of our daily behaviour [13]). Upon correspondence with the study authors, I learned that the final stop-locations were found by distance grouping using Density-based spatial clustering of applications with noise (DBSCAN) [23], and were labelled per individual user.

Clearly the spatiotemporal patterns of face-to-face encounters carry important information about

[17] There is redundant information in the WIFI, and CDR data traces that were observed for all study participants: both could be used to further pinpoint the location of the users [9]. However, for this thesis only the GPS traces were used.

the type of relationship between two individuals. Thus, initially I intended to investigate how the different online activity patterns relate to timing and location of offline encounters. In particular, I was interested to see if the multiplexity of the pair, i.e. the number of different contexts they share, is reflected in distinct online traces. To investigate the colocation of students, I intended to use the stop-location data. However, to compare the stop-locations of distinct users, these needed to be matched into a universal location vocabulary. The combination of stop-locations of different users proved unexpectedly challenging. Thus, I set this dimension of the analysis aside for future work and focussed on the contact between two individuals (as reflected in the Bluetooth data) rather than their colocation.



(a) User locations at DTU.

(b) The DTU campus on Openstreetmap.

Figure 13: The locations (red) and stop-locations (blue) of 100 users at DTU during February 2014. There is a clear accumulation of stop locations in areas corresponding to dormitories (Ostenfeld Kollegiet and Kampsax Kollegiet), Building 101 which houses the canteen, and several lecture buildings across campus.

3.2 Possible Data Biases

The population under consideration consists of undergraduate students at a middle sized western European university. As such, all results presented in this thesis are found for a predominantly young, white, well-educated, and mostly non-working population. These subjects are likely to have different daily rhythms, contact patterns, and online behaviour than the rest of the population. Although this does not limit the validity of the results in this study, it should be kept in mind when extrapolating the results to the Danish population at large.

Because it is not possible to relate Bluetooth media access control (MAC) addresses to specific Facebook ID's, unless we have a priori knowledge on the fact that they belong to the same user, we can relate the different networks only for study participants. As such, we have significantly more data on the interactions of study participants that have a high number of friends within the study. For all others we know roughly how socially active they are, but can only use this as control/baseline, not as object of study. This could be called a boundary specification problem: we have comparably little information on how the small community of study participants is embedded in the larger system of social life at DTU and beyond [16].

Furthermore, it is not possible to account completely for the influence that Facebook's site-specific algorithms have on the results. There is evidence that eg. the 'Birthday-reminder' function drives a large portion of interactions between weak Facebook ties [22], and the site may also expose some pairs to each other's newsfeed content more than others [1]. However, these platform-based mechanisms that drive user interactions, will mostly introduce noise on the data, and is unlikely to affect the result on a more systematic level.

4 Online to Offline Mapping

This chapter investigates which characteristics of Facebook interaction are most useful to predict offline interactions. There exist only few studies that investigate individuals' co-presence based on knowledge of their online interactions, and the most relevant predictor for offline interactions has not been identified so far. Thus, we start by extracting meaningful features from the online interaction data, where we focussed on measures of the total intensity and of the time-structure of interaction between a pair of Facebook users. Next, section 4.2 presents the results of experiments to classify offline behaviour based on the features that describe the online interaction. We have used binary classification to predict not one but several offline behaviours. Specifically, three different target variables were tested:

- 'BT(θ)' stands for Bluetooth, and designates the network of people that meet at least θ times in a given month. Without further specification 'BT' refers to 'BT(1)'. This variable is one of the most basic units of the face-to-face interaction network, and also the most important for the prediction of disease spread.
- 'OH' stands for off-hour and refers to those pairs that meet during non-class times, i.e. on weekends or weekdays between 17-8 o'clock. Eagle et al. have shown that contact during off-hours correlates with friendship [11], and thus the ability to classify this variable correctly would offer a strong indication that we can distinguish meaningful offline friendships.
- 'EVE' stands for evenings, and refers to meetings between 18-24 o'clock. Even more stringent than the off-hour variable, this variable was included in the analysis to investigate how precisely we can distinguish between friendship-related offline behaviours, i.e. how far we can push the limit of classification based on online interaction.

We find that it is possible to predict the occurrence of an offline encounter between an interacting pair on Facebook with an accuracy of 78 %. The total amount of interaction on Facebook was deemed the most important feature, as long as we control for user's differing activity levels on Facebook. The accuracy of predictions of offline meetings is likely limited by chance encounters, however we can predict more stringent target variables with an accuracy of roughly 70 % also. Furthermore, our analysis reveals some fundamental differences between online and offline interactions, most interestingly, we find that the timing of interactions is much more meaningful offline than online.

4.1 Aspects of Online Interaction

In its raw form, the Facebook interaction data is essentially time-series data, which records when and how a pair interacts. The first aim of this thesis is to investigate which aspects of the high-dimensional interaction data are most important to capture the essence of a Facebook friendship. This is also a key first step for the classification of offline behaviour. The task of extracting features is not well posed, and in principle an infinite number of features could be derived for every user-pair. Choosing the features is a process of hypothesis-driven testing, guided by hints from the literature and knowledge of the subject area. Hereby the classifier gives valuable insight regarding which features are important to predict the target variable.

4.1.1 Interaction: Activity Level and Directionality

Previous research has shown that the amount of interaction between a pair of Facebook users is a strong - if not the strongest - indication of friendship between them [28]. There are different types of Facebook interaction (see section 3.1.2), and the correlations between them hint at the existence of typical sequences or combinations of interaction. In table 2 we see that the exchange of comments and likes is moderately correlated, which suggests that users who comment on each other's posts

also 'like' more posts. These results correspond well to those reported by Jones et al. [28], although the strength of the correlations is somewhat lower. The difference in magnitude of the correlation coefficients is likely due to our shorter observation window (one month), since preliminary results for a combination of all nine months of data indicate values that lie closer to those reported by Jones et al.

	Comments	Message_to	Liked_Story	Message_Tag	Tagged_Story
Comments	1.0	0.31	0.44	0.24	0.39
Message_to		1.0	0.26	0.29	0.19
Liked_Story			1.0	0.14	0.15
Message_Tag				1.0	0.12
Tagged_Story					1.0

Table 2: Correlations of different user-pair interactions in February 2014.

Counter to our intuition, we initially observed very little importance and effect of the 'total number of interactions' variable in the prediction of offline contact. The types of interaction (as subdivided according to θ_1 from section 3.1.2: 'comments', 'message_to', 'liked_story', 'message_tag', and 'tagged_story') were not deemed important separately either.

Thus, we investigated whether the effect may be obscured by the varying activity levels of Facebook users, i.e. the importance different Facebook users attribute to an interaction, and/or the shear volume of interactions they generate. To study this, all 6 measures of interaction strength were rescaled according to the activity-levels of the pair involved in the interaction. Let $INT_\vartheta(i,j)$ be the number of interactions of type ϑ between pair (i,j) and let $a_{\vartheta,x}$ denote the total activity of user x with respect to type ϑ, then:

$$Disc\_INT_\vartheta(i,j) = \frac{1}{2}\left(\frac{INT_\vartheta(i,j)}{a_{\vartheta,i}} + \frac{INT_\vartheta(i,j)}{a_{\vartheta,j}}\right) \tag{4.1}$$

is the 'discounted' strength of interactions of type ϑ between pair (i,j). This corrected interaction will be low if both users are highly active in general but share only few interactions, and high if the interactions between i and j form a large portion of their total Facebook activity[18]. Since these activity-corrected variables are rescaled based upon the interaction-specific activity level per user, most of the correlation of the original interaction variables is lost (see Table 3).

	Comments	Message_to	Liked_Story	Message_Tag	Tagged_Story
Comments	1.0	0.09	0.23	0.04	0.2
Message_to		1.0	0.02	0.09	0.0
Liked_Story			1.0	-0.04	0.01
Message_Tag				1.0	-0.01
Tagged_Story					1.0

Table 3: Correlations of different activity corrected user-pair interactions in February 2014.

Upon rescaling, the interaction variables instantly became some of the most important features for classification (see section 4), which is in line with the results from Jones et al. [28]. However, our results go even further and suggest a difference in the relative importance of Facebook interactions depending on the total activity level a user exhibits on the social networking platform.

This raises two interesting questions for further study: Firstly, do more active FB users have different interaction patterns, or a different way of using the medium? To study this effect further it would be good to include the total activity levels of user i and j as separate features in the classification. Furthermore, a second type of activity-measure could take into account the number of distinct users

[18] Here the Facebook activity level also includes the interactions with users outside the study, and all values are calculated on a monthly basis.

that a given user i interacts with (rather than the total number of interactions they engage in). The second area of further investigation would be to quantify whether a tie or interaction is in some sense abnormal for a certain user. For example, future research could investigate interactions that occur outside of 'normal interaction times', or take into account a possible imbalance in the relation between i and j. This imbalance could be characterised by comparing the number of interactions user i initiates $int(i \to j)$ with those that are initiated by j, i.e. $int(j \to i)$. Some steps have been taken to implement both the directionality of interaction and the second type of activity-measure, but neither have been included in the analysis as of yet. This is a promising area for future work.

4.1.2 Temporal Entropy

A common method to characterise the predictability of interactions is to investigate the entropy of the time of day an interaction takes place, or of the time between interactions [32]. In a recent study, Sapiezynski et al. showed that the entropy of when two people meet (as measured using Bluetooth data) is one of the most important features of offline data in predicting contact on Facebook [9].

Thus, we investigate the relation between the temporal entropy of on- and offline contacts, for two notions of the time-distribution of contact: (i) the entropy of the waiting time distribution of the pair-interactions[19], which captures information about the regularity of the time between interactions, and (ii) the entropy of the distribution of interactions into 168 hour-bins of the week, which describes the regularity of the time of meeting. For both types of entropies, we expect that pairs who have a regular meeting each week (such as a shared class, taking lunch together every day, or meeting for beer on Fridays) will have a low entropy, whereas pairs who interact more randomly will have higher entropies.

Let the set of distinct times - either waiting times between interactions of pair (i, j) or the 168 hour-bins - be denoted by $W = \{\Lambda_1, \ldots, \Lambda_n\}$. Then the entropy is:

$$S(W) = -\sum_{k=1}^{n} P(\Lambda_k) \ln \left(P(\Lambda_k)\right) \tag{4.2}$$

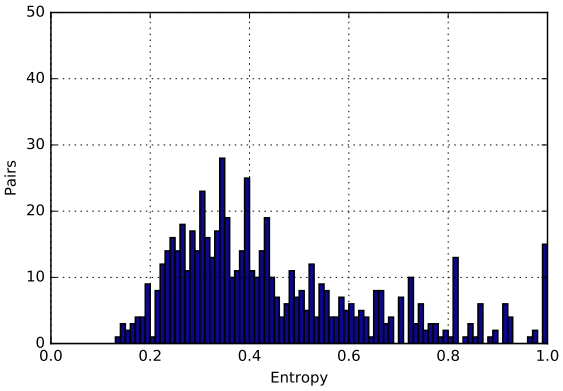
where $P(\Lambda_k)$ is the probability that W takes the value Λ_k. These probabilities were approximated by dividing the number of observations of event Λ_k by the total number of observations for the pair (i, j). Since the entropy scales as a function of the number of observations N_{obs}, we normalise by the maximal possible entropy which corresponds to the case where $P(\Lambda_k) = 1/N_{obs}$ for all k.

These entropies were calculated both for the Facebook and face-to-face interactions between each pair in the month of February 2014, whereby only those pairs that shared at least 3 interactions in this month were included in the analysis[20]. Figure 14 shows the distribution of entropy values for all included pairs. For longer observation windows, e.g. 9 months, the form of these distributions stays the same.

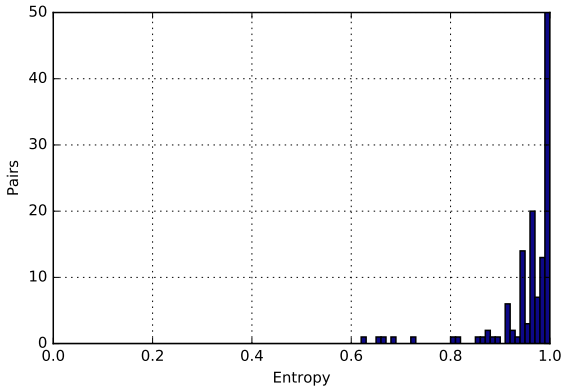
Figure 14 shows that face-to-face meetings have lower entropies and are thus more regular than interaction times on Facebook. This is the case both for waiting times and the hour of the week. Interestingly enough, the correlation between the waiting time and the hour of the week entropy of the Bluetooth traces reaches $(-0.19 \pm 2.0 \cdot 10^{-6})$. Presumably, this is because it is much more probable offline that a pair will interact again in the next time-bin than it is online. Furthermore, the correlation between on- and offline interaction entropy is rather small or non-existent: only (-0.11 ± 0.11) for the hour of the week entropy, and (0.05 ± 0.09) for the entropy of the waiting times. These results point to a central difference between online and offline interactions: The time of interaction does not carry as much weight on Facebook as it does offline.

[19] To reduce the space of possible waiting times somewhat, these were grouped into 4-hour bins.

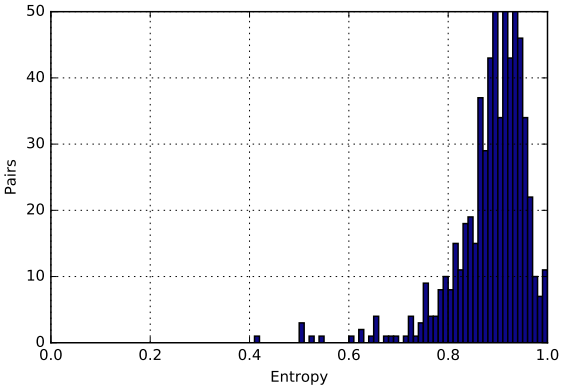
[20] The restriction to at least 3 interactions is prompted by the fact that at least two values are need to compute a non-trivial entropy, and waiting times are calculated as the time difference between two interactions, which shortens the total number of values by one.



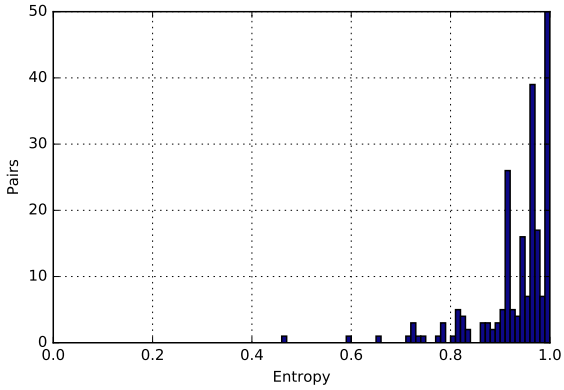
(a) Bluetooth Waiting Times.



(b) Facebook Waiting Times.



(c) Bluetooth Hour of the Week.



(d) Facebook Hour of the Week.

Figure 14: The temporal entropy of Bluetooth and Facebook interactions between user-pairs (threshold at 3 interactions). On the top row, the temporal entropy of waiting times offline (a) and online (b) is shown. The second row shows the hour of the week entropy, for the Bluetooth (c) and Facebook (d) interactions. The y-axis of the Facebook distributions extends much further (due to the large number of pairs with very high entropy), but has been cut to 50 for comparison with the Bluetooth histogram. In total, the Bluetooth histograms includes 630 pairs, whereas the Facebook histograms feature only 280 because the threshold of 3 interactions proved more stringent here. However, for longer observation windows (9 months) and a larger number of pairs the form of these distributions stays the same.

4.1.3 Features for Classification

For each pair (i, j) that shared at least one Facebook interaction in a given month, the following features were extracted from their Facebook interactions and included in training the classifier:

- **Tot_int**: The total number of interactions.
- **Disc_int/comment/message_to/liked_story/message_tag/tagged_story**: The total number of exchanged interactions, 'comment', 'message_to', 'liked_story', 'message_tag', and 'tagged_story', while accounting for the activity of the users involved (as defined in section 4.1.1).
- **CNN**: The number of common nearest neighbours, i.e. the number of shared Facebook friends.
- **CNN_int**: The number of shared interaction friends, i.e. persons k that both i and j have interacted with on Facebook in the given month.
- **Min/Max/Mean_Wait**: Waiting times: the minimal, maximal, and mean time that passed between two Facebook interactions (in hours).
- **Resp_rate**: The order of response, which measures whether user j is typically the first, second,

third etc. person to comment on a post by user i (and vice versa). To compute this variable, first all unique message ID's of interactions between i and j are extracted, and it is recorded in which order other users commented on these same messages. The order of response variable is then found by averaging over the rank of the first comment per unique Message ID.

- **Prev**: Previous: the total number of exchanged interactions in the previous month, discounted for activity. 'Prev' has no value for the month of September 2013, thereafter it is equivalent to 'Disc_int'$[t-1]$ where t is the given month.
- **First**: the date of the first observed interaction between a pair, in the number of days relative to 01.09.2013 (the start of the semester). This is a proxy for the 'age' of a Facebook relationship.

	Tot no int	(0)	(i)	(ii)	(iii)	(iv)	(v)	CNN	CNN Int	Max wait	Min wait	Mean wait	Resp rate	Prev	First
Tot_no_int	1.	0.32	0.4	0.23	0.31	0.2	0.15	0.22	0.24	0.37	-0.06	0.08	-0.1	0.11	-0.18
Disc_int (0.)		1.	0.28	0.29	0.22	0.18	0.17	-0.09	-0.09	0.11	-0.01	0.03	-0.09	0.	-0.09
Disc_comment (i)			1.	0.09	0.23	0.04	0.2	-0.03	-0.06	0.2	-0.04	0.05	-0.17	0.23	-0.16
Disc_message_to (ii)				1.	0.02	0.09	0.	-0.11	-0.08	0.09	0.03	0.04	0.03	0.16	-0.2
Disc_liked_story (iii)					1.	-0.04	0.01	-0.06	-0.13	0.21	0.03	0.1	-0.18	0.26	-0.19
Disc_message_tag (iv)						1.	-0.01	-0.11	-0.1	0.	-0.07	-0.07	-0.08	0.	0.1
Disc_tagged_story (v)							1.	-0.18	-0.18	0.08	0.	0.03	-0.24	0.11	0.13
CNN								1.	0.62	0.15	0.02	0.09	0.19	-0.06	-0.16
CNN_Int									1.	0.2	0.04	0.12	**0.43**	-0.07	-0.08
Max_wait										1.	0.5	0.84	-0.07	0.08	-0.17
Min_wait											1.	0.86	-0.01	0.01	-0.06
Mean_wait												1.	-0.04	0.04	-0.11
Resp_rate													1.	-0.08	-0.05
Prev														1.	-0.2
First															1.

Table 4: Correlations of all features for February 2014. Not surprisingly, the interaction variables are all weakly correlated with another. There is also a weak correlation between the total number of interactions and the number of friends or shared interaction friends (CNN_Int). Furthermore, there is a moderate positive correlation between CNN_Int and the Response rate.

4.2 Classification

4.2.1 Implementation

All features mentioned in section 4.1.3 were included in training the classifier. First, the user-pairs and corresponding features were extracted from monthly slices of the Facebook interaction data. Then the classifier was trained either on each of these dataframes of monthly pairs individually, or all monthly sets were combined (concatenated). The resulting number of pairs is listed in Table 5. We see that concatenation greatly increases the number of observations that can be used for training and testing, however it treats user-pairs - even between the same two users - as being independent from month to month. Furthermore, it tacitly assumes there are no strong yearly- or semester-cycles which change the predictive power of certain variables during certain months.

Since the absolute values of the features differs greatly, it is important to standardise them before performing any further machine learning tasks. As is common in the literature, this was done using the Z-score, which compares a sample of the random variable X, with sample mean μ and standard deviation σ, to a standard normally distributed variable [23]:

$$Z = \frac{X - \mu}{\sigma} \tag{4.3}$$

The model was trained and tested using a 10 fold cross-validation. In each fold the Accuracy, ROC AUC, Precision, Recall, F1 score, and Matthew's correlation coefficient were calculated (as described in section 2.2.2). To compute the overall ROC curve, the predictions (scores) on each test set were recorded and combined [68]. For the other metrics a generalisation error was calculated by averaging the scores over all folds.

All code was written in Python, using Ipython notebooks that run on a secure Jupyter server at DTU [83]. For the implementation of the different classification algorithms, I used the scikit-learn library for Python [84].

	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Concat
Pairs	433	633	792	1014	781	960	803	589	685	6690

Table 5: The number of pairs the classifier was trained on for each month. 'Concat' refers to the concatenation of all 9 monthly dataframes.

4.2.2 Classification Results

In Figure 15 the value of different classification metrics is shown as a function of the month the classifier was trained on[21]. These plots depict the predictive strength of the classifier for each monthly dataframe. First of all, we see that the accuracy for predicting 'BT(1)' is 78%, with very high precision and recall. Secondly, the predictive power (i.e. the value of the performance metrics) is mostly constant over time, except for a dip around December/January. From section 3 we know that in these months users had less offline contact, and as such the prevalence of the positive class decreases sharply. The classifier can not fully compensate for this, and thus the performance decreases.

The performance results of both classifiers can be read more clearly in Table 6, both for the month of February 2014 and the concatenated months. The month of February 2014 was chosen because it features the best combination of the number of pairs (see Table 5) and stability of the offline social interactions. Table 6 shows that the difference between the performance of a classifier trained

[21] The Recall is 1.0 for the Majority Vote Classifier when the positive class is in the majority, and 0 when not. Also, the ROC AUC will always be 0.5 for a Majority Vote (Random) Classifier.

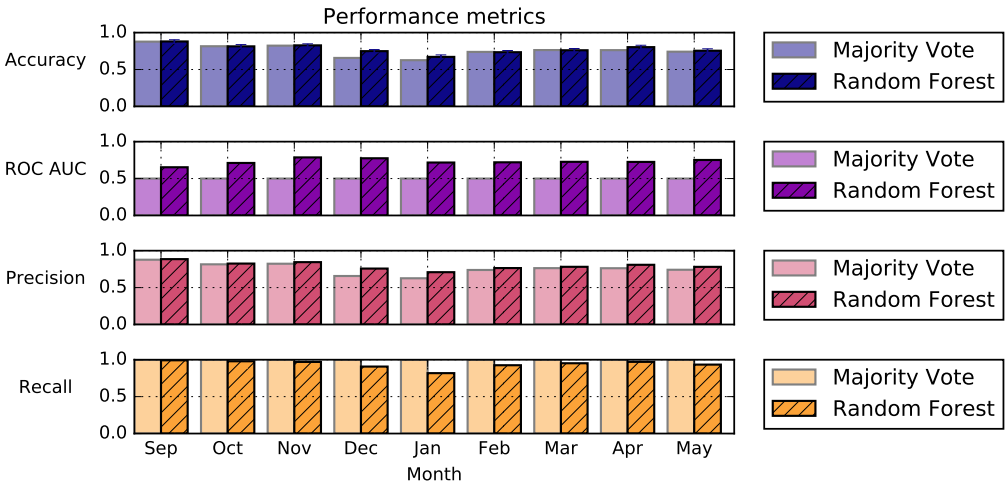


Figure 15: The classification performance for a classifier trained to predict the target BT(1). In most months the accuracy of the Random Forest Classifier is quite high, although not significant with respect to the Majority Vote Classifier.

on one month of data or the concatenation is minor, except for the reduced error bands on the accuracy (as introduced in section 2.2.5). Concatenation thus greatly reduces the uncertainty on the generalisation error.

Since the prevalence of the positive class is quite high for the target 'BT(1)', the Majority Vote Classifier performs almost equally well as the Random Forest Classifier. To investigate the relation between predictive power and the target variable, we study the sensitivity of the classifier performance to the target variable BT(θ) as a function of the threshold θ. Figure 16 shows that the performance of the Random Forest classifier stays relatively constant as a function of the target threshold, whereas the Majority Vote accuracy follows the class imbalance in the sample. The biggest difference between the Random Forest and the Majority Vote Classifiers is achieved for a threshold of thirty meetings per month, i.e. where the positive class transfers from majority to minority (as seen by the fact that the Majority Vote Classifier's Recall switches from +1 to 0). As such, as we increase the threshold of the number of encounters per month, the quality of the classifier becomes more apparent.

These results indicate, that our classifier does not merely 'get lucky' because of the high probability a pair will encounter each other at least once. Rather, we extract meaningful offline relationships that meet offline for quite a substantial amount of time (30 observations per month). For applications in prediction or recommender systems, the performance of the classifier should likely be improved further. However, we have shown some very promising hints both of the results and feasibility of such an undertaking.

		February 2014		Concatenated Months	
		Random Forest Classifier	Majority Vote Classifier	Random Forest Classifier	Majority Vote Classifier
BT(1)	Accuracy	0.74 ± 0.025	0.74	0.78 ± 0.009	0.74
	ROC AUC	0.71	0.5	0.74	0.5
	Precision	0.77	0.74	0.79	0.74
	Recall	0.93	1.0	0.95	1.0
	F1-score	0.85		0.86	
	Matthew	0.26		0.33	
BT(27)	Accuracy			0.69 ± 0.005	0.53
	ROC AUC			0.77	0.5
	Precision			0.70	0.53
	Recall			0.75	1.0
	F1-score			0.72	
	Matthew			0.39	

Table 6: Performance metrics for the classification of BT(1) and BT(27). The difference between the performance of a classifier trained on one month of data (960 pairs) or the concatenation (6690 pairs) is minor, except for the error bands on the accuracy.

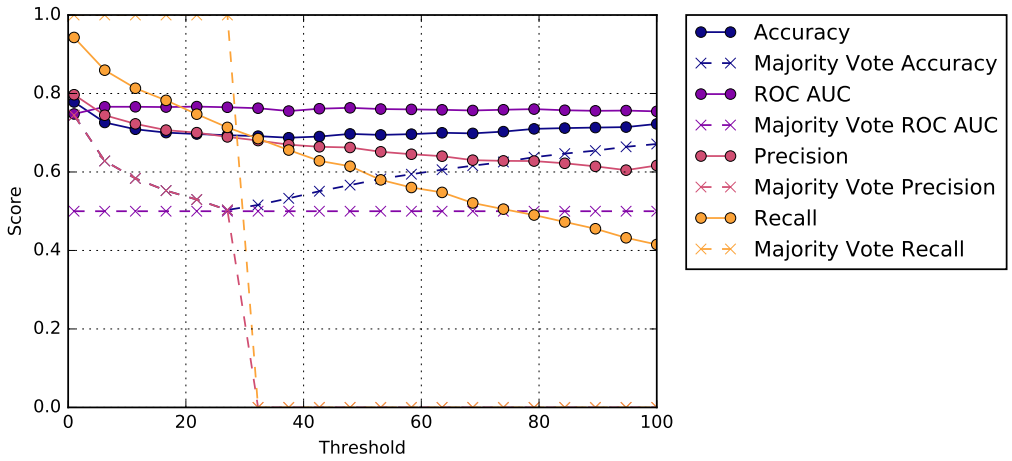


Figure 16: Performance metrics of the Random Forest and the Majority Vote Classifiers as a function of the target threshold θ. The difference in classifier performance is maximised at a threshold of roughly 30 meetings per month, which is where the amount of samples in both classes is balanced.

4.2.3 Feature Importance

In this section, we study the importance of the different features in training the classifier. Figure 17 shows the feature importances for the classifier that predicts offline contact BT(1). Some features are clearly more important in determining offline behaviour than others. Especially the variables related to the number of interactions are deemed important, which is in line with the results obtained by Jones et al. [28]. However, these are only significant once we account for the activity level per user. This can be seen clearly by the low importance of the Tot_no_int variable, which was included as comparison. Furthermore, the variables that hold information about the waiting times are almost negligible. The fact that i.e. the date the pair first interacted ('First'), is such a strong predictor of offline interaction further points to some characteristic temporal development of a Facebook friendship.

Note that for highly correlated features, most notably CNN and CNN_int, one must take care not to make strong statements about the relative importance. This is because the classifier will pick any of

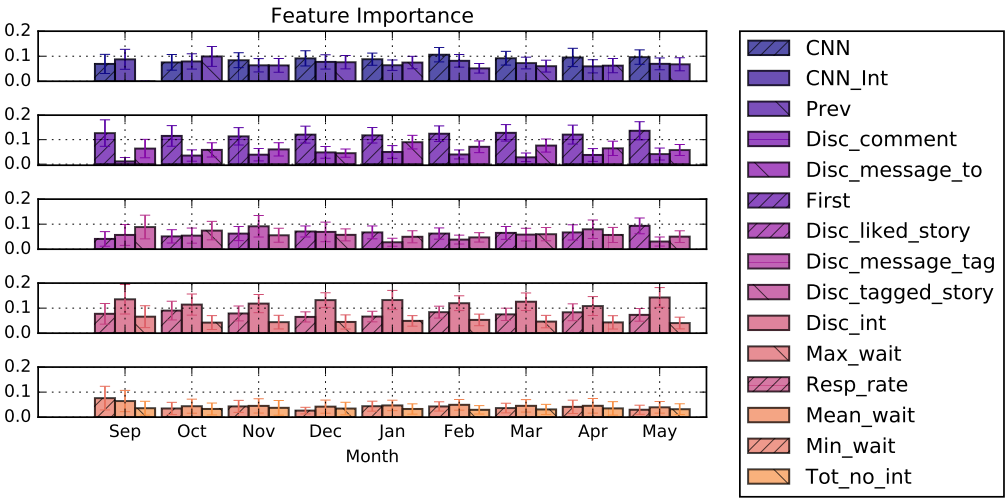


Figure 17: The feature importances for nine classifications of BT(1). Feature importance and standard error were extracted from the Random Forest Classifier. The width of the slivers indicates the feature importance, whereby all feature importances always sum to one. The relative importance of the features varies somewhat between months, however the discounted interaction measures are much more important than the waiting times throughout (further explanation of the variables is given in section 4.1.3).

the correlated features as predictor more or less at random, but once one is used the importance of the others is strongly reduced, even though the features may have a similar relation with the predicted outcome (see Table 4 for all feature correlations). Given the moderate correlation between almost all features, the method to determine relative feature importance will have to be refined. In particular, future work will investigate whether e.g. ten moderately correlated features could 'cancel each other out', and thus seem less important than one completely uncorrelated feature.

To determine whether all features contribute to the accuracy of the classifier, one should compare the classifier performance for different sets of input features. Testing all possible feature sets is hardly feasible, which is why recursive feature elimination (RFE) selects features by recursively considering smaller and smaller sets of features [23]. With cross-validation RFE can also determine the optimal number of features [69]. We performed RFE with cross-validation on the concatenated Facebook data, and found that all features important were deemed important.

To study how well one can generalise the results of training the classifier on one month of Facebook data, we compare the feature importance over nine classifications. These classifications are mostly independent, and present a way to investigate how robust our findings are to monthly changes in the data. Figure 18 shows that there is some difference between features regarding their variability from month to month. All in all the importances are relatively stable, which is an argument in favour of using the concatenated dataframe.

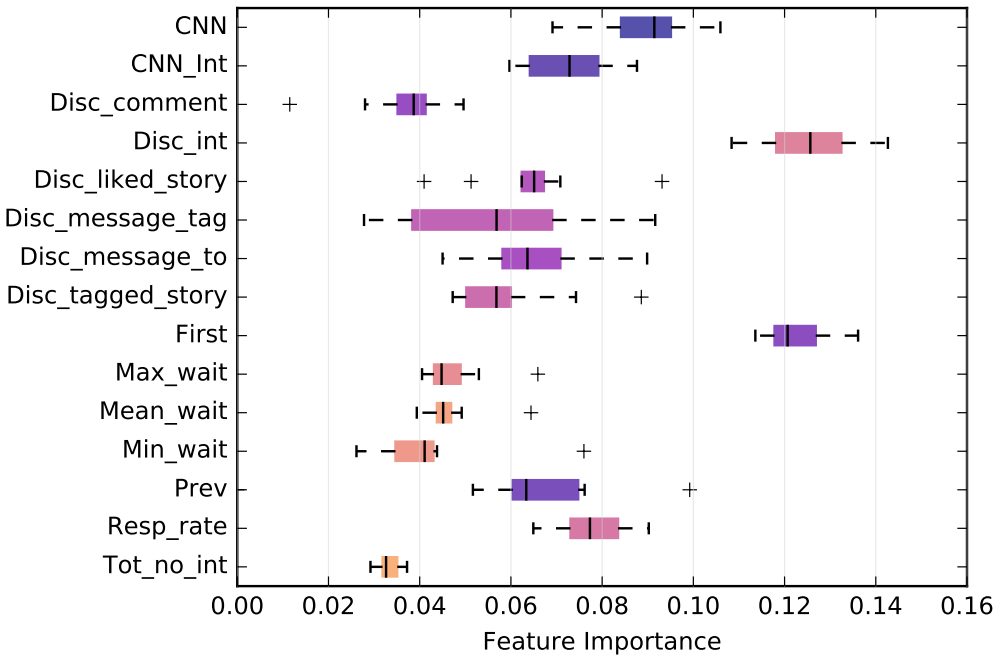


Figure 18: The average feature importances for nine classifications of BT(1). Further explanation of the variables is given in section 4.1.3. We clearly see that 'Disc_int', i.e. the total level of interaction - once accounted for the activity per user -, is most important. It is followed by 'First', i.e. the date the pair first interacted, and 'CNN' which denotes the shared number of friends. The waiting times and total number of interactions (not accounting for activity levels) are deemed least important.

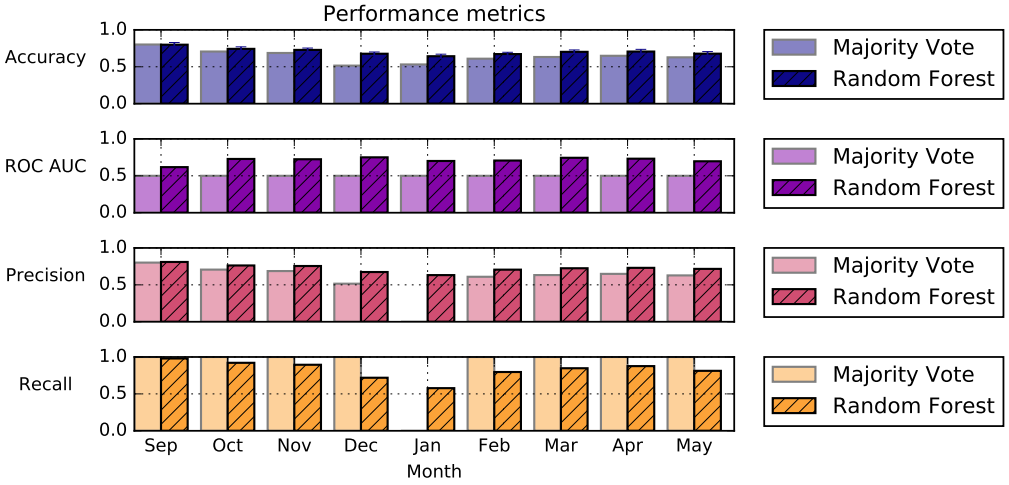
4.2.4 Off-Hour and Evening Meetings

To better understand which aspects of the offline social tie can be learned based on Facebook interaction only, we investigate two further target variables: whether a pair meets during 'off-hours' (OH) or in the evening time (EVE).

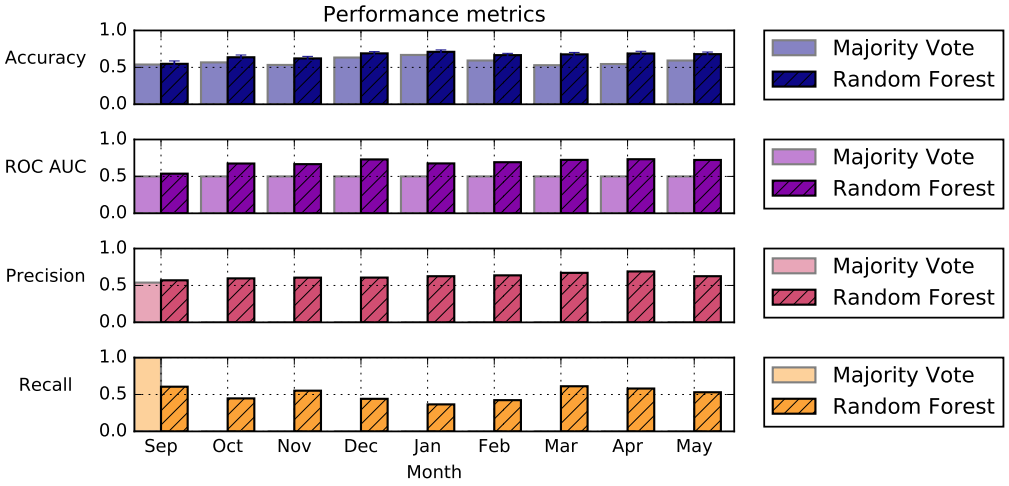
The performance metrics of the classification of these target variables are listed in Table 7, both for the month of February 2014 and the concatenated months. We find that the target OH can be predicted with an accuracy of 69%, and EVE with 65% accuracy. The accuracy and other performance metrics are lower for the classification of off-hour and evening meetings than for 'BT(1)'. Yet the difference between the Random Forest and Majority Vote Classifiers is higher in these cases, similar to the case for BT(27)[22]. This suggests that the more stringent target variables allow for training of the classifier function, and that the Majority Vote Classifier performs worse on the more balanced datasets.

Besides the off-hour and evening target variables, some further target variables were investigated, most notably the archetypes that DTU master student Linards Kalnins extracted from the Bluetooth data. Each choice of target variable is complicated by the careful balance one has to strike between specificity of the target behaviour, and the resulting class imbalance in the sample. This activity was put on hold due to time restrictions, however it is an interesting direction for future research to compare target variables more systematically.

[22]The ratios of positive to negative class of the OH and EVE target variable are roughly 5:3, and 7:10 respectively.



(a) OH

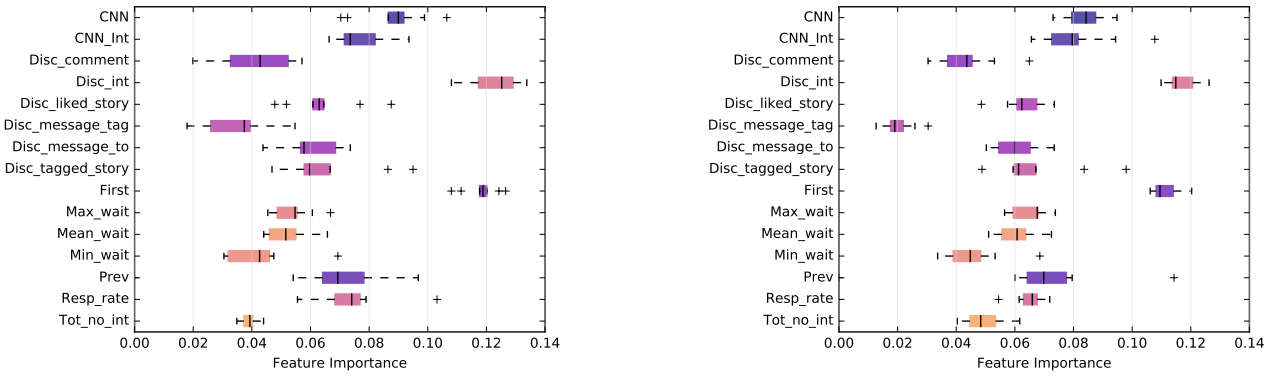


(b) EVE

Figure 19: The classification performance for a classifier trained to predict the (a) OH and (b) EVE target variables. The accuracy of the Random Forest Classifier is significantly better than the Majority Vote Classifier in most months, but much lower than for the 'BT(1)' variable overall. This means the target is harder to predict, yet the classifier more strongly profits from knowledge about the pair-features. Furthermore, although the 1-class is clearly in the minority in the case of EVE (Precision and Recall of 0 for the Majority Vote Classifier), the Random Forest achieves a Precision of roughly 65%.

		February 2014		Concatenated Months	
		Random Forest Classifier	Majority Vote Classifier	Random Forest Classifier	Majority Vote Classifier
OH: Off-Hour	Accuracy	0.68 ± 0.024	0.61	0.69 ± 0.007	0.62
	ROC AUC	0.72	0.5	0.73	0.5
	Precision	0.71	0.61	0.72	0.62
	Recall	0.81	1.0	0.84	1.0
	F1-score	0.74		0.77	
	Matthew	0.29		0.32	
EVE: Evening	Accuracy	0.67 ± 0.03	0.59	0.65 ± 0.01	0.58
	ROC AUC	0.71	0.5	0.70	0.5
	Precision	0.63	0.0	0.62	0.0
	Recall	0.44	0.0	0.47	0.0
	F1-score	0.51		0.54	
	Matthew	0.27		0.27	

Table 7: Performance metrics for the classification of OH and EVE. For both target variables, the difference between the performance of a classifier trained on one month of data (960 pairs) or the concatenation (6690 pairs) is minor, except for the error bands on the accuracy. For the **EVE** target the performance metrics of the Random Forest Classifier are significantly better than for the Majority Vote Classifier.



(a) OH **(b) EVE**

Figure 20: The average feature importances for nine classifications of the (a) OH and (b) EVE target variables. Further explanation of the variables is given in section 4.1.3. We clearly see that 'Disc_int', i.e. the total level of interaction - once accounted for the activity per user -, is most important. It is followed by 'First', i.e. the month in which the pair first interacted, and 'CNN' which denotes the shared number of friends. The waiting times and total number of interactions (not accounting for activity levels) are least important.

4.3 Discussion

In this chapter, we set out to better understand the relationship between online and offline interactions, and to use Facebook interaction data to predict different offline behaviours. We found that it is possible to predict the occurrence of an offline encounter between an interacting pair on Facebook with an accuracy of 78 %. More stringent target variables, such as a meeting during off-hours or in the evening time, could be predicted with 69 % and 65% accuracy respectively. This accuracy is lower than the study of Jones et al., which used Facebook to classify self-reported friendships [28]. However, their investigation was restricted to a dataset which included the closest friend and one random Facebook tie of slightly more than 700 users (i.e. a balanced dataset of 1500 pairs). In our case the accuracy is presumably limited by the fact that the presence of a single encounter in a given month is an event highly affected by chance. A classifier which assigns all pairs a 100% probability of meeting, performs almost as well as a classifier that was trained on features from the Facebook interaction. Given the comparably small size of the DTU campus, freshman social circles, and the inner city of Copenhagen, it is not surprising that many study participants meet at least once per month. This points at an important limitation of the use of Facebook to predict offline contact: the study of Facebook interactions can predict friendships and corresponding regular or planned offline meetings, but carries very little information about chance encounters.

Furthermore, we found that the same features were deemed important for the prediction of normal, off-hour, and evening contact. The number of interactions was most important to predict offline contact, as long as we account for the activity level of a given user. Secondly, it was important when a pair first started interacting on Facebook. This may be related to previous results by Viswanath et al., which show that the relative strength of a Facebook tie is related to its age [22]. Not surprisingly, how many friends a pair has in common on the medium also carried a lot of weight. However, the frequency of interaction (as measured by inter-event times) had an almost negligible effect on the prediction of offline meetings. This is supported by a separate investigation in which we showed that the temporal entropy of interactions is much more informative offline than for online interactions. Combined, these results indicate that although the timing of interactions is important offline, online the mere presence of interaction is the more relevant object of study.

These results give some clear first hints regarding the relative importance of features and the feasibility of the prediction of offline contact. However, the predictive strength of the classifier can likely be improved by including further features to describe the Facebook interaction. For example, this study has not taken any socio-demographic information of the participants into account, which could be extracted from their Facebook profiles. Additionally, we have shown that a promising avenue for further study would be to investigate how different types of Facebook use affect the predictive strength of variables.

How to relate the classification of ties to the structure of the resulting network is a non-trivial question. The above results tell us very little about the applicability of this method for network-based investigations. Therefore, chapter 5 will introduce our novel way of assessing the quality of the network prediction.

Given the importance of tie strength, it would be a logical step to use a weighted network for the investigation of offline disease spread. The predicted network could easily be made a weighted network by letting the link weights depend on the classification probability, or using e.g. the presence of a link in the ‘EVE’ or ‘OH’ networks to attribute a higher weight. However, in the long run predicting binary links itself will not be sufficient, rather we will want to predict interaction intensity. This thesis has focussed on binary link prediction because we started with the desire to identify different kinds of interaction, that cluster Facebook ties according to their offline analogue. Furthermore, binary prediction it is the most simple case, which lays the basis for future work that includes tie strength.

5 Spreading on Social Networks

The standard performance metrics for machine learning methods reflect the goodness of a trained classifier only at the level of pairs, i.e. the correct prediction of single links. These metrics give an indication of the usability of the classifier for applications that target a particular class. However, they do not consider the network structure that emerges through the combination of many of such pair-level predictions. To do this, one must consider the influence the complex network structure has on the dynamics of processes that unfold on the network. Of course, the complex interaction structure leads to different consequences given different processes on the network. In this section, we propose to measure the quality of classification by modelling the initial breakout of a disease in the student population, simulated with a SIR model.

We have used the Gillespie algorithm to simulate the evolution of a SIR model on five different networks: (i) the actual offline contact network ($BT(1)$ for February 2014), as well as (ii) the network predicted by the Random Forest Classifier (for the same target variable), (iii) the network of pairs that interact on Facebook, (iv) the Erdös-Renyi random graph with the same number of nodes and average degree as the actual network (also called 'the random network' in the following), and (v) the configuration model with the exact degree distribution of the actual network. Both (iv) and (v) are synthetic networks that serve as different benchmarks or 'null models'. Furthermore, we compare against the Facebook interaction network, to investigate more carefully the benefit of using our trained classifier rather than the simple sum of all interacting pairs on Facebook [23].

It is important to note that the actual and predicted networks refer to the target variable of meeting at least once during February 2014. The actual network corresponds to the true class labels; the predicted network is put together from the prediction on the independent test sets of each of the 10 cross-validation folds, with a classification threshold set at 0.5. All pairs that are predicted to belong to the 1-class are included in the network, whereas the 0-class is deemed not to interact.

We find that a simulation on the predicted network can accurately predict the total number of infected for nearly all values of R_0, in particular the critical threshold is similar for both networks. Furthermore, comparisons with the random and configuration model network indicate that the disease spread is fundamentally driven by the degree distribution for large R_0, whereas for slower transmission other aspects of the topology play the most important role.

5.1 Comparison of the Network Structure

An essential first step to compare the 5 networks defined above, is by looking at standard, aggregate network properties. Figure 21 shows the degree histogram of the predicted and actual network, and Table 8 lists a few structural properties of these networks. The structural properties of the actual and predicted network are quite well-aligned for all three target variables. The classifier slightly over-predicts the number of edges in the largest connected component of the network for BT(1) and OH, but under-predicts EVE[24]. However, the average clustering coefficient, the diameter of the largest connected component (LCC) and the average shortest path in the LCC all closely match between the actual and predicted networks. Furthermore, a Kolmogorov-Smirnov-test on the actual and predicted degree distribution for $BT(1)$ has a KS-statistic of 0.13, and p-value of 0.90, i.e. we can not reject the hypothesis that the samples of node-degrees were drawn from the same distribution. Together these results offer an important first indication that our predicted network closely resembles the actual structure of offline interaction.

[23] The network of all people who interact on Facebook corresponds to the network that would be predicted by the Majority Vote Classifier trained for BT(1).

[24] The amount of overlap (45%) of edges in the LCC of the EVE network is even lower than the precision of the classifier for the total network, which suggests some parts of the LCC may have become disconnected in our prediction.

By contrast, there are some important structural differences between the actual and synthetic networks. Table 9 shows that - by construction - the Erdös-Renyi random graph and the configuration model networks are much less clustered than the actual BT(1), predicted BT(1), and Facebook interaction networks (featuring average clustering coefficients of 0.01 and 0.04 rather than 0.2). Furthermore, the diameter - the maximum shortest path between any two nodes - in the largest connected component of these real-world networks is much bigger than for the random and configuration model (11 rather than 7 or 8 respectively). It is not surprising that the real-world networks feature more local structure than the synthetic graphs (see section 1.2). However, these local differences will critically affect the early stages of the disease spread. Nodes with a small degree will have a slower transmission rate per node than higher degree nodes, and clusters may trap the disease in a remote region of the network - surrounding infected individuals with recovered ones before they have a chance to spread the disease to others.

These results indicate that the predicted network more closely resembles the actual network than the other models. However, yet again we run into problems of comparability. What does it mean that the two networks have the same average degree, or that the degree distributions are not significantly different? Only a dynamic simulation can determine what combination of structural differences has a marked effect on the quality of prediction of an epidemiological application.

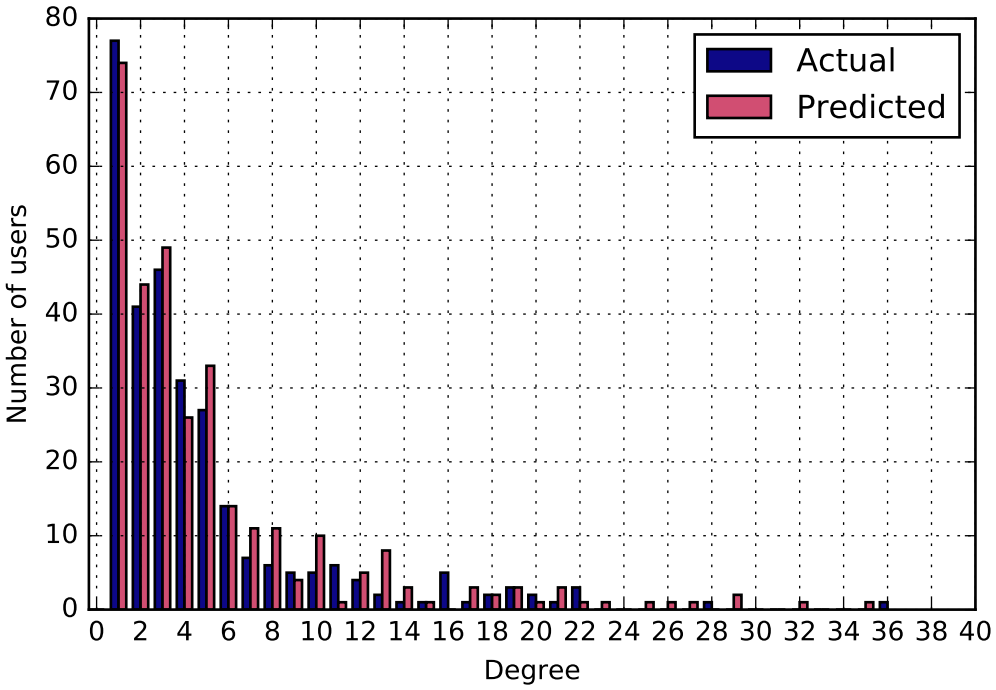


Figure 21: The degree histograms of the predicted and actual networks. The network of users meeting at least once in February 2014 was predicted using the Random Forest Classifier from section 4.2.1.

	Actual 'BT'	Predicted 'BT'	Actual 'OH'	Predicted 'OH'	Actual 'EVE'	Predicted 'EVE'
Average clustering coefficient	0.19	0.18	0.20	0.22	0.19	0.20
Largest connected component (LCC)	292 (90%)	315 (94%)	252 (87%)	275 (91%)	172 (76%)	156 (72%)
Diameter of LCC	11	11	12	14	9	9
Average shortest path in LCC	4.2	4.0	4.4	4.3	3.9	4.1
# Edges in LCC	691	835	563	634	358	277
Overlap	621 (89% of actual)		423 (67% of actual)		162 (45% of actual)	
# Nodes in LCC	292	315	252	275	172	156
Overlap	281 (96% of actual)		221 (88% of actual)		113 (66% of actual)	

Table 8: Structural network properties of the predicted and actual networks. 'BT' stands for Bluetooth, and designates the network of people that meet at least once in a given month. 'OH' is off-hour, and 'EVE' is evenings. All results are for the month February 2014.

	Actual	Predicted	FB interaction	Random	Configuration
Largest connected component (LCC)	292	315	345	288	286
Average clustering coefficient of LCC	0.21	0.20	0.18	0.01	0.04
Diameter of LCC	11	11	13	7	8
Average shortest path in LCC	4.2	4.0	4.0	3.7	3.6
Mean degree in LCC	4.7	5.3	5.5	4.7	4.7

Table 9: Structural network properties of the actual, predicted, Facebook interaction, random, and configuration model networks. The random and configuration model networks are much less clustered (by design). Furthermore, the Facebook interaction network has 10-20% more nodes than any of the other networks. Lastly, the diameter - the maximum shortest path between any two nodes - in the largest connected component of the real-world networks is much bigger than for the random and configuration model. All results are for the month February 2014.

5.2 Simulating Disease Spread

5.2.1 Implementation

To simulate the evolution of a SIR model, we implemented the Gillespie algorithm as described in section 2.3.2. To verify that this implementation was correct, we compare the results of an ensemble of Gillespie simulations on a fully connected graph against a numerical simulation of the mean-field ODE's (equations 2.11, 2.12, 2.13). The fully connected graph was used for comparison because this should behave equivalent to a well-mixed system. Both the Gillespie and the numerical ODE simulation are initialised with the same parameters: $N = 1000$, $I(0) = 10$, $R_0 = 6.0$. The result is shown in Figure 22: the theoretical results lie within the confidence bands of the Gillespie simulation for all times. This shows that our Gillespie algorithm correctly simulates the dynamics of disease spread.

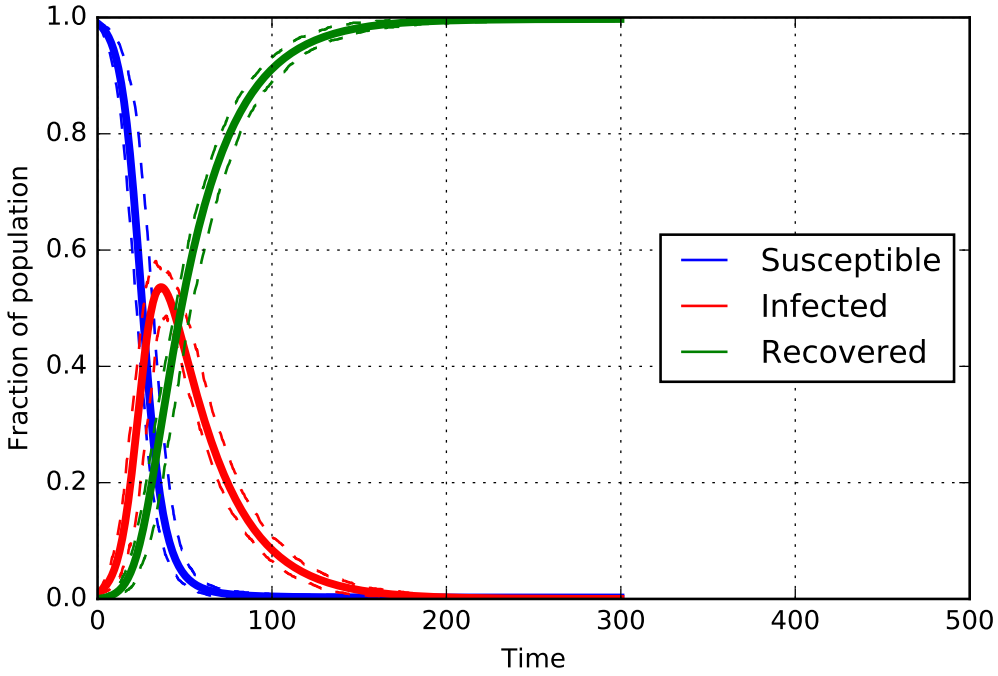
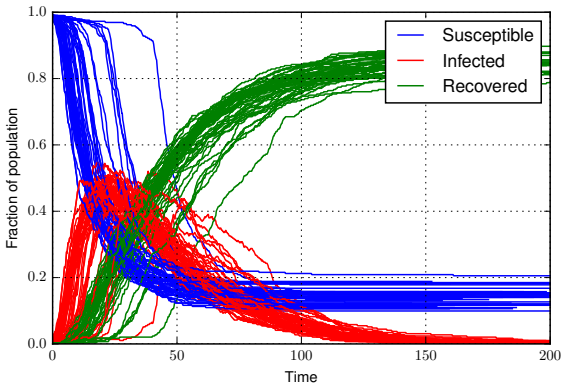


Figure 22: The dynamics of infection on a fully connected graph. Here $N = 1000$, $I(0) = 10$, $R_0 = 6.0$. The result of the numerical integration of the the mean-field ODE's (equations 2.11, 2.12, 2.13) is shown as a thick solid line. The result of 200 runs of the Gillespie simulation is shown as the thin solid line, with dashed 95% confidence bands. The theoretical results lie within the confidence bands of the Gillespie simulation for all times.

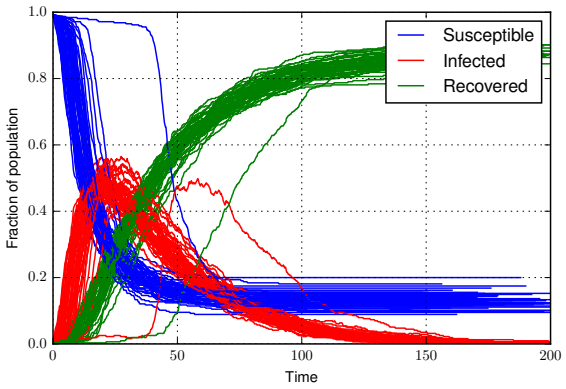
5.2.2 Simulation Results and Discussion

After comparing the static network structure, we now proceed to the comparison of dynamic simulations of epidemics on the 5 networks defined above. First of all, we can visually compare the infection curves, i.e. the time-development of the disease on the network. In Figure 23 it is clear that the infection curves are similar for all networks, but there are important visible differences[25]. Compared to the simulations on the actual network (Figure 23a), the simulations on the predicted network (Figure 23b) reach a slightly higher maximum (5%) slightly earlier (< 10 time-steps). However, the total number of recovered individuals lies very close to the actual value. Furthermore, both networks

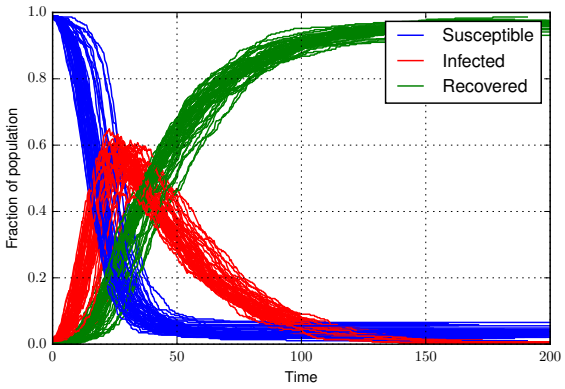
[25]Note that the values used to simulate these infection curves are different than those used in the comparison of ensemble averages. Therefore the timescales here are much shorter than in the later Figures.



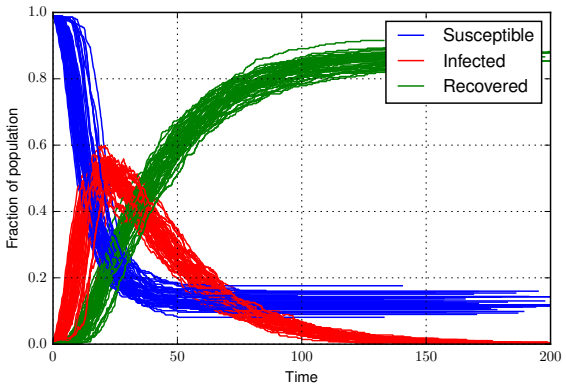
(a) Actual Network.



(b) Predicted Network.



(c) Random Network.



(d) Configuration Network.

Figure 23: Examples of the infection curve for the (a) actual, (b) predicted, (c) random, and (d) configuration model networks. The simulations were initialised with $I(0) = 3$. The transmission per S-I link is $\beta_l = 0.074$, which corresponds to an $R_0 = 10.0$ on the actual network. The infection curve is somewhat steeper for the configuration network than for the actual, and the fraction of infection at maximum is higher.

show a few cases where the disease spread took off very slowly - as characterised by the delayed infection curve. Here the disease was likely initially trapped in an area of the network with only low degree nodes, thus keeping the transmission rates equally low. The simulations on the predicted network perform better than those on the Erdös-Renyi random graph (Figure 23c), which show a much higher maximum of the infection curve than simulations on the actual and predicted networks. Furthermore, at the end of the epidemic (i.e. where the infection curve reaches 0), the population of the random network is almost fully recovered. This means that nearly every individual has contracted the disease in course of the epidemic, an effect which is not observed for the actual, predicted, or configuration model networks.

In the following, we are interested in the mean behaviour of an ensemble of simulations on the network[26]. We randomise over the choice of the initial infected node, to simulate the case where we do not know which individual will be patient zero. Taking into account only the mean could clearly overlook other characteristics of simulations on the network, such as the variability between runs of the ensemble. More regular structures - such as the random model, where the dynamics of the simulation depend less on the initial condition, are likely to have a smaller variance of the distribution of infection curves. However, characterising the behaviour of the mean of the distribution is clearly the first step before investigating higher order moments.

To compare the disease spread we want to determine (i) which fraction of the network becomes infected (I_{tot}), (ii) how fast the disease is eradicated (T_e), and (iii) whether an epidemic will be critical

[26] In the following we consider an ensemble of size 300.

or not given a value for the transmission per link β_l. To do this, we compute the quantities I_{tot} and T_e for each run and find their ensemble averages $\langle I_{tot} \rangle$ and $\langle T_e \rangle$. Figures 24 and 25 display the values of $\langle I_{tot} \rangle$ and $\langle T_e \rangle$ for each of the five networks. The results shown correspond to a range of different β_l values, while the recovery rate was kept constant at $\gamma = 0.001$.

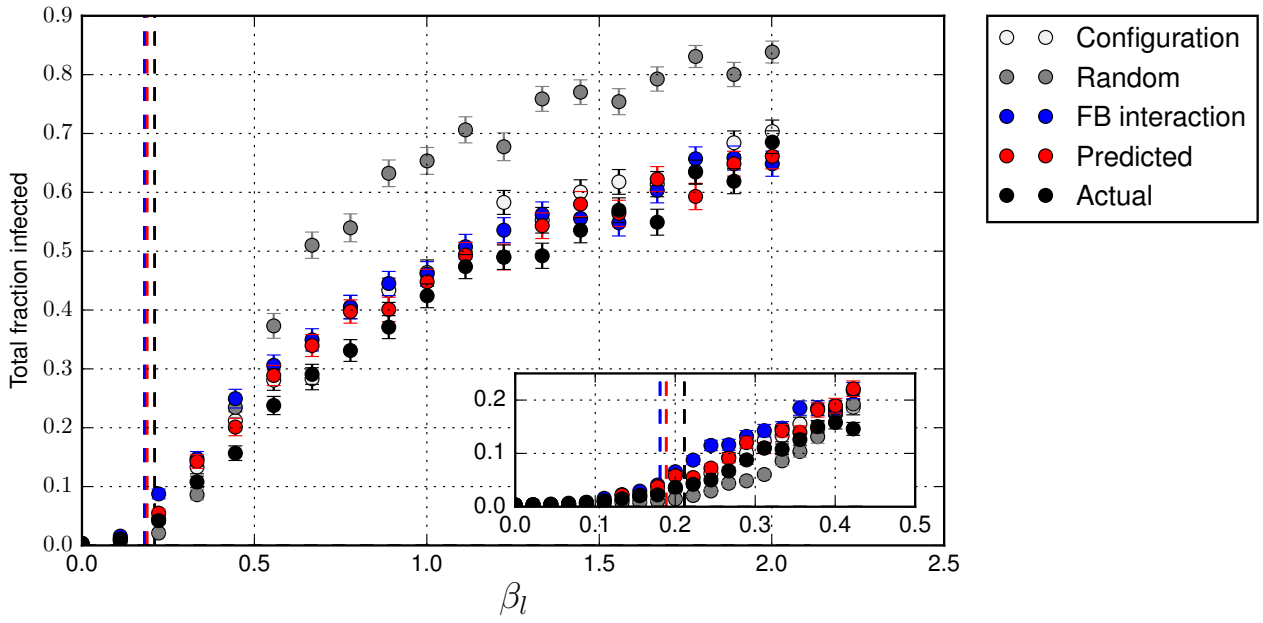


Figure 24: The total fraction of infected individuals, shown as a function of the transmission rate per link. The disease statistic was averaged over 300 disease simulations. The β_l values listed are 10^{-3}, and correspond to $R_0 = [0.0, 10.0]$ on the real network (the inset covers $R_0 = [0.0, 2.0]$). The errors are standard errors of the mean.

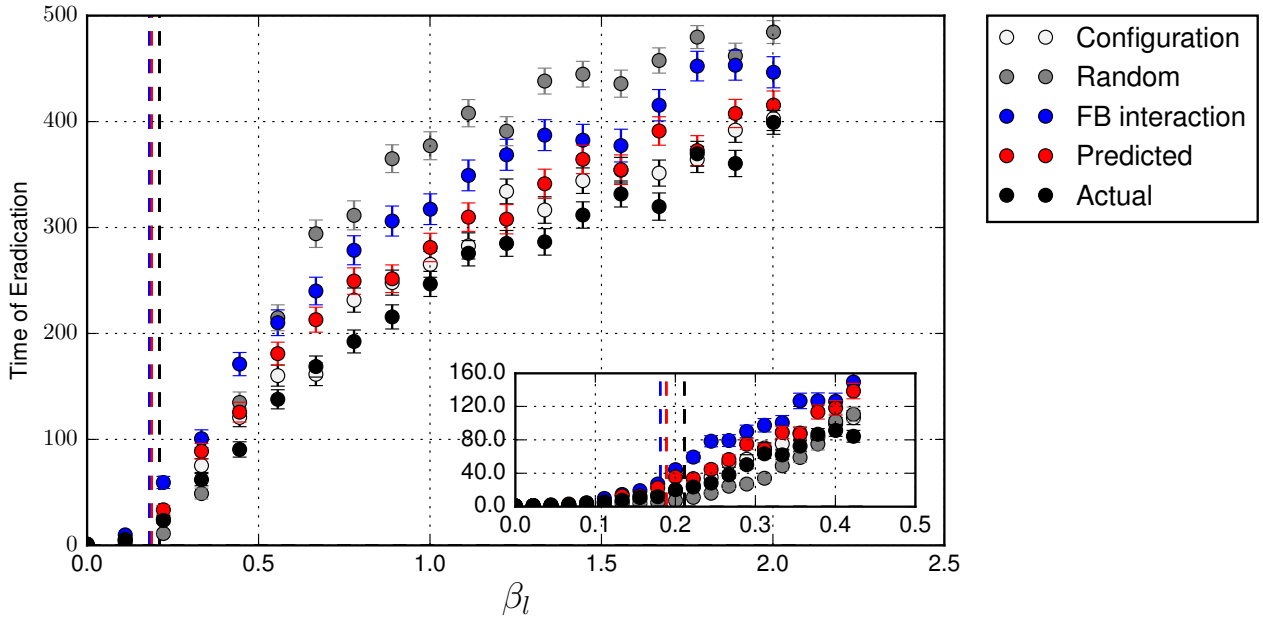


Figure 25: The time until eradication of the disease, shown as a function of the transmission rate per link. The disease statistic was averaged over 300 disease simulations. The β_l values listed are 10^{-3}, and correspond to $R_0 = [0.0, 10.0]$ on the real network (the inset covers $R_0 = [0.0, 2.0]$). The errors are standard errors of the mean.

From Figure 24 we see that a simulation on the predicted network can accurately predict $\langle I_{tot} \rangle$ for

nearly all values of the transmission rate β_l. The predicted network generally predicts slightly larger infections, which are eradicated some tens of time-steps later. For most values of β_l there is a significant difference between the actual and all other networks. However, for the predicted and configuration model networks these differences are small. This becomes more clear when looking at the relative difference between the actual and all other predictions. Let x be an ensemble average, e.g. $\langle I_{tot} \rangle$, on one of the comparison networks. Then the relative difference between x and the same statistic for the actual network x_{act} is:

$$x_{rel} = \frac{x - x_{act}}{x_{act}} \tag{5.1}$$

where - using Gaussian error propagation [85] - the error is given by:

$$\sigma_{x_{rel}} = \sqrt{\sigma_x^2 + \sigma_{x_{act}}^2 \left(\frac{x}{x_{act}^2} \right)^2} \tag{5.2}$$

The relative difference is shown in Figure 26. For $\beta_l > 0.8$ ($R_0 > 4$ on the actual network) the relative difference between $\langle I_{tot} \rangle$ on the predicted and actual network is less than 10 % and stays mostly constant. Around the epidemic threshold of $R_0 = 1$ on the actual network, the difference is up to 40% of the total fraction of infected on the actual network. The simulations on the predicted network consistently predict a higher proportion of infected individuals, i.e. the simulations become supercritical for lower values of the transmission per link. An explanation for this observation could be that the $R_0 = 1$ on the predicted network is reached for smaller values of β_l because the average degree $\langle k \rangle$ is higher (see Table 9).

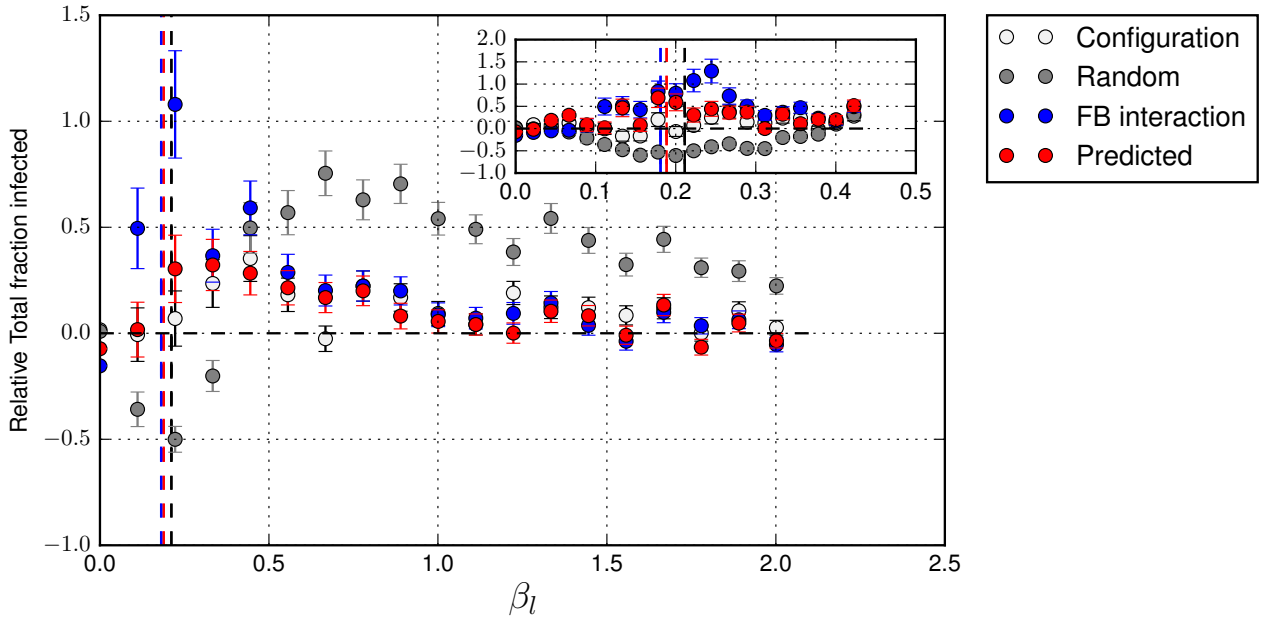


Figure 26: The relative difference in the total fraction of infected individuals, shown as a function of the transmission rate per link. The difference between the actual and other networks are shown for the total fraction of infected individuals. The statistic was averaged over 300 disease simulations. The β_l values listed are 10^{-3}, and correspond to $R_0 = [0.0, 10.0]$ on the real network (the inset covers $R_0 = [0.0, 2.0]$). The errors are standard errors of the mean. The vertical dashed lines correspond to R_0 on the different networks (from left to right: Facebook, predicted, actual respectively).

To interpret the strength of the above results, we compare against simulations on other networks. First of all, to verify the benefit of training a classifier rather than using the raw Facebook interactions, we compare against the network distilled from these interactions. In Figure 24 we see that the

estimate of $\langle I_{tot} \rangle$ on the Facebook interaction network does not significantly differ from the results on the predicted network. However, Figure 25 shows that the time of eradication is overestimated by a wide margin for all values of R_0. This presumably stems from the fact that the Facebook network is much bigger than the face-to-face network (see Table 9). Lastly, Figure 26 shows that the simulations based on the Facebook interaction network more strongly deviate around the critical transition.

Secondly, we compared against the Erdös-Renyi random graph with the same number of nodes and the same average degree as the actual network. This random network shows significantly different values of $\langle I_{tot} \rangle$ and $\langle T_e \rangle$ than the actual network, for nearly all values of β_l. In particular, for $\beta_l > 0.4$ simulations on the random network overestimate the effects of the disease, whereas for smaller values of β_l both the number of infected individuals and the time until eradication are underestimated. Thus, the random network exhibits a large subcritical region where the actual disease would be critical ($R_0 > 1$). A failure to predict disease outbreak for these transmission values can have disastrous consequences. Combined, these observations show that the random network fails to capture the aspects of the network structure that govern the spreading dynamics on the actual network. Since the random network mirrors the mean degree of the actual network, these results suggest that higher order information about the topological features of the network is needed to recreate the spreading process in a useful way.

To understand how much of the disease spread is degree driven, the last network we compare to is the configuration model, with the same degree distribution as the actual network. On this network the simulations reach results which do not significantly differ from those on the predicted network for most values of β_l. However, the critical transition around R_0 more closely matches that of the actual network, since both have the same average degree and degree distribution (by construction). These results indicate that the disease spread is fundamentally driven by the degree distribution for most values of R_0, which fits well with our expectation of the processes driving spreading dynamics on networks. Previous research has shown that the degree distribution is a crucial factor to describe the dynamics of spread [7, 52]: as such, it is not surprising that the configuration model closely matches the simulations for the actual network. However, knowledge of the exact degree distribution of a network is hard to acquire. It is more likely that a national health organisation will be able to guess the number of persons in a population and their average degree of interaction accurately, than the full degree distribution. This would lead to the choice of a random model, which could fail to predict disease outbreak, with possibly lethal consequences. Compared to the random model, both our prediction and the simple Facebook interaction network do much better.

The next step in this investigation will be to understand how much the higher order network properties affect more subtle characteristics of disease spread, such as the variability between runs and the dependence on the initial condition. Preliminary investigations suggest that the results of the disease simulation on the predicted network more closely resemble those on the actual network for a higher number of initially infected individuals. This effect is particularly important around the epidemic threshold, since local variations in the predicted network structure are more likely to make the difference between the disease dying out or spreading to a sizeable fraction of the network. A further valuable investigation will be to perturb the degree distribution of the configuration model slightly, to see how robust the above results are to noise in the degree distribution.

Additionally, an interesting direction for further work is to compare and combine the results for classification on 9 different months (September 2013 to May 2014). This will allow us to make more general statements about the use of this method for estimating the quality of classification, by comparing the results for several different predicted and actual networks. In this case, we will also study the correlation between the relative quality of prediction as measured by the ROC AUC, and how well the disease spread on the actual network can be approximated on the predicted network.

6 Conclusion and General Discussion

The aim of this thesis was to understand the relationship between online and offline networks, and to use data of online interactions to predict the outcome of spreading processes on the face-to-face network.

In this study, we found that it is possible to predict the occurrence of an offline encounter between an interacting pair on Facebook with an accuracy of 78%. More stringent target variables, such as meeting at least 27 times per month or during off-hours, could be predicted with 69 % accuracy (and a precision of 70% and 72%, respectively). This represents a 19% accuracy increase with respect to the Majority Vote Classifier, which strongly suggest that online interactions contain valuable information about offline ties and the corresponding face-to-face contact structure. Our performance is considerably lower than the 82 % accuracy with which Jones et al. were able to determine a Facebook user's closest friends. This discrepancy is likely due to the use of offline behaviour as target variable, which is considerably more subjected to the noise introduced by chance encounters, than the designation of a close tie. In a study which predicted new contacts based upon CDR data, Wang et al. had a precision of 73.5% when restricting the prediction to nodes that shared common neighbours [12]. Our results already lie within this range, although they can likely be improved by including further aspects of the Facebook interaction data or additional sources of data such as CDR.

Accuracy is one of the most commonly used parameters to describe the performance of a classifier, but it does not take the predicted network structure into regard. Since we are interested in the approximation of the structure of the offline network rather than the prediction of a single tie, new methods were needed to quantify prediction performance. So far, no appropriate performance test for network prediction has been described in the literature. We chose to adopt the dynamic simulation of disease spread, which is widely regarded as one of the most successful applications of the study of dynamics on complex networks [52], as functional validation of our predicted network. Using disease simulations of a SIR model, we found that a simulation on the network predicted by our trained Random Forest Classifier closely approximates the total number of infected on the actual network, for nearly all investigated disease virulences. This predicted network shows dynamic evolution of the disease outbreak that lies closer to that on the actual network and configuration model network, than the network derived from Facebook interactions does, especially around the critical epidemic threshold. Comparison with an Erdös-Renyi random graph shows that the predicted network performs significantly better. Furthermore, simulations on a configuration model network with the same degree distribution as the actual network indicate that the disease spread is fundamentally driven by the degree distribution for large β_l. For smaller values of the transmission rate, higher order structure plays an important role, allowing for highly sensitive sensing of the quality of network prediction.

The proximity of the results of disease simulations on our predicted network to those on the actual network, implies that knowledge of Facebook interaction between individuals allows one to accurately simulate the outbreak of a disease on this population. It is important to study how the results obtained from a student population extend to the Danish population at large (see section 3.2 for a discussion of the data bias in this study). For larger populations, the need to obtain access to the full Facebook interaction data of each individual raises privacy concerns and limits the feasibility of our approach. Additionally, the use of online social network data will be fundamentally limited by the prevalence of Facebook use in the population. If we assume a fraction η_{FB} of the population uses Facebook, we can maximally sample a fraction η_{FB}^2 of the possible communication links in the population. According to a report published by DR Media Research in 2014, almost 3.1 million Danes used Facebook each month (55% of the population) [86], which suggests a full-scale study would sample 30 % of all possible communication links. In younger populations, this percentage is likely to be even higher. As such, this limitation should not keep us from further investigating the applicability of online social network data.

6.1 Future Directions

The current work can be extended in a number of ways. Concerning the classification of offline behaviours, it is an obvious choice to further investigate the effect different types of Facebook use have on the prediction of offline contact (as mentioned in section 4.1.1). Additionally, it would be interesting to study the role of individual features and Facebook profile information vs. the interaction between a pair. This might have significant impact on the classification of types of Facebook relations, and assigning offline meanings to them. Secondly, by including additional sources of information such as CDR data, the applicability of our predictions of offline contact can be extended to include a larger proportion of the population. Our current focus on the prediction of offline contact between pairs who interact on Facebook, may namely paint a biased picture of what happens offline, since there may be many pairs who did not interact online, but did meet each other face-to-face. Lastly, with respect to the evaluation of the quality of network prediction, it will be essential to take further account of the variability of simulations on less regular networks, and to investigate the critical behaviour for small per link transmission rates. Furthermore, the robustness of the classification should be tested by predicting on several independent datasets (months).

Since a fully time-aware investigation would have exceeded the scope of this thesis, this study has not taken into account the dynamic nature of the social network. However, the high temporal resolution of both on- and offline interactions is one of the most exciting aspects of the Copenhagen Network Study dataset. Taking causality, or temporality - i.e *when* a pair will meet next - into account while investigating the prediction of offline meetings, could give vital insights about the co-evolution of Facebook and face-to-face interactions, and may be crucial for the correct prediction of dynamics on the social network (as discussed in section 1.2). A first starting point would be to compare our simulations of offline disease spreading against a simulation that takes into account the full temporal structure of the offline contact network (which is the closest approximation we can make to reality).

Lastly, the investigations in this thesis originated in the study of friendships and other pair-based interactions. However, there is great potential in taking a more network-based research approach to study changes in the social network at a grander scale. For example, Peel and Clauset have shown that it is possible to detect e.g. the occurrence of a social event in a freshman population, based on qualitative change in the social network structure [87]. Other alternatives would be to investigate the possibility to detect ‘waves’ or phases of friendship formation, increased off-hour Bluetooth interaction, or exploratory spatial behaviour [32].

7 Acknowledgments

First, I would like to thank Dr. Olivia Woolley-Meza for the many engaging discussions, as well as suggestions and feedback to my work and the contents of this thesis. I look forward to continuing to work with her and to bring this research to the next level.

Further, I would like to thank Prof. Dirk Helbing for the opportunity to work in his group, and for giving me the scientific freedom that enabled me to research and thrive at ETH. Then I would like to thank the entire COSS team, in particular Dietmar, Isa, Caleb, Lloyd, Matthias, and Nino, for the pleasant working environment, and their help with a wide range of research and Switzerland related questions.

Third, I would like to thank Prof. Sune Lehmann, whose unfailing optimism and kind words of support never failed to reach me in the cloud. Also, I would like to thank his team at DTU, especially Radu, Enys, Piotr, Daniel, and Linards for useful discussions about the dataset, and for sharing their expertise about its analysis.

Lastly, I would like to thank my friends, here and abroad, who have allowed me to create 'home' all over Europe; my sister Helle, for being an amazing human being, whose existence helps ground me; my parents for being a safe-haven and sounding board throughout my study and writing this thesis; and my girlfriend Vera whom I love.

8 References

[1] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[2] Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

[3] Dirk Brockmann and Dirk Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 2013, forthcoming.

[4] Lijun Sun, Kay W Axhausen, Der-Horng Lee, and Manuel Cebrian. Efficient detection of contagious outbreaks in massive metropolitan encounter networks. *Scientific reports*, 4, 2014.

[5] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[6] Marcel Salathe, Linus Bengtsson, Todd J Bodnar, Devon D Brewer, John S Brownstein, Caroline Buckee, Ellsworth M Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L Mabry, et al. Digital epidemiology. *PLoS Comput Biol*, 8(7):e1002616, 2012.

[7] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W Feldman, and James H Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

[8] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.

[9] Piotr Sapiezynski, Arkadiusz Stopczynski, David Wind Kofoed, Jure Leskovec, and Sune Lehmann. Inferring human mobility from sparse low accuracy mobile sensing data.

[10] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[11] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[12] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.

[13] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA, 2011. ACM.

[14] Mitra Baratchi, Nirvana Meratnia, and Paul JM Havinga. On the use of mobility data for discovery and description of social ties. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1229–1236. ACM, 2013.

[15] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[16] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978, 2014.

[17] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[18] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[19] Talayeh Aledavood, Sune Lehmann, and Jari Saramäki. On the digital daily cycles of individuals. *arXiv preprint arXiv:1507.08199*, 2015.

[20] European Commision. Commision Staff Working Document - Advancing the Internet of Things in Europe. `https://ec.europa.eu/digital-single-market/en/news/staff-working-document-advancing-internet-things-europe`, 2015. [Online; accessed 19-06-2016].

[21] Bernardo A Huberman, Daniel M Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*, 2008.

[22] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.

[23] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. *Introduction to data mining*, volume 1. Pearson Addison Wesley Boston, 2006.

[24] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM, 2009.

[25] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[26] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[27] Yves-Alexandre de Montjoye, Arkadiusz Stopczynski, Erez Shmueli, Alex Pentland, and Sune Lehmann. The strength of the strongest ties in collaborative problem solving. *Scientific reports*, 4, 2014.

[28] Jason J Jones, Jaime E Settle, Robert M Bond, Christopher J Fariss, Cameron Marlow, and James H Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.

[29] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[30] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[31] Mark T Rivera, Sara B Soderstrom, and Brian Uzzi. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *annual Review of Sociology*, 36:91–115, 2010.

[32] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. The fundamental structures of dynamic social networks. *arXiv preprint arXiv:1506.04704*, 2015.

[33] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[34] Steven H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[35] Bernhard Korte, Jens Vygen, B Korte, and J Vygen. *Combinatorial optimization*, volume 2. Springer, 2012.

[36] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[37] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[38] Luís MA Bettencourt. The origins of scaling in cities. *Science*, 340(6139):1438–1441, 2013.

[39] Michael Batty. The size, scale, and shape of cities. *Science*, 319(5864):769–771, 2008.

[40] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[41] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[42] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[43] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[44] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(4):1–16, 2012.

[45] Arkadiusz Stopczynski, Piotr Sapiezynski, Sune Lehmann, et al. Temporal fidelity in dynamic social networks. *The European Physical Journal B*, 88(10):1–6, 2015.

[46] Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):1–30, 2015.

[47] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[48] Bruno Ribeiro, Nicola Perra, and Andrea Baronchelli. Quantifying the effect of temporal resolution on time-varying networks. *Scientific reports*, 3, 2013.

[49] Lazaros K Gallos, Diego Rybski, Fredrik Liljeros, Shlomo Havlin, and Hernán A Makse. How people interact in evolving online affiliation networks. *Physical Review X*, 2(3):031014, 2012.

[50] Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

[51] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 2001.

[52] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[53] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[54] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[55] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[56] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[57] Samuel Alizon. Co-infection and super-infection models in evolutionary epidemiology. *Interface focus*, 3(6):20130031, 2013.

[58] Weiran Cai, Li Chen, Fakhteh Ghanbarnejad, and Peter Grassberger. Avalanche outbreaks emerging in cooperative contagions. *Nature physics*, 11(11):936–940, 2015.

[59] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170, 2011.

[60] Jérôme Kunegis, Damien Fay, and Christian Bauckhage. Spectral evolution in dynamic networks. *Knowledge and information systems*, 37(1):1–36, 2013.

[61] Zan Huang and Dennis KJ Lin. The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2):286–303, 2009.

[62] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[63] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *SDM?06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[64] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[65] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009.

[66] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):9, 2012.

[67] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*. AAAI, 2009.

[68] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[69] Scikit learn developers. Documentation of scikit-learn. `http://scikit-learn.org/stable/documentation.html`, 2014. [Online; accessed 20-06-2016].

[70] Peter Bühlmann. Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics*, pages 985–1022. Springer, 2012.

[71] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[72] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[73] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[74] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[75] Geoffrey Grimmett and Dominic Welsh. *Probability: an introduction*. Oxford University Press, USA, 2014.

[76] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

[77] Piotr Sapiezynski, Arkadiusz Stopczynski, Radu Gatej, and Sune Lehmann. Tracking human mobility using wifi signals. *PloS one*, 10(7):e0130824, 2015.

[78] Vedran Sekara and Sune Lehmann. The strength of friendship ties in proximity sensor data. *PloS one*, 9(7):e100915, 2014.

[79] Enys Mones, Arkadiusz Stopczynski, Alex Pentland, Nathaniel Hupert, and Sune Lehmann. Vaccination and complex social dynamics. *arXiv preprint arXiv:1603.00910*, 2016.

[80] Arkadiusz Stopczynski, Alex Sandy Pentland, and Sune Lehmann. Physical proximity and spreading in dynamic social networks. *arXiv preprint arXiv:1509.06530*, 2015.

[81] Facebook. The Graph API - Facebook for developers. `https://developers.facebook.com/docs/graph-api`, 2016. [Online; accessed 26-04-2016].

[82] Andrea Cuttone, Sune Lehmann, and Jakob Eg Larsen. Inferring human mobility from sparse low accuracy mobile sensing data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 995–1004. ACM, 2014.

[83] Sensible DTU. API Documentation - Sensible DTU. `https://www.sensible.dtu.dk/wiki/index.php/API_documentation`, 2016. [Online; accessed 24-05-2016].

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[85] University of Rochester. Error Analysis. `http://teacher.pas.rochester.edu/phy121/Laboratory/ErrorAnalysis/ErrorAnalysis.htm`, 2016. [Online; accessed 16-06-2016].

[86] DR Media Research. Media Development 2014. Technical report, 2014.

[87] Leto Peel and Aaron Clauset. Detecting change points in the large-scale structure of evolving networks. *arXiv preprint arXiv:1403.0989*, 2014.