Applications of graph theory to real-world networks

Natalya Levenkova

A thesis in fulfilment of the requirements for the degree of
Masters by Research

School of Mathematics and Statistics
Faculty of Science
March 2014

Acknowledgements

We thank Betsy Foxman for providing the data from the survey [35] conducted in the Seattle area in 2003-2004.

Also we thank UK Data Service for providing access to the Natsal data from UK Data Archive [33].

We thank David Bright from the School of Social Sciences (UNSW) for a opportunity to work on the criminal networks project.

Abstract

We apply graph theory to two problems involving real-world networks. The first problem is to model sexual contact networks, while the second involves criminal networks.

The structure of an underlying sexual contact network is important for the investigation of sexually transmitted infections. Some measures are very difficult to estimate for real-world contact networks. Therefore, mathematical models and simulations can be used for estimating these measures. In this paper we introduce the *spatially embedded evolving network* model. We compare simulated results to real-world data from two surveys against three measures of sexual contact networks: the number of partners; duration of partnerships; gaps and overlaps lengths. We found that each of these measures can be captured independently by our model by choosing suitable values of the input parameters.

Investigation of drug markets and the criminal syndicates groups that operate within them is important in order to target drug law enforcement interventions in the most effective ways. We explore the effectiveness of four different hypothetical intervention strategies that aim to dismantle a criminal network: interventions which target individuals based on degree; interventions which target individuals based on role; interventions which combine the first two strategies; and random intervention. The results of our research shows that the most effective strategy is targeting individuals based on high degree and roles within the networks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preface

In this thesis we consider two applications of graph theory to real-world networks.

The first application is a modelling of a contact network.

The second application is an investigation of the structure of a drug trafficking network.

The first part of this chapter is devoted to an introduction for the thesis. The second part includes definitions of some important terms used in the thesis.

The reminder of the thesis is structured as follows. The first application is the main topic of the thesis and is discussed in Chapters 2 – 5. Chapter Two surveys the epidemiological background of mathematical modelling of contact networks. Chapter Three describes several different random graph models. In Chapter Four we introduce a new discrete model of a random network which can be used as a model of a contact network. We called it the spatially embedded evolving network model (the SEEN model). Chapter Five is on the simulation of a network and contains three subsections. The first

subsection contains parameters for the simulations. The second subsection is a comparison of the simulated data with real-world data. The last subsection investigates some further properties of static binomial case of the model.

Chapter Six describes the the second application and is based on the book chapter [51] by Bright, Greenhill and myself. The main aim of this study is to examine the impact of different strategies of dismantling of criminal networks.

Chapter Seven contains a brief conclusion on both projects.

1.2 Definitions

We now give some definitions which will be needed in the thesis.

1.2.1 Graphs

The main reference used in this section is [28].

A *graph* is a pair $G = (V, E)$ of finite sets such that the elements of E are 2-element subsets of V, where $V(G)$ is a set of *vertices*, also called *nodes* or *points* of graph G and $E(G)$ is a set of *edges*, also called *links* or *connections* of graph G. We write vw to represent an edge between vertices v and w. From this point onwards we will mostly use terms "vertex" and "edge", but sometimes we will also use other names for these notions, for example, "node" and "link" in context of social networks.

Two vertices are *adjacent* or *neighbours* if they are connected by edge.

An *empty graph* is a edgeless graph.

A graph is called a *complete graph* if all its vertices are parwise adjacent.

The *degree* of a vertex v denoted by $\deg(v)$ is the number of edges connected to a vertex or the number of neighbours of v. We also write $deg_G(v)$ if we wish to emphasise

the graph G.

A vertex of degree 0 is called an *isolated vertex*.

The *neighbourhood* $N(v)$ of a vertex v of undirected graph G can be defined as follows:

$$N(v) = \{w \in V : vw \in E\}.$$

The *degree distribution* of a graph G is the probability distribution defined so that the probability of a non-negative integer k is proportional to the number of vertices in G with degree k.

The *mean degree* of a graph G is the number

$$\deg(G) = \frac{\sum\limits_{v \in V} \deg(v)}{|V|}.$$

The *minimum degree* of a graph G is the number $\delta(G) = \min\{\deg(v) : v \in V\}$.

The *maximum degree* of a graph G is the number $\Delta(G) = \max\{\deg(v) : v \in V\}$.

A *path* of length k is a non-empty graph $P = (V, E)$, where

$$V = \{x_0, x_1, ..., x_k\},\ E = \{x_0x_1, x_1x_2, ..., x_{k-1}x_k\}$$

and all x_i are distinct.

A *connected component* is a maximal non-empty sub-graph in a graph G, such as any two vertices in the sub-graph are linked by a path.

The graph $C = P + x_{k-1}x_0$, where $P = x_0...x_{k-1}$ is a path, is called the *cycle* of length k or k-*cycle*, $k \geq 2$.

A *multigraph* (V, E) consists of a finite set V of vertices and a multiset E of edges,

where each edge is a multiset $\{v, w\}$ with $v, w \in V$. If $v = w$ then the edge $\{v, v\}$ is a *loop*.

A *regular ring lattice* is a graph with n vertices $0, 1, \ldots, n-1$ where vertex i is connected to vertex j if and only if $|i - j| \leq k \pmod{n}$. Another name for this graph is the *circulant graph* $C_n^{1,2,\ldots,k}$.

1.2.2 Directed graphs

The main reference used in this section is [10].

A *directed graph* is a pair $G = (V, E)$, where V is a set of vertices and E is a set of ordered pairs of vertices called *directed edges*. We say that u is the *tail* and v is the *head* of the directed edge (u, v). The edges (u, v) and (v, u) may be both present.

A directed edge whose tail and head coincide is called a *loop*.

The *out-degree* of a vertex v (denoted by $\deg^{\text{out}}(v)$) is the number of directed edges in G whose tail is the vertex v.

The *in-degree* of a vertex v (denoted by $\deg^{\text{in}}(v)$) is the number of directed edges in G whose head is the vertex v.

If you take a directed graph and forget the edge directions, you obtain a multigraph where each edge has multiplicity at most 2.

1.2.3 Network theory

The main references used in this section are [35, 58].

A *contact network* is a set of individuals with different patterns of interactions between them. In terms of graph theory we can say that a contact network can be represented as a graph, individuals in this case are vertices and interactions between them are edges.

In the social sciences, the degree of a vertex is called the *degree centrality score* of that vertex. We will use graph-theoretic terminology for this.

We define the *local clustering coefficient* of a vertex v for undirected graphs as

$$C(v) = \frac{2|\{wu : w, u \in N(v), wu \in E\}|}{|N(v)|(|N(v)| - 1)}.$$

In other words, the clustering coefficient of a vertex v is the average probability that two neighbors of v are neighbours themselves.

The *average network clustering coefficient* for all n vertices is defined by

$$\overline{C} = \frac{1}{n} \sum_{v \in V(G)} C(v).$$

We will be particularly interested in sexual contact networks.

The *number of partners* in the case of a sexual network is an equivalent to a vertex degree. Therefore, the *number of partners distribution* is equivalent to the degree distribution.

Consecutive partnerships are defined as partnerships separated in time.

Concurrent partnerships are defined as overlapping partnerships where interaction with one partner occurs simultaneously with interaction/interactions with another partner/partners. In terms of graph theory we can describe concurrent partnerships as follows: two edges are concurrent at the particular moment if they have a vertex in common.

A *gap* is a period of time between two consecutive partnerships. An *overlap* is a negative gap, i.e. the period of time that two concurrent partnerships are overlapped. These concepts are discussed further in Section 2.2.

The *duration of a partnership* is a period of time between the first and the last sexual

contacts between two partners.

1.2.4 Probability theory

The main reference used in this section is [38].

The *expected value* of a discrete random variable X with mass function f is defined as follows:

$$\mathbb{E}(X) = \sum_{x:f(x)>0} x\, f(x)$$

whenever this sum is absolutely convergent.

The *expected value* of a continuous random variable X with density function f is defined as follows:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x$$

whenever this integral exists.

Let A be an event and let $\mathbb{1}_A$ be an *indicator variable* for the event A. That is,

$$\mathbb{1}_A(w) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c \end{cases}$$

The expectation of the indicator variable $\mathbb{1}_A$:

$$\mathbb{E}(\mathbb{1}_A) = 0 \cdot P(A^c) + 1 \cdot P(A) = P(A).$$

The *binomial distribution* $Bin(n, p)$ is a discrete probability distribution with two parameters: n is the number of trials and $p = 1 - q$ is probability of success. The

probability of getting exactly k successes in n trials is given by the probability mass function:

$$f(k) = \binom{n}{k} p^k q^{n-k} \quad \text{if} \quad 0 \leq k \leq n.$$

A *power law distribution* on the nonnegative integers is one in which

$$Pr(X = n) \propto n^{-\alpha} \tag{1.1}$$

for some constant $\alpha > 0$. Often the power law is only assumed to hold in the tail of the distribution, which means that (1.1) only needs to hold for n above some threshold.

The *normal* or *Gaussian distribution* $\mathcal{N}(\mu, \sigma^2)$ is a continuous probability distribution with two parameters: μ is the mean and σ^2 is the variance. It has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

Lognormal distribution $\ln\mathcal{N}(\mu, \sigma^2)$ is a continuous probability distribution with two parameters: μ and σ^2 are the mean and variance of the natural logarithm of the variable. It has density function

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

Chapter 2

Epidemiological background

2.1 Motivation

Epidemiology of HIV and other sexually transmitted infections (STIs) has a long history tightly related to mathematical modeling [6]. Mathematical modelling provides insight into dynamics of infectious diseases which is very important for their control and prevention. One of the factors having impact on the transmission dynamics of STIs is the structure of the underlying sexual contact network [44, 45].

The spread of sexually transmitted infections within a population depends on the various patterns of sexual contacts between individuals. Diversity of human sexual behaviour causes a high variety of such patterns and produce a complex dynamic and heterogeneous network of sexual contacts. Patterns and a contact network itself evolving over time. Epidemiologists use contact networks to explore propagation of sexually transmitted infections.

There are global and local network structure characteristics. For example, degree distribution has effect at the individual level. This characteristic can be measured using

surveys. Clustering appears at the local neighbourhood level, while the number of connected components appears on a network level. The last two measures are very difficult to estimate for real-world contact networks [44]. Therefore, mathematical models and simulations can be used for estimating these measures.

The structure of underlying contact network is extremely important for the investigation of disease transmission [52, 55].

There are some studies on the impact of the degree distribution on disease spread [52, 55]. But there is a lack of models or frameworks which can help to investigate how clustering can impact epidemic dynamics [65].

Mathematical models are valuable and helpful in understanding of complex processes despite of their simplifying assumptions. However, the assumptions should be defined precise and clearly. Even a small set of input parameters can produce complicated behaviour of system. But interpretation of this behaviour can be a challenge. A precisely defined model combined with reliable interpretation of results can be used for a prediction of model behaviour. Mathematical approach should be combined with understanding of epidemiology to gain accurate model [6].

2.2 Measures of contact networks

The structure of a contact network can be described using many parameters. Some of them which are of particular interest for contact networks are listed below.

2.2.1 Clustering coefficient

Clustering describes the network structure within neighbourhood of a vertex. Social networks are often highly clustered [41, 53]. Level of clustering has a great impact on the

speed of disease transmission [65].

2.2.2 The number of connected components

If network consists of more than one connected component, the epidemic will spread within the connected component(s) which contain initially infected vertices and will not spread to the connected components without them.

2.2.3 The number of partners

Both the degree of the vertex itself and the degrees of its neighbours have a great impact on the risk of infection. For example, a monogamous individual might be in contact with a individual who has many contacts and belongs to a high-risk group. So, degree distribution reflects one of the most important factors: heterogeneity of sexual behaviour.

2.2.4 Duration of partnerships

This parameter is of our interest because contacts are more effective for infection spreading in the case of stable relationships [65].

2.2.5 Gaps and overlaps lengths

Gap length is an important factor of spreading STIs through population because the duration of infectivity for some disease can be as long as months or even years. Therefore, infections can spread across population even without concurrent partnerships. In the case of overlaps, infection can easily spread among the partners of infectious person [8, 35, 43].

Let us consider types of gaps and overlaps. See Figure 2.1, adapted from [58].

Four cases are shown for two partnerships, involving a given individual, Partnership A and Partnership B. We have $t_{s_A} < t_{s_B}$. Partnership A starts at time t_{s_A} and ends at time t_{e_A}, similarly for Partnership B. In the first case, there are two partnerships with a gap between them. The next situation shows consecutive partnerships, with zero gap. In the third case, we have concurrent partnerships with complete containment. The overlap in this case is the length of Partnership B, while the gap is $t_{s_B} - t_{e_A}$, a negative quantity. In the fourth case, two partnerships are concurrent without complete containment. The overlap is $t_{e_A} - t_{s_B}$ and the gap, which is negative, is $t_{s_B} - t_{e_A}$.

Therefore, positive or zero gap lengths between two partnerships for the pair of individuals characterise consecutive partnerships. Negative gap lengths characterise concurrent partnerships where two partnerships either overlap or one partnership fully contains another one.

Patterns for three and more partnerships are more complex.

Gaps are safe only if they are longer than the infectious period. This is particularly important for HIV when an individual can spread infection during an indefinitely long period with different infectious rates. The duration of infectivity for some diseases can be as long as months or even years. The more concurrent partnerships exist in a network, the faster an epidemic grows [50].

2.3 Existing models

There is a huge amount of literature on different mathematical models of contact networks. In this section we will give a brief idea of some types of models. Researchers use different types of network models to study disease dynamics. All of them have their advantages and disadvantages.

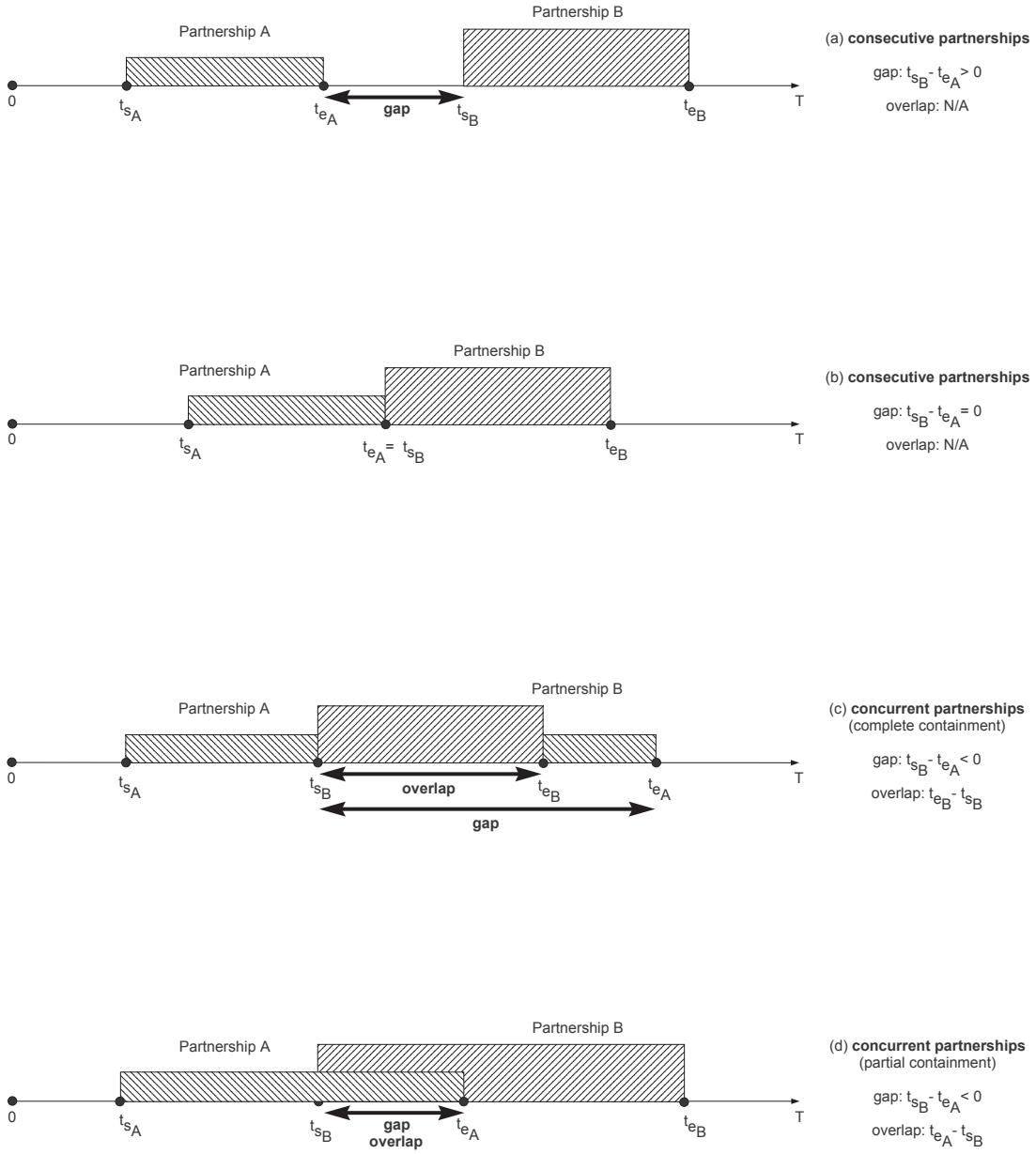


Figure 2.1: Gaps and overlaps

2.3.1 Agent-based models

Agent-based or individual-based models (IBMs) are used extensively in different fields such as real-world business process, ecology, social network analysis and epidemiology [60].

IBMs are based on computer simulations of individuals, their characteristics, interactions and behaviours [21]. Therefore, micro-level parameters allow to simulate a behaviour of the whole system on macro level. IBMs give a complex description of a system and provide possibility to tune the number an characteristics of individuals.

There are some difficulties related to IBMs. It can be hard to find the right characteristics for individuals proceeding from the backgroung of problem. Some micro-level parameters can generate irrational behaviour too complicated to analyze. The micro-level description of individuals goes with simulations of behaviour of many individuals which can be computation intensive.

2.3.2 The SIR model

The classic epidemic model of the spread of a disease over a network is the SIR model [53]. In this model the population is divided into three classes:

- S (susceptible) is a state of individuals when they do not have the disease but can get it from infective individuals;
- I (infective) is a state when they have the disease and can transmit it;
- R (recovered/removed) is a state of individuals which were removed from infective pool on the score of death or immunity.

The fractions s, i and r of individuals in states S, I, R are defined by following expressions:

$$\frac{ds}{dt} = -\beta is,$$

$$\frac{di}{dt} = \beta is - \gamma i,$$

$$\frac{dr}{dt} = \gamma i,$$

where

β is the contact rate (the average number of contacts of an individual sufficient for transmission of disease per unit time),

γ is the recovery rate (infective individuals recover and become immune or removed at this constant rate).

Models of this type are called *fully mixed*, because they make an assumption that population is "fully mixed", that is an infective individual is equally likely to spread the disease to any other individual as they are chosen at random from the whole population. Another assumption is that all individuals have approximately the same number of contacts over the same time period and, hence, the same probability of infection. Both these assumptions are unrealistic because, in real life, disease can only be transmitted in the case of physical contact of some sort between individuals. So, the structure of the contact network is an important aspect of spread of disease.

These models do not take into account duration of contacts and assume that every new partnerships appears with a new partner [7]. Also they do not involve topology and do not represent a real contact pattern.

Models that take into account duration of partnership are called pair-formation models [29]. Individual in these type of models can be in susceptible, infected or removed state and also can be single or in a partnership. Pair-formation models provide a dynamic pro-

cess of pair formation and separation, but they do not take into account concurrentship (an individual cannot have more than one partnership simultaneously).

2.3.3 The SIR model on networks

The SIR model can be generalized to an epidemic taking place on a network. Grassberger in [37] made the first attempt to do it. He showed that the SIR model can be mapped exactly onto *bond percolation* on the same network.

In bond percolation each vertex represents an individual. The edges (bonds) are either occupied or unoccupied, that is, with probability p an edge is occupied, independently for all edges. An occupied edge is able to transmit the disease, while an unoccupied edge is not. For example, consider an outbreak of disease on a network. It starts with a single infected individual and spreads across the network. Edges which are able to transmit disease connect the set of vertices representing the individuals infected in this outbreak and form a connected percolation cluster of occupied edges.

Warren *et al.* [67] considered a more general SIR model for the case where different individuals may have different rates of infection or different recovery rates. They showed that this case can also be mapped to bond percolation.

In one of his papers [52] Newman shows that a large class of SIR models can be solved exactly on networks of different kinds. The resulting model is equivalent to uniform bond percolation on the same network representing the community with some edge occupation probability, which varies because it depends on varying rates of infection and recovery rates.

There is other class of models describing disease spreading, namely, the class of SIS models. The SIS model is a model of endemic diseases, that is, where individuals can

be infected many times and after recovering from the disease becomes susceptible again. The SIS model cannot be solved exactly on a network unlike the SIR model [53].

2.4 Two datasets for real-world sexual networks

Foxman et al. [35] has conducted a random digit dialing survey in the Seattle area in 2003 – 2004. In the survey researchers selected 1051 participants who reported being involved in sexual intercourse. The participants answer the questions on sexual history including gaps, overlaps and lengths of sexual partnerships. The authors used descriptive statistics and graphical representation to describe their results.

Each participant was asked about the lifetime number of partners. The authors state that the cumulative distribution of lifetime number of sex partners approximately follows a Pareto distribution or power law.

Participants were asked about their 5 last partnerships to explore patterns observed in their sexual behaviour. The distribution of the lengths of gaps and overlaps for the most recent partnerships was approximately the same as for previous partnerships. About 25% of relationships were reported as overlapping. The distributions of separately negative and positive gaps are almost exponential.

The cumulative distributions of the lengths of completed relationships follow an exponential distribution. The cumulative distribution of the most recent partnership has a broader tail whereas other the four look similar.

The authors of [35] also explored how the lengths of gaps and overlaps correlate with sociodemographic parameters such as age, income, gender etc.

Another example is a National Survey of Sexual Attitudes and Lifestyles (Natsal 2000), conducted in Britain using computer-assisted interviews [33]. Natsal 2000 sampled

12,110 adults aged 16-44 in 2000. The survey provided new data on sexual behaviour patterns in Britain as well as data on HIV/AIDS propagation. Participants were asked questions about learning about sex, their first sexual experiences, use of contraception, sexual practices and behaviour etc, and socio-demographic questions.

2.5 Discussion

As we mentioned above, the SIR model can be used for describing an epidemics propagation on networks. Therefore, a possible approach to construct a contact network for study of epidemic propagation is a simulation of a model describing the structure of the sexual partnership network. But what kind of network is suitable?

A wide range of real-world networks have been described using random graphs. There are a number of classical random graph models which can be used for building a prototype of contact network. There is a more detailed survey on random graph models in Chapter 3.

Important requirements for the simulated contact network are such characteristics of real-world networks as spatial embedding, degree-inhomogeneous structure, evolution over time [24].

Some properties of different kinds of networks seem to be common to both types, but there are some differences between social and nonsocial networks [54]. Firstly, it has been observed empirically in studies of the network structure that in social networks highly connected nodes tend to be connected to other nodes with high degree. This tendency is called assortative mixing. Secondly, social networks have a higher level of clustering than the corresponding random model. In nonsocial networks highly connected nodes tend to be connected to low degree nodes. So, they show disassortative mixing. The level of clustering in nonsocial networks is no higher than that of the random model.

Newman and Park explain these differences in terms of community structure of social networks. Therefore, model should possess tunable parameters: degree distribution, level of clustering etc [65].

The model also should be kept simple, with few parameters. But even a small set of input parameters can produce complicated behaviour of system.

One of the most important factors is heterogeneity of sexual behaviour. So, a model should take into account this factor in some way. Even a small group of infectious individuals can cause the outbreak of disease if they have high degree in the network [11].

Output results can be compared to survey data, for example [35, 33]. We focus on the two datasets discussed in Section 2.4, particularly, we study the cumulative distribution of lifetime number of sex partners, the cumulative distributions of the lengths of relationships, and the distribution of the lengths of gaps and overlaps [35].

Next, we survey existing random graph models to see if they are suitable.

Chapter 3

Random graph models

In many fields of contemporary science the target complex systems are treated by representing them as graphs. For example, epidemiologists use the different types of contact networks to investigate epidemic propagation [63].

Traditionally, a wide range of networks have been described using random graphs. A random graph is a graph generated by some random process. There are a number of classical random graph models. One of them is the random graph theory of Erdős and Rényi [32]. They proposed using random graphs as a simplified model of communication nets (e.g. road or electric network system) or more complex systems (e.g. organic structures) possessing such properties as growth and inhomogeneity. The random graph model of Erdős and Rényi is based on the assumption that every pair of distinct vertices in the graph is independently connected with probability p [17].

Another approach was proposed by Watts and Strogatz [68]. They started with a regular ring lattice and randomly reconnected the edges with probability p. Investigation of the region between regular p=0 and random p=1 graphs resulted in "small-world"

graphs which were highly clustered like a regular lattice but had short path length like a Erdős-Rényi random graph.

Many real world networks differ from the previous two models in two important aspects: the growing character of real networks and the preferential attachment process ("rich get richer"). In contrast, the Barabási-Albert model defines systems with complex structure which possess these two key characteristics [12, 13].

None of the models mentioned above devote attention to spatial embedding apart from the Watts-Strogatz model, where the nodes of the ring lattice are equally spaced. But distance is also a natural feature of real-world networks. Hence, spatially embedded scale-free networks can be very useful in constructing of real-world network models [24]. We describe some spatially embedded random graph models in Section 3.3.

Undoubtedly, there are more models describing different types of networks. But the models mentioned above are most popular approaches to describe random graphs. More detail on each of these models is given below.

3.1 The classical models

Different disciplines, ranging from natural science to computer science, need precise tools to describe systems with complex topology. Such systems form networks (graphs) where nodes (vertices) represent elements of system and connections (edges) represent interactions between them. Until recently, these networks have been described using random graphs, which were first defined by Erdős and Rényi in 1959. Strictly speaking the first serious attempt to construct random networks was the "random net" of Rapoport and Solomonoff in 1951 [64]. But Erdős and Rényi rediscovered the "random net" independently a decade later, gave it the name "random graph" and studied it rigorously. It

should be mentioned that at the same time Gilbert introduced a random graph model corresponding to the $\mathcal{G}(n,p)$ Erdős-Rényi model [36].

We now describe several classical random graph models.

3.1.1 The Erdős-Rényi model

Random graph theory is based on using probabilistic methods in the study of graphs. Erdős and Rényi proposed using random graphs as a simplified model of communication nets [31]. There are two frequently occuring models of random graphs [17].

The first model $\mathcal{G}(n,M)$ is defined as follows. Let $\Omega(n,M)$ be a set of all graphs on vertex set $V=[n]=\{1,...,n\}$ and with precisely M edges, which can be selected from the $N=\binom{n}{2}$ possible edges. Hence $\Omega(n,M)$ has $\binom{N}{M}$ elements. Then $\mathcal{G}(n,M)$ is the uniform probability space on $\Omega(n,M)$.

If $G_0 \in \Omega(n,M)$, $G \in \mathcal{G}(n,M)$ then

$$\Pr(G=G_0)=\frac{1}{\binom{N}{M}}.$$

The second, more interesting, model assumes that every pair of distinct vertices in graph is connected independently at random. Let $\Omega(n)$ be a set of all graphs on the vertex set $[n]$. Hence, $\Omega((n)$ has 2^N elements. Then $\mathcal{G}(n,p)$ is the probability space on $\Omega(n)$ defines as follows. Each graph $G(n,p)$ is obtained by starting with n vertices and no edges. Each decision about the appearance of the edge is made independently and for each pair of distinct vertices i and j the probability of accepting ij as an edge is equal to p [17, 28]. Here p may be a constant or can depend on n, for example $\mathcal{G}(n,1/n)$.

If $G_0 \in \Omega(n,p)$, $G \in \mathcal{G}(n,p)$ then:

$$\Pr(G = G_0) = p^{M_0} q^{M-M_0},$$

where

$$M_0 = |E(G_0)|, \quad M = \binom{n}{2}, \quad q = 1 - p.$$

In the Erdős-Rényi model, the probability that a vertex has degree k tends to a Poisson distribution with the mean degree λ, as n tends to infinity [5]:

$$\Pr(k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

where

$$\lambda = (n-1)p.$$

Erdős-Rényi random graphs reproduce one of the principal features of real-world networks - the short distance between any two vertices. This feature is often called the "small-world effect". However, in almost all other respects the properties of these random graphs do not match the properties of real-world networks. For example, classical Erdős-Rényi random graphs are constructed on a fixed vertex set when real world networks have a tendency to grow. Hence, we can say that the classical random graph model is not appropriate for describing real-world networks.

3.1.2 Random graph process

A random graph process is one more way to model random graphs. Let $\tilde{\Omega}_n$ be a nested sequence of graphs $G_{n,0} \subset G_{n,1} \subset ... \subset G_{n,N}$, such that $G_{n,t}$ has n vertices and precisely

t edges, and t is time, $t = 0, ..., N$. Hence, $\tilde{\Omega}_n$ has $N!$ elements. Then $\tilde{\mathcal{G}}_n$ is the uniform probability space on $\tilde{\Omega}_n$. To obtain $G_{n,t}$ from $G_{n,t-1}$ we add one edge at a time, which is chosen uniformly at random from all edges not yet present. The distribution on $G_{n,t}$ is uniform on $\tilde{\Omega}_n$ if no extra conditions (e.g., max degree constraint).

Studying random graph processes allows to find out more about the evolution of random graphs. For example, the hitting times of an appearance of different properties [17, 18, 39].

3.1.3 Other models

We mention just a few other graph-theoretical models.

Random regular graphs Denote by $\mathcal{G}(n, d)$ the uniform probability space of d-regular graphs on $[n]$, where $n \in \mathbb{N}$ and dn is even. Every d-regular graph of this space has the same probability, so it is a random d-regular graph. Consider a set of dn points which divided into n parts $v_1, ..., v_n$ of d points. A perfect matching (a set of independent edges) of dn points into pairs is a *pairing*. So, each d-regular graph on $[n]$ corresponds to exactly $(d!)^n$ possible pairings randomly with no loops or multiple edges. This model is called pairing model or configuration model. There is also the pairing model with fixed degree sequence $d_1, ..., d_n$, i.e. each part v_i consist of d_i points [17, 53, 69].

Random bipartite graphs

Denote by $\mathcal{G}(n, n, p)$ the probability space of random bipartite graphs, where n is the cardinality of two disjoint subsets of vertices V_1, V_2. Each graph $G(n, n, p)$ of this space can be obtained as follows. Every pair of vertices $v_i \in V_1$ and $v_j \in V_2$ are connected by an edge independently at random with probability p. Bipartite graphs can be used for

describing networks with two types of vertex, where edges only join vertices of different type [53].

Random planar graphs

A graph is *planar* if it can be drawn in such a way that no two edges intersect. Let $\mathcal{P}_n$ be the probability space of all planar graphs on $[n]$. Then the random planar graph P_n is a graph chosen uniformly at random from this class [28, 48].

3.1.4 Directed random graphs

All the models described above can be defined for the case of directed random graphs. Each vertex of a directed graph has both an in-degree and an out-degree. So, there are also two degree distributions. Directed graphs can be useful for the cases when one needs to take into account the directions of the edges. For example, a directed edge can show a direction of a disease spreading.

3.2 Models of real-world networks

3.2.1 The Watts-Strogatz model

A new approach to modelling networks was proposed by Watts and Strogatz in 1998 [68]. They were inspired by the famous experiment carried out by Stanley Milgram in the 1960s [49]. Some letters were given to the participants of the experiment to be sent to recipients identified by name, address and occupation. The letters passed from person to person to reach a target individual in a small number of steps. The result was that everybody is only six handshakes from anybody else. This fact that any pair of

vertices in most networks are connected by a short path was the first demonstration of the small-world effect.

Watts and Strogatz proposed that topologies of many real-world networks are neither completely regular nor completely random. So, they studied a model of network which lies between these two states. To construct such a network they started with a regular ring lattice with n vertices and k edges per vertex. Then they randomly reconnected every edge independently with probability p over the ring in a particular order and obtained graphs with different levels of regularity.

Figure 3.1 is taken from [68].

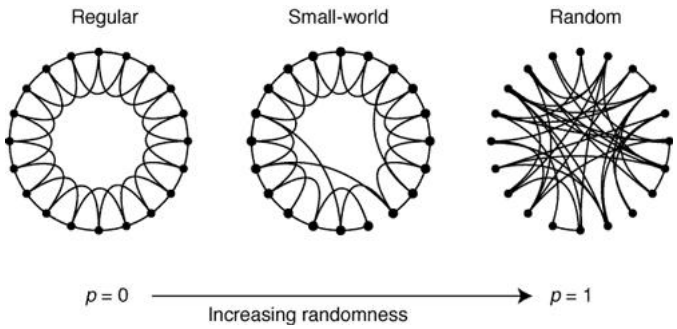


Figure 3.1: Models of networks in the order of increasing randomness

The main result of Watts and Strogatz's study is that for intermediate values of p, not too close to 0 or 1, the graph is a small-world network. Watts and Strogatz defined two characteristics of these graphs: the path length $L(p)$ and the clustering coefficient $C(p)$. Path length is the average for all pairs of vertices of the number of edges in the shortest path between the two vertices. The clustering coefficient is the average number of edges between the neighbours of any vertex. Let k_i be the degree of vertex i. Then $\binom{k_i}{2}$ edges exist if all the neighbours are fully connected. The clustering coefficient shows what proportion of these edges are present. A high clustering coefficient means that the neighbours of a node in the network are likely to be neighbours of each other. In other

words, the friend of your friend is likely also to be your friend.

Thus, Watts and Strogatz found that for intermediate values of p, the graph produced by their method is highly clustered like the original lattice but has small path length like a random graph. That is, most pairs of vertices are not neighbours of each other but can be linked by a small number of steps. Many real world networks possess small-world network characteristics [68].

3.2.2 Scale-free networks

Many real-world networks differ from the previous two models in two important aspects: the tendency of real-world networks to grow in size over time and the process of preferential attachment. These features are exhibited by scale-free networks. The exploration of scale-free networks started with the study of "power-law" behaviour of different kinds of networks, where many nodes have very low degree while a few have very large degree. For instance, the graph of the Internet [34], the graph of citations in academic literature [47, 59], the graph of telephone calls [1]. The probability that vertex in such network has degree k follows a power law distribution: $\Pr(k) \sim k^{-\gamma}$ for some positive constant γ. One of the properties of a power law is a scale invariance. That is, a scaling by a constant of the argument of the power-law function doesn't change the shape of the function. So, a network whose degree distribution follows a power law or, more generally, any heavy-tailed distribution is called a scale-free network.

3.2.3 The Barabási-Albert model

In 1999 Barabási and Albert introduced the concept of scale-free networks [12]. They found that some real world networks such as the World Wide Web, the networks of actors

linked by movies, the citation network in science have a few number of nodes which have many more connections than other nodes. This motivated Barabási and Albert to define a preferential attachment process, which they claimed produced a power-law distribution.

That is, the probability that a new vertex i will be connected to one of the earlier vertices is proportional to the degree of these vertex:

$$\Pr(k_i) = \frac{k_i}{\sum\limits_j k_j}.$$

Hence, in contrast to the classical models and the "small-world" model, a scale-free network grows with time and consists of few nodes which have a high number of neighbours and many nodes connected to a small number of neighbours [12].

But this simplified model cannot be used for practical applications because of the following problems.

Firstly it is impossible to define probability proportional to the degree of the vertices at the first step. Barabási and Albert says that we start with a small number of vertices. That means we start with a graph with no edges, hence, the initial degrees are zero. To address this problem we can choose some non-empty graph instead. In general, the choice of the initial graph plays a significant role. This problem also has been solved by Dorogovtsev, Mendes, Samukhin in [30]. In their model each vertex has an constant "initial attractiveness" a.

The preferential attachment part of the model also has a weakness. Each new vertex has to be connected to m earlier vertices. But Barabási and Albert give their formula without m. But the formula without m cannot be used for each edge added according to the description of the model. Barabási and Albert have solved this problem with assumption that all vertices added at the same step have equal degrees. There is also

a solution without this assumption. It has been given by Krapivsky et al. [42] and independently by Dorogovtsev et al [30]. For the case when a new vertex must be joined to a random set of m earlier vertices the formula is much more complicated and must take into account each of the possible sets of vertices [18].

3.2.4 The LCD model

There is a model which fits the BA model description but is much more precise [19]. To describe this model, firstly, define a random graph process $(G_{1,t})_{t\geq 0}$, where $G_{1,t}$ is a graph on vertex set $v_1, \ldots, v_t$. Start with the empty graph $G_{1,0}$ or with the graph $G_{1,1}$ which has one vertex and one loop. To obtain $G_{1,t}$ from $G_{1,t-1}$ add the vertex v_t with one edge connected to v_i which is chosen randomly with probability:

$$\Pr(i = s) = \begin{cases} \frac{\deg_{G_{1,t-1}}(v_s)}{2t-1} & 1 \leq s \leq t-1 \\ \\ \frac{1}{2t-1} & s = t. \end{cases},$$

Now we define the process $(G_{m,t})_{t\geq 0}$ for the case when we add vertex with $m > 1$ edges. Run the process $G_{1,t}$ on a sequence $v'_1, v'_2, ...$ and form the graph $G_{m,t}$ from $G_{1,mt}$ as follows. Identify the vertices $v'_1, ..., v'_m$ to form v_1, $v'_{m+1}, ..., v'_{2m}$ to form v_2 and so on, allowing multiple edges.

The process $G_{1,t}$ is dynamic, but the distribution of the graph is static. The process $G_{1,t}$ at a particular time t has a static description, the *linearized chord diagram (LCD)* description. An LCD consists of $2n$ distinct points at arbitrary positions along the x-axis, which have been joined in pairs by semi-circular chords contained in the upper half-plane [19]. The conception of LCD is based on n-pairing which is a partition of $[2n]$ into pairs. There are $\frac{(2n)!}{(n!2^n)}$ n-pairings. The LCD model allows to construct $G_1^{(n)}$ without

using the process. To form a graph from an LCD we start with the left and identify all left endpoints up to the first right endpoint and merge them into the first vertex. To form the second vertex we identify all further endpoints up to the second right endpoint, and so on.

Figure 3.2 of an LCD with the corresponding graph is taken from [22].

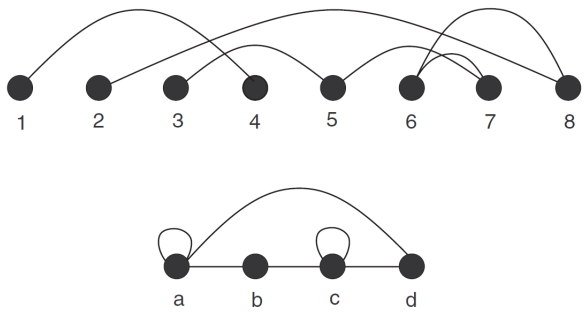


Figure 3.2: An LCD with the corresponding graph

The graph constructed in this way for an LCD chosen uniformly at random from all $\frac{(2n)!}{(n!2^n)}$ LCDs has the same distribution as a random graph $G_1^{(t)}$ [18].

3.2.5 Directed scale-free graphs

As mentioned above, sometimes the graph describing the real-world network should take into account edge directions. Such a model was introduced by Bollobás, Borgs, Chayes and Riordan [20]. The model can be described as follows. The parameters α, β, γ, δ_{in} and δ_{out} are non-negative real numbers such that $\alpha + \beta + \gamma = 1$. Let G_0 be any fixed initial directed graph which grows by adding a single edge at each discrete time step using the following rules.

1. With probability α add a new vertex v and a directed edge from v to an existing vertex w, where w is chosen with probability proportional to $\deg_{in}(w) + \delta_{in}$.

2. With probability β, add an edge from an existing vertex v to an existing vertex w, where v and w are chosen independently with probability proportional accordingly to $\deg_{out}(v) + \delta_{out}$ and $\deg_{in}(w) + \delta_{in}$.

3. With probability γ add a new vertex w and an edge from an existing vertex v to w, where v is chosen with probability proportional to $\deg_{out}(v) + \delta_{out}$.

Bollobás et al prove that the resulting in- and out-degree distributions follow power-law distribution with different exponents.

3.3 Spatially embedded networks

Many real-world networks come with a spatial embedding, as they exist in 3-dimensional euclidean space and have a natural notion of distance. Particularly, in the network of human contacts a connection often means a physical contact, so it is important to take into account a geographical aspect.

None of the models described above devote attention to spatial embedding apart from the Watts-Strogatz model, where the nodes of the ring lattice are equally spaced and have some sort of spatial organisation. The Erdős-Rényi and the Barabási-Albert models do not have any spatial embedding at all.

There exist a number of relatively recent studies with respect to spatial embedding, which we now discuss.

3.3.1 Scale-free networks embedded in lattices

In 2002 – 2003 several very similar models describing methods for embedding scale-free networks in lattices were proposed [9, 61, 67, 70].

For example, the method by Rosenfeld et al. [61] is based on the natural constraint that the total length of links in the network is minimal. The nodes are ordered on a d-dimensional lattice of size R and have an initial degree k taken from the power-law distribution. Each node selected at random connects to its closest neighbours until it has the prescribed degree.

3.3.2 Random geometric graphs

Let $\mathbb{R}^d$ be d-dimensional space. Let $\|\cdot\|$ be some norm on $\mathbb{R}^d$, for instance, the Euclidean norm. Given a finite set $\mathcal{X} \subset \mathbb{R}^d$ and some positive parameter r, denote by $G(\mathcal{X}, r)$ the undirected graph with vertex set $\mathcal{X}$. Each pair of vertices v_i, v_j in the graph connected by an edge if and only if the distance between them is at most r, i.e. $\|v_i - v_j\| \leq r$. The random geometric graph is constructed on vertices distributed at random uniformly and independently on $\mathbb{R}^d$. This model can be more useful and can more realistically describe some types of real-world networks, especially in the cases when the triangle property (if vertex v_i is close to v_j, and v_j is close to v_k, then v_i is close enough to v_k) plays a more important role than the independence of edges which characterises the Erdős-Rényi model [31, 56].

For more information on random geometric graphs see [56, 66].

3.3.3 The SERN model

The newest studies of spatial embedded networks aim to define how spatial embedding affects the structure of networks. Recent research along these lines has been performed by Bullock and coauthors [14, 24].

Bullock et al. [24] recognized the importance of spatial embedding and gave the follow-

ing general definition. A *spatially embedded network* is a network with nodes distributed in some metric space where the probability of a pair of nodes being connected depends on the distance between them. Bullock and coauthors constructed the spatially embedded random networks model (SERN). Nodes in this model are placed independently at random on some metric space according to a specified distribution, not uniformly as in random geometric graphs. For each pair of vertices, the decision about appearance of the edge is made only on the distance between the nodes. The model was created to analyse of the effects of spatial embedding on network structure. The result was that scale-free spatial networks are possible where the node distribution is inhomogeneous. That is, spatially embedded random networks can have scale-free distribution if there is a singularity in the spatial distribution of nodes. So, homogeneity or inhomogeneity of the node distributions impacts on network structure, particularly, on degree distribution [24].

3.4 Time-evolving models

Most of network models described above are static and do not evolve over time. The Barabási-Albert model is the only model which possesses such a feature of real-world networks as growth. But growth in this context means that the number of vertices and edges is increasing. So, we can consider a network built on the principle of preferencial attachment as a cumulative network which stores information about all connections between vertices over time. But the model does not provide a mechanism for edge deletion. Thus, the BA model is not suitable for building a network which evolves over time, where edges can disappear as well as appear.

3.4.1 Mobile geometric graphs

One of evolving models is the dynamic Boolean model introduced by van der Berg et al. in [15]. This model has been extensively studied by Peres et al. [57] with application to wireless networks and it has been named the *mobile geometric graph model*.

The model can be defined as follows. We start with a random geometric graph at time 0 where nodes are distributed as a Poisson point process in $\mathbb{R}^d$ with a fixed intensity. Vertices move independently according to some stochastic process (Brownian motion in this case) in continuous time.

The authors focused on a wireless network application and considered the following properties: detection time, coverage time, percolation time and broadcasting.

3.5 Summary

There are a great number of different models of random graphs and their variations. The major models are described above. Most research on these random graph models has focused on asymptotic properties: that is, properties of the model which hold as the number of vertices tends to infinity.

Each of these models has its advantages and disadvantages and distinguishing features. The current review aims to define which of these models is more suitable to build a prototype of an underlying contact network. The most important factors to consider are spatial embedding, inhomogeneity and time evolution.

The fact that many real-world networks including social networks have spatial organization should not be neglected. For example, sexual contacts are more likely to occur locally within a neighbourhood. Thus, it is very important to take into account a notion of spatial embedding between nodes of network to define a geographical neighbourhood.

Inhomogeneity of a network can become apparent in a few ways. It can be provided by spatial inhomogeneity, that is, where the spatial distribution of nodes is inhomogeneous [24]. Another option is a network with a spatially homogeneous distribution of nodes where inhomogeneity is provided in some other way.

Most real-world social networks are not stable: they evolve over time. Evolution is related to the number of individuals in a network, some of them connect to or disconnect from a network. Evolution also shows on the level of connections between nodes. Connections appear and disappear, and sometimes appear again. All these changes should be represented in a model of contact networks.

Apart from these properties, some analysis of real data have indicated that cumulative degree distribution of different real-world networks follows a power law [16, 35, 46, 62]. However, other research shows that such conclusions might be hasty as the degree distribution may follow other heavy-tailed distributions [26].

Model	Inhomogeneous degrees	Evolving over time	Spatially embedding
Classical models	✓	×	×
Watts-Strogatz	✓	×	✓
Barabási-Albert model	✓	✓	×
SERN model	✓	×	✓
Mobile geometric graphs	✓	✓	✓

Table 3.1: Comparison of some random graph models

As we can see from Table 3.1, most of random graph models do not satisfy the most important factors of real-world networks apart from the mobile geometric graph model. Researchers working on mobile geometric graphs [57] have not studied some aspects relevant to contact network, e.g. cumulative degree distribution. In next Chapter we introduce a new model which is a generalization of mobile geometric graphs, adapted to a discrete setting.

Chapter 4

Spatially embedded evolving network (SEEN) model

4.1 Model description

We now describe a new discrete model of a random network based on ideas of random geometric graphs and random motion of vertices. We focus on the one-dimensional case, since it is efficient for computations and produces a broad range of output behaviour. We generalize random geometric graphs to consider different distribution of radii, not just uniform [56, 66]. To make the network evolve over time we add random motion of vertices. In that case it will be a discrete generalization of mobile geometric graphs. Firstly, we describe the static model and then we describe how it evolves over time.

4.1.1 Static model

Let $d \geq 1$ be an integer and let M be a large fixed even positive integer.

Let L be a circle in 1-dimensional space $\mathbb{R}$, which we take as the interval $[0, M]$ with endpoints identified. A metric on L can be defined as follows. For any two points x_i and x_j on L, let

$$d(x_i, x_j) = \min(|x_i - x_j + u| : u \in (-M, 0, M)).$$

So, we set periodic boundary conditions for the interval $[0, M]$ to be able to simulate points scattered on the circle. That means that when a point passes a boundary (left or right end of the segment), it reappears on the opposite side immediately.

We divide the circle into M disjoint buckets (intervals) of equal lengths (see Figure 4.1). The jth bucket is $[j-1, j)$ for $j = 1, .., M$. Now fix $N \in \mathbb{N}$ with $N << M$. We will place N points on the circle L randomly, as follows. A bucket b_i for the ith point is chosen uniformly at random from $\{1, \ldots, M\}$ independently for all $i = 1, .., N$. Then we calculate the midpoint of the bucket containing x_i:

$$x_i := b_i - \frac{1}{2}.$$

Therefore, we obtain a *position vector* $(x_1, \ldots, x_N)$. Points represent nodes of the future network.

Let $Y : \mathbb{Z}^+ \to [0, 1]$ be a probability distribution on positive integers and let $r_1, .., r_N$ be independent and identically distributed random variables with distribution Y. We call $(r_1, .., r_N)$ the *radius vector* and Y the *radius distribution.*

For each $i \neq j$, let $(x_i, x_j) \in E$ be a directed edge if and only if $d(x_i, x_j) \leq r_i$.

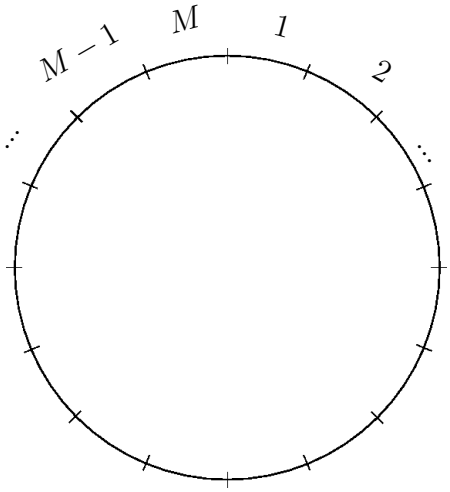


Figure 4.1: A circle divided into M buckets

Thus, we obtain a random directed graph. Then we transform the random directed graph to undirected one on the same vertex set, with edges defined as follows: for all $1 \leq i \neq j \leq N$, let $v_i v_j$ be an edge of the undirected graph if at least one of $(v_i, v_j), (v_j, v_i)$ is an edge of the directed graph (Figure 4.2).

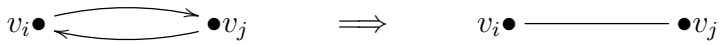


Figure 4.2: Directed edges replacing by a single undirected edge

Note: If radius r_i is greater than $\frac{M}{2}$ then vertex i is connected to all other vertices. In the case that $r_i = r$ for all i we obtain the random geometric graph model.

Therefore, the input parameters for the model are the fixed number of buckets M, the number of nodes N and the probability distribution parameters. Changing the ratio N/M or parameters of the probability distribution, we can change the density of the produced network.

4.1.2 Evolving network

Now we want the network to evolve over time. The process of evolving of the network starts at the time $t = 0$ with N vertices and e_0 edges, where $N, e_0 \in \mathbb{N}$. The graph G_0 obtained at the time step $t = 0$ is the static network described in Section 4.1.1.

Random walk of the vertices induces evolution of network. At each time step, each vertex moves independently one bucket clockwise or anticlockwise with probability $\frac{1}{2}$. So, the vertices change the distance between them with probability $\frac{1}{2}$ at each time step. Therefore, this process leads to appearance of new edges and to disappearance of existing ones, as vertices move in or out of range of other vertices.

Therefore, we obtain a sequence of "snapshots" of the evolving network. Let G_t denote the snapshot at time t. Each "snapshot" contains edges present at current time step only.

Also we consider a sequence of cumulative networks H_t. At each time step t we obtain a cumulative network H_t containing all edges which were present in at least one of $G_0, ..., G_t$.

4.1.3 Example

Let us consider a small example of how our model works. First, we construct a small static network with $N = 4$, $M = 8$ at time step $t = 0$.

Suppose that the random bucket b_i and random radii r_i are as in Table 4.1.

Vertex $\boldsymbol{v_i}$	v_1	v_2	v_3	v_4
Bucket index b_i	3	7	2	5
Radius r_i	1	1	2	2

Table 4.1: Vertex positions and radii

The positions of the vertices are shown in Figure 4.3.

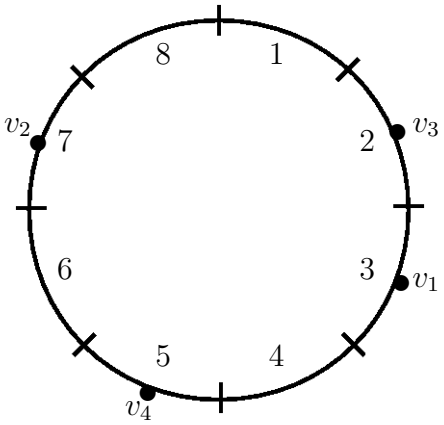


Figure 4.3: The snapshot of the vertex positions in the network G_0 at the time step $t = 0$

Radii of vertices are shown in Figure 4.4. A radius specifies the range of impact of a vertex.

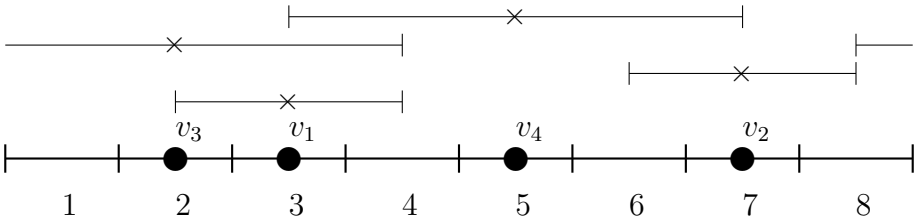


Figure 4.4: The snapshot of the network G_0 at the time step $t = 0$

The vertex v_1 contains the vertex v_3 in its range of impact, the range of the vertex v_2 doesn't contain any vertices, the range of the vertex v_3 contains v_1, and the range of v_4 contains v_1 and v_2.

Therefore, we obtain a directed graph (Figure 4.5).

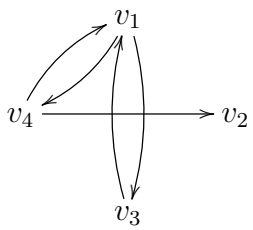


Figure 4.5: The directed graph at the time step $t = 0$

Replacing 2-cycles by a single edge and ignoring directions on all edges, we obtain an undirected graph G_1 (Figure 4.6).

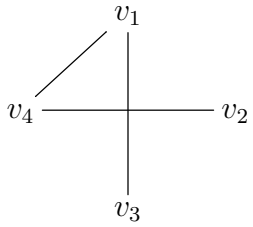


Figure 4.6: The undirected graph G_0 at the time step $t = 0$

Now suppose the vertices have moved at time step $t = 1$ according to Table 4.2.

v_i	v_1	v_2	v_3	v_4
Direction	clockwise	clockwise	anticlockwise	clockwise
New b_i	4	8	1	6

Table 4.2: Directions of vertex movement and new vertex positions

The new positions of the vertices are shown at Figure 4.7.

At time step 1, some vertices have entered or left the range of impact of others.

Radii of vertices are shown in Figure 4.8.

Now the vertex v_1 contains no vertices in its range of impact, the range of the vertex

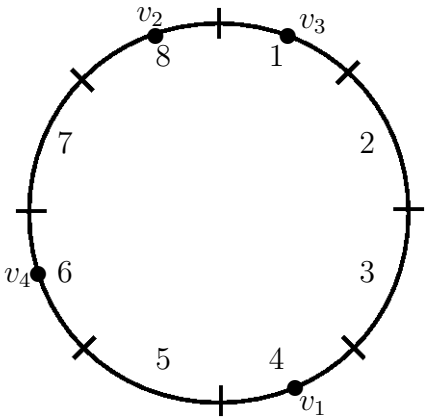


Figure 4.7: The snapshot of the vertex positions in the network G_1 at the time step $t = 1$

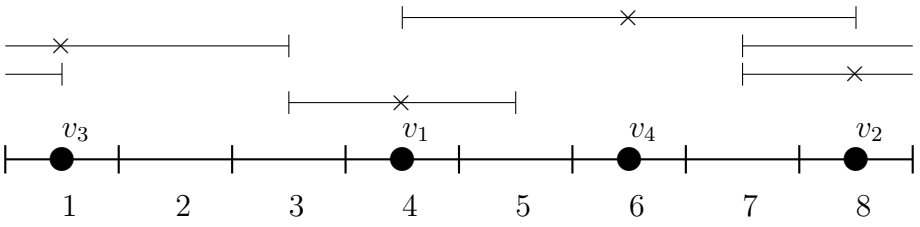


Figure 4.8: The snapshot of the network G_1 at the time step $t = 1$

v_2 contains v_3, the range of the vertex v_3 contains v_2, and the range of v_4 contains v_1 and v_2 as at the previous time step. We obtain a new directed graph (Figure 4.9).

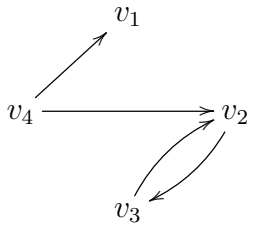


Figure 4.9: The directed graph at the time step $t = 1$

Following the procedure described above we obtain an undirected graph G_2 (Figure 4.10).

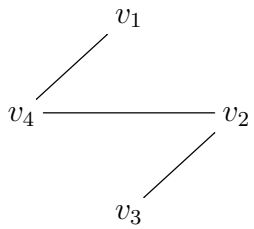


Figure 4.10: The undirected graph G_1 at the time step $t = 1$

From G_0 and G_1 we can produce the cumulative graph H_1 which contains all edges which were present in at least one of G_0 or G_1 (Figure 4.11).

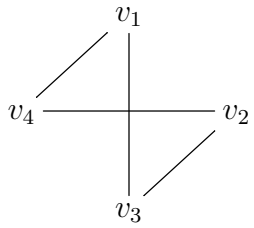


Figure 4.11: The cumulative graph H_1 obtained after two time steps

Chapter 5

Simulation of the network

5.1 Parameters for the simulations

5.1.1 Input parameters

For the simulation we use the following parameters:

- M - the number of buckets for the unit circle
- N - the number of points distributed on the unit circle, $N << M$
- T - the number of time steps
- $(x_1, \ldots, x_N)$ - the position vector
- $(r_1, \ldots, r_N)$ - the radius vector drawn from binomial, lognormal or power-law distribution

The values of the parameters are described in Subsections 5.2 and 5.3.

We run the simulation, producing graphs $G_0, \ldots, G_T$ as described in the previous chapter.

5.1.2 Output measures

Then we calculate the following output measures for the static network ($t = 0$):

- Mean, minimum and maximum degree, mean outdegree
- The number of connected components
- Network average clustering coefficient

We also calculate the following measures for the cumulative network after T time steps:

- Cumulative degree distribution
- Length of partnerships distribution
- Gap/overlap distribution

First we focus on the last three measures, Subsection 5.2 contains the results and comparison with real-world data.

Results on the static output measures for the static binomial model are described in Subsection 5.3

5.2 Comparing simulated and real-world data

In this section we will describe the results of simulations and compare them to real-world data. The simulated data has been obtained according to the SEEN model de-

scribed in Chapter 4 with a radius vector chosen from one of considered distributions, described below.

5.2.1 Methods

We compare the simulated data from the SEEN model to the real-world data against three measures of sexual partnerships: lifetime number of partners, lengths of the most recent or the second most recent partnership, duration of gap or overlap.

For the simulation we use the following parameters. The number of buckets for the unit circle $M = 20480$. The number of points distributed on the unit circle $N = 1024$. A position vector $(x_1, \ldots, x_N)$ was chosen independently at random according to uniform distribution on $\{1, ..., M\}^N$ and was then held fixed throughout. A radius vector $(r_1, \ldots, r_N)$ was chosen independently at random for N vertices from the following distributions:

- Binomial distribution $\mathrm{Bin}(\frac{M}{2}, p)$, where $p \in [\frac{1}{8192}, \frac{1}{2}]$,
- Power-law distribution $\mathrm{P}(\alpha)$, where $\alpha \in [2, 5]$,
- Lognormal distribution $\ln \mathcal{N}(0, \sigma^2)$, where $\sigma \in [0.5, 2.4]$.

Our choice of the ranges of the parameter for each distribution is conditioned on the density of network. (Since M is fixed, each distribution has a single parameter.) Too sparse or too dense networks cannot reproduce characteristics of real-world networks, so we choose parameters which avoid these extremes.

The best way to avoid artefacts of particular choices of radii is to use the Monte Carlo method. Therefore, for each of the above distributions we simulated a hundred vectors of radii, then we sorted them in descending order and calculated the averaged vector of

radii over all hundred vectors. These averaged radii were randomly assigned to vertices. This smooths the output (compared to the results of a single run) while maintaining heterogeneity.

We ran all simulations for the following number of time steps: $T \in \{256, 512, 1024\}$.

Let us introduce a shorthand terminology for the models as follows, indicating both the radii distribution and the number of time steps for the simulation:

- Binomial (p, T)
- Power law (α, T)
- Lognormal (σ, T).

5.2.2 Method of quantifying closeness of the SEEN model to the real-world data

From each of the two real-world data sets described in Section 2.4 we extracted the cumulative distributions for the following measures:

- lifetime number of sex partners (logarithmic scale);
- duration of the most recent and the second most recent sexual partnerships (semilogarithmic scale);
- duration of gap or overlap between the most recent and the second most recent partnerships.

For each combination of values of input parameters in our model, we calculated the number of partners, lengths of the most recent partnership and the second most recent partnership, duration of gap or overlap between the most recent and the second most

recent partnerships for each vertex. By combining these data for all 1024 vertices, we obtained the cumulative distributions for the three measures described above. The last two measures are presented as percentages of the total number of time steps to allow comparison between the simulated and real-world data.

To understand which values of input parameters give results closest to the empirical data, we calculated the differences between areas under the curves of cumulative distributions for the simulated data and for both empirical data sets. Since all our curves are monotonically non-increasing, we can use this difference as a measure of the distance between the output of the model and the real-world data. For each family of distributions we found the values of the input parameters which minimized this distance and hence gave the best approximation to the empirical data.

5.2.3 Lifetime number of sex partners

Figure 5.1 shows the best result for the distributions we considered. The closest approximation for the lifetime number of partners was obtained for Lognormal $(2.4, 256)$ (see Figure 5.1). The best result for power law distribution is Power law $(2.1, 512)$. Binomial radii distribution failed to give any good approximation for all values of input parameters considered, however, the closest result was obtained for Binomial $(\frac{1}{8192}, 256)$.

5.2.4 Length of sexual partnerships

Investigating the lengths of sexual partnerships, we found that for all radii distributions considered, the simulated distribution of length of the most recent partnerships is closer to the empirical distribution of the second most recent partnerships and not to the empirical distribution of the most recent partnerships (see Figure 5.2). Al-

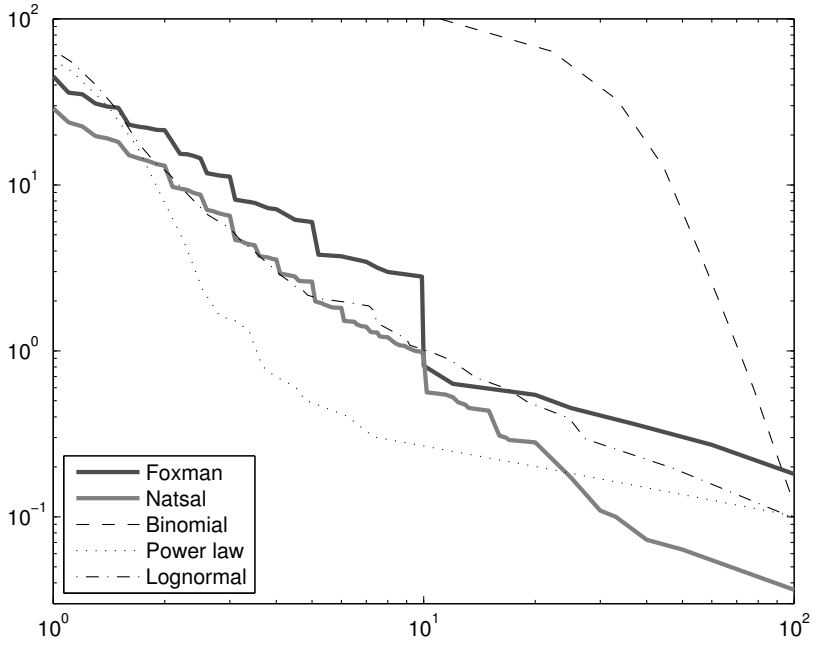


Figure 5.1: Cumulative distribution of lifetime number of sex partners

though we obtained good approximation of the empirical data for all radii distributions: Binomial $(\frac{1}{4096}, 256)$, Power law $(3.4, 256)$, Lognormal $(0.7, 1024)$, the best result overall was for lognormal distribution.

Also we calculated the percentage of short-term (single) partnerships for the same values of input parameters as above. This percentage is 24.74% for the Foxman's data and 22.65% for Natsal 2000. Simulated data gives 30.86% for binomial distribution of radii, 33.79% for power-law distribution and 28% for lognormal distribution with the same parameters as stated in the previous paragraph.

In comparison with the most recent partnerships, the best result has been achieved for Lognormal $(0.5, 512)$ (see Figure 5.3). The best values for the other radii distributions were Binomial $(\frac{1}{128}, 256)$ and Power law $(5, 256)$, though these gave less accurate results.

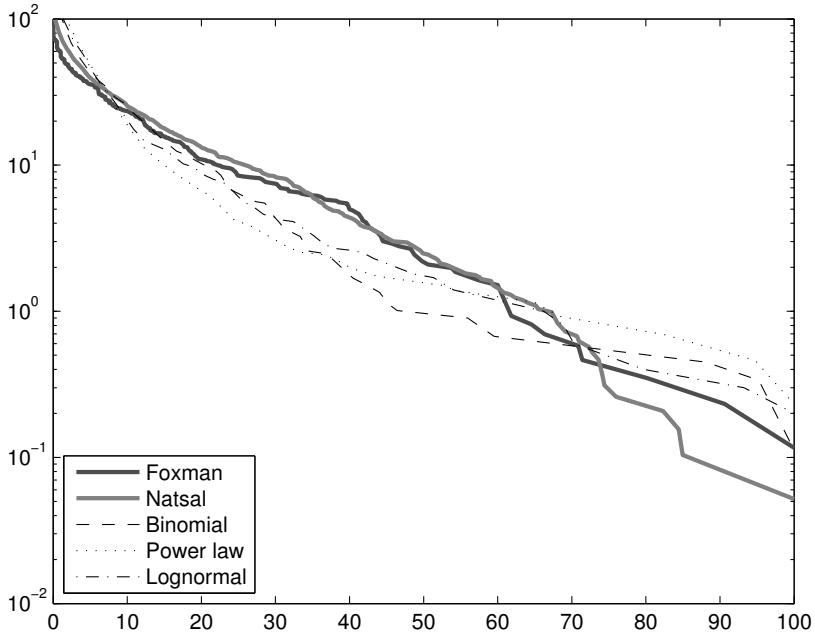


Figure 5.2: Cumulative distribution of length of partnerships, as a percentage of T (compared to the second most recent partnerships in empirical data)

5.2.5 Length of gap or overlap

Figure 5.4 shows the best result for the distributions we considered. Cumulative distributions of length of gap or overlap for the most recent partnerships of the empirical data are almost equally well approximated by all three radii distributions considered: Binomial $(\frac{1}{512}, 256)$, Power law $(2.1, 256)$, Lognormal $(1.6, 256)$. The model with lognormal distribution gave the result which was closest to both empirical data sets simultaneously.

Also we calculated the percentage of overlaps (concurrent partnerships) for the same input parameters as above. This percentage equals to 22.69% for Foxman's data, 14.58% for Natsal, 23.12% for binomial distribution, 12.94% for power law, 15.12% for lognormal. So, the model with binomial distribution of radii produces the closest result to the Foxman's data, For this measure power-law and lognormal distributions better approximate

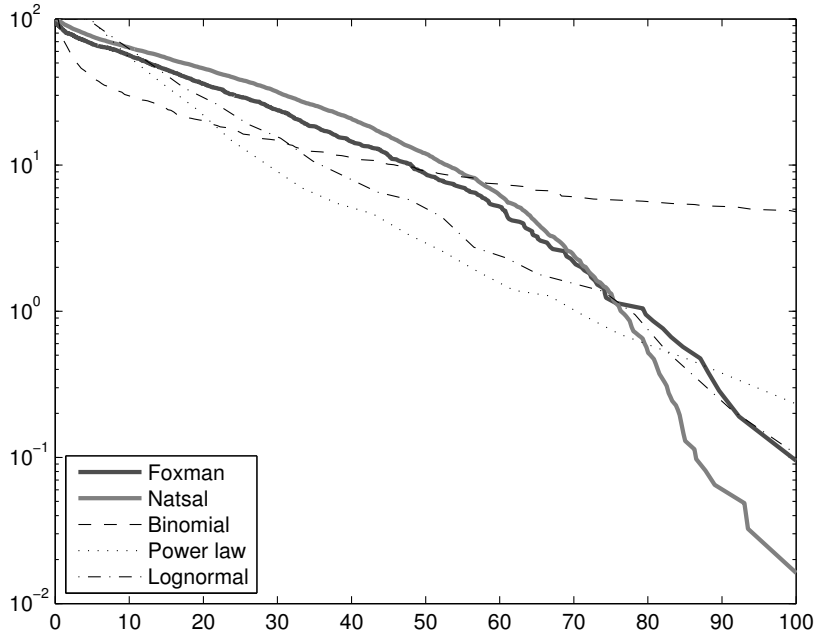


Figure 5.3: Cumulative distribution of length of partnerships, as a percentage of T (compared to the most recent partnerships in empirical data)

the Natsal data.

5.2.6 Discussion

We found that each of these important measures of sexual partnerships can be captured by our model to some extent by choosing suitable radii distribution and number of time steps.

Relatively dense networks give better results for lifetime number of partners, and give a realistic degree distribution with a few vertices of very high degree. For example, Lognormal$(2.4, 256)$ (as shown in Figure 5.1) gave mean degree 8.1, compared with 16.4 in the Foxman data and 10.4 in the Natsal data. The highest number of partners in the same lognormal model was 543, compared with 1000 in both real-world data sets. (Note,

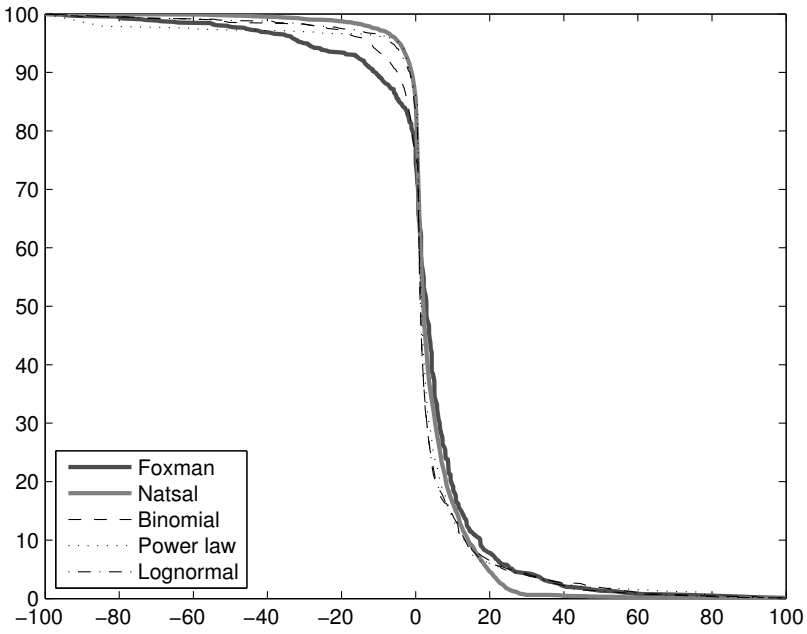


Figure 5.4: Cumulative distribution of gap/overlap lengths, as a percentage of T

however, that respondents tends to round off larger number of partners [35]).

Relatively sparse networks gave better results for lengths of partnerships and for gaps and overlaps. The best results for lengths of partnerships was for the network with mean degree 3.8 and highest degree 12, while for gaps and overlaps we obtained the best approximation from a network with mean degree 2.8 and highest degree 34. These numbers are not realistic.

The best fit for length of partnerships was lognormal with a long simulation ($T = 1024$ for second most recent partnership, $T = 512$ for most recent). For all other measures, the best fit was given by shorter simulations ($T = 256$).

Overall the closest results to the real-world data has been obtained for a contact network produced from the model with radius vector drawn from lognormal distribution. The most successful match (for all distributions) was to gaps and overlaps, with

a difference in area under the plot (compared to real-world data) of less than 1% when compared to the Foxman data, and around 1% when compared to the Natsal data, for all simulations shown in Figure 5.4.

No set of parameters captures all measures simultaneously but we are able to tune the input parameters to obtain needed output for one chosen measure.

5.3 Some results for the static binomial case

We now perform some further simulations for the case of binomially-distributed radii. We consider the initial graph G_0 only.

We used $M = 2560$ buckets. Starting with $N = 256$, we chose a position vector $(x_1, \ldots, x_{256})$ uniformly at random from $\{1, \ldots, M\}^{256}$. Then we pruned this vector, deleting points independently uniformly at random to get vectors of 128, 64 and 32 points to see how the changes in density of produced network will affect output. For each N we performed a hundred simulations: one run for each position vector drawn from $Bin(\frac{M}{2}, p)$. Here p takes various values in the range

$$p = \left[\frac{1}{8192}, \frac{1}{2}\right].$$

We calculated and ploted values for the output parameters averaged over a hundred runs. We also calculated a line/curve of best fit using least squares, which is shown as a dotted line on each plot, for comparison.

5.3.1 Vertex degrees

Mean degree as a function of p (line of best fit): $f(p) \sim (2.02p + 0.003)(N - 1)$. This

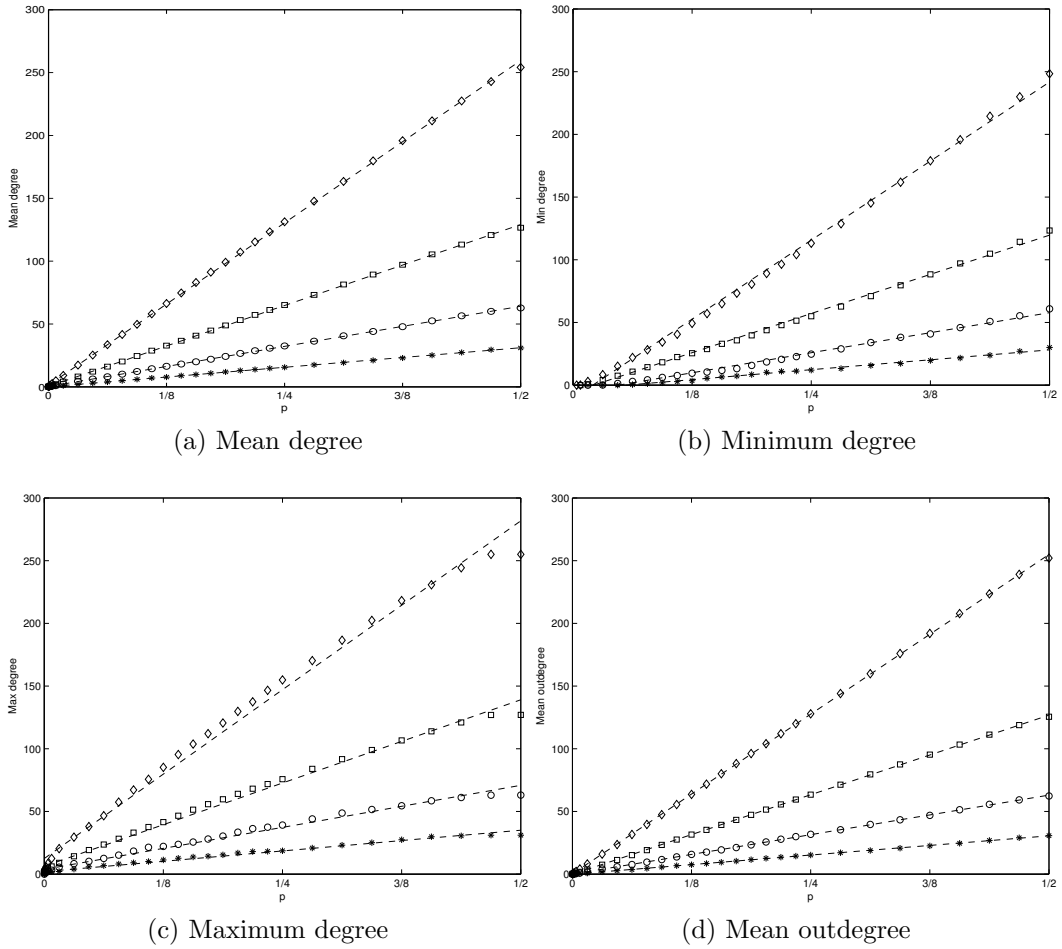
result is similar to the expected degree of a vertex in a random geometric graph with radius r, which equals $2rN$.

Minimum degree as a function of p (line of best fit): $f(p) \sim (2.01p - 0.07)(N-1)$.

Maximum degree as a function of p (line of best fit): $f(p) \sim (2.12p + 0.06)(N-1)$.

Mean outdegree as a function of p (line of best fit): $f(p) \sim (1.99p - 0.001)(N-1)$.

The data points and the lines of best fit are shown in Figure 5.5.



(a) Mean degree (b) Minimum degree

(c) Maximum degree (d) Mean outdegree

Figure 5.5: Vertex degrees for $N = 32$ ($*$), $N = 64$ ($\circ$), $N = 128$ ($\square$), $N = 256$ ($\diamond$)

5.3.2 The number of connected components

We found a curve of best fit of the form: $f(p) \sim c_1 N \exp(-c_2 4N^2 p)$. The best fitting curve of this form was $f(p) \sim 0.98N \exp(-1.144N^2 p)$.

The data points and the lines of best fit are shown in Figure 5.6.

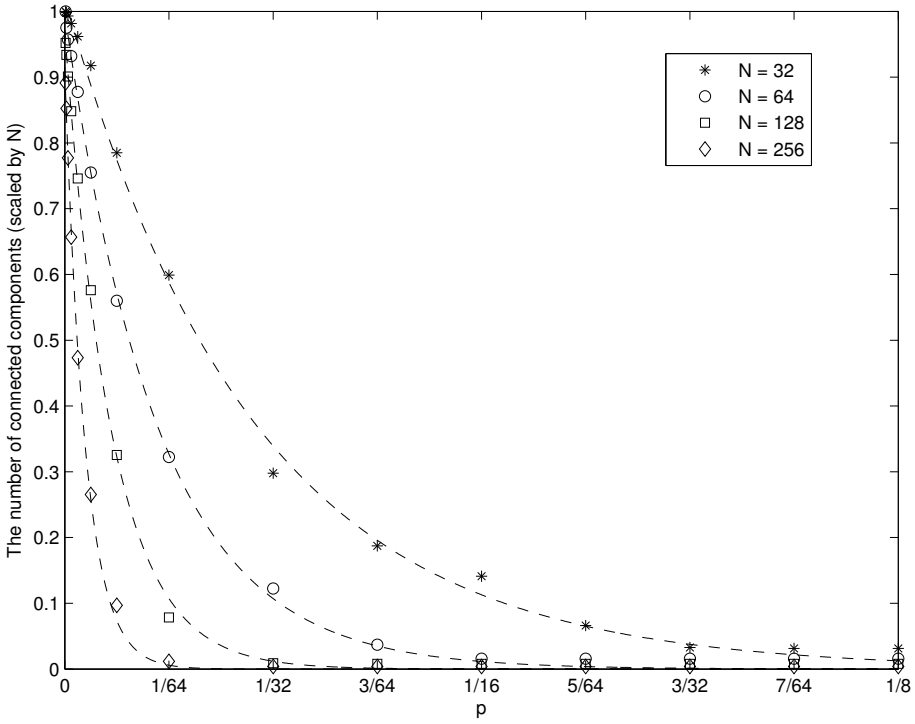


Figure 5.6: The number of connected components

5.3.3 Clustering coefficient

We found a curve of best fit of the form: $f(p) \sim c_1 - \exp(-c_2 p)$. The best fitting curve of this form was $f(p) \sim \frac{3}{4} - \exp(-1.56p)$.

The data points and the lines of best fit are shown in Figure 5.5.

We now give a heuristic argument to explain the limiting value $\frac{3}{4}$.

Let u, v and w be vertices.

If p is small then standard deviation of radii is very small. Therefore, assume that the radii of u, v and w are equal to $r \in \mathbb{N}$.

Let $B_r(v)$ be a ball with a centre in v and a radius equal to r.

Let i be a possible position (a bucket) for w and j be a possible position for u such as $i, j \in B_r(v)$.

Then the probability that $u \in B_r(w)$ can be calculated as follows:

$$\begin{aligned}
\text{Prob}(u \in B_r(w)) &= \sum_{i=-r}^{r} \sum_{j=-r}^{r} \text{Prob}(w \text{ is at position } i \,|\, w \in B_r(v)) \\
&\qquad \times \text{Prob}(u \text{ is at position } j \,|\, u \in B_r(v)) \\
&\qquad \times \mathbb{1}(u \in B_r(w)) \\
&= \sum_{i=-r}^{r} \sum_{j=-r}^{r} \frac{1}{(2r+1)^2} \mathbb{1}(u \in B_r(w)) \\
&= \frac{1}{(2r+1)^2} \sum_{i=-r}^{r} \{\text{number of values for } j \in \{-r, .., 0, .., r\} \text{ which lie in } B_r(w)\} \\
&= \frac{1}{(2r+1)^2} \sum_{i=-r}^{r} (2r + 1 - |i|) \\
&= \frac{1}{(2r+1)^2} \left(2r + 1 + 2 \sum_{i=1}^{r} (2r + 1 - i) \right) \\
&= \frac{1}{(2r+1)^2} \left(2r + 1 + 2r(2r+1) - 2\frac{r(r+1)}{2} \right) \\
&= \frac{3r^2 + 3r + 1}{4r^2 + 4r + 1} \\
&\approx \frac{3}{4}, \qquad \text{as } r \text{ is an integer.}
\end{aligned}$$

Therefore, the probability that two neighbours of the vertex v are neighbours themselves is approximately equal to $\frac{3}{4}$.

Dall and Christensen [27] derived a formula for the clustering coefficient for different

dimensions of the random geometric graph model. They obtained $\frac{3}{4}$ for the 1-dimensional case.

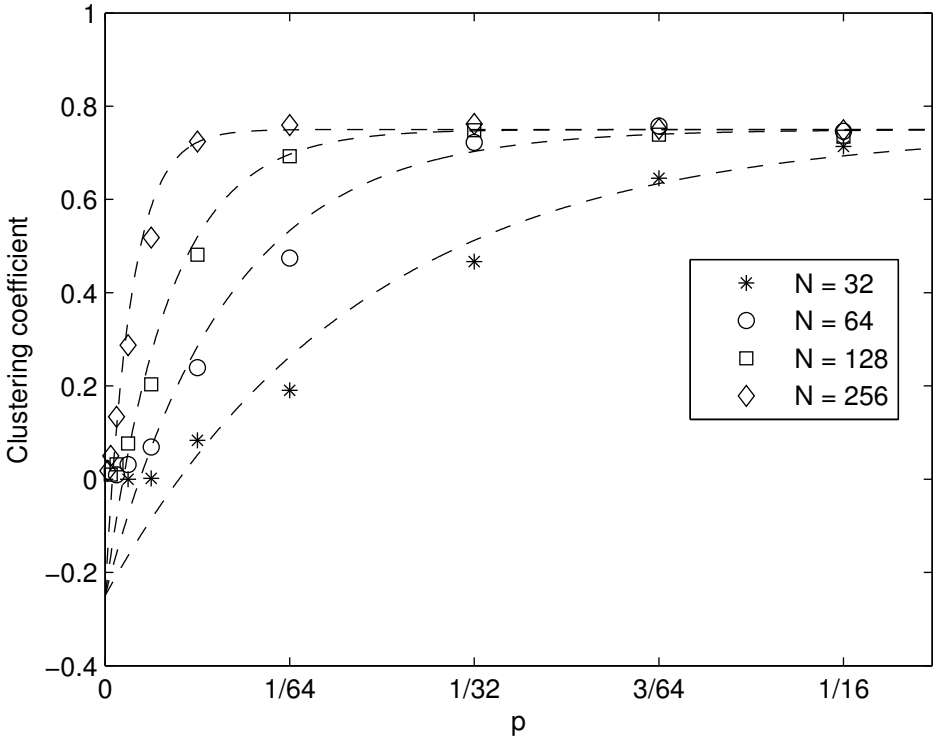


Figure 5.7: Clustering coefficient

5.4 Summary

In this chapter we described the simulations and discussed the results. We calculated cumulative degree distribution, length of partnerships distribution and gap/overlap distribution for different values of input parameters and different distributions of vertex radii. Then we compared the simulated data with real-world data from two sources. We also described some results on the static binomial case of the SEEN model.

Chapter 6

Criminal networks

The content of this chapter is a condensed and edited version of the book chapter [51] by Bright, Greenhill and myself. The chapter was originally written for a social sciences and criminology audience. Here we changed to graph-theoretic terminology rather than SNA (social network analysis) terminology.

6.1 Introduction

In this section, we examine a case study of a criminal network by describing the global structure of the network. The main aim of our research is to investigate different strategies for dismantling the networks using computer simulation.

For the current study, we used an existing data set described in a previous paper [23]. The method used to extract and analyse the data was as follows. A search was conducted on the NSW Lawlink website for criminal cases between January 1999 and May 1999 using two search terms: "methamphetamine" and "methylamphetamine". The aim was to find

a criminal network involved in the manufacture and trafficking of methamphetamine for a related project. Cases were included if they involved the manufacture and distribution (including importation) of the drug.

Previous research using the same data set identified seven roles for 35 of the 36 network members. There was insufficient information on one individual, so this vertex was not given a role (see [23]). Roles were determined based on the main set of activities or responsibilities for the different network participants, as described in judges' sentencing comments. Table 6.1 shows the seven roles and a description for each.

A network map is displayed in Figure 6.1 with roles indicated by shape and shading.

We use simulations to investigate different strategies for dismantling the networks, and apply two "measures of disruption" to evaluate the effectiveness of these strategies. Specifically, we compare law enforcement strategies which focus on degrees with those which focus on attributes of individuals in the network. The motivation for this work is to explore which factor or combination of factors law enforcement should focus on when targeting vertices with the aim of dismantling the network.

Our first aim was to examine whether the structure of the network could be classified as scale-free or exponential. We applied a goodness-of-fit test to check whether the degree distribution of the network is consistent with a power-law distribution (with cut-off).

Next, to examine the impact of law enforcement interventions, we conducted four sets of simulations. A computer simulation needs numerical data in order to perform calculations. Therefore, in order to take the role of each vertex into account, the roles must somehow be quantified. Our approach to this was to assign a weight to each vertex, where the weight is inversely proportional to the number of individuals in the syndicate with the same role. Hence the weight is a proxy for how difficult it might be to replace that individual, were they to be removed from the network (i.e., arrested). For example,

Role	Descriptor
Managers	Designated tasks to others, provided the funds for parts of the drug trafficking operation, or to whom other individuals reported.
Clan lab managers	Managed the operation of clandestine laboratory sites.
Wholesale dealers	Responsible for selling methamphetamine in single to multiple kilogram lots.
Resource providers	Sourced chemicals and equipment required for the manufacture of the drug.
Specialists	Possessed specialist knowledge and skill in the manufacture of methamphetamine.
Workers/labourers	Paid a wage to complete tasks or follow orders.
Corrupt officials	Occupied government positions and received bribes to behave in corrupt ways.

Table 6.1: Descriptions of the roles played by individuals in the network

there are two managers in the network, so they are both assigned a weight of 0.5. There are 10 workers in the network, so they all receive a weight of 0.1. While we acknowledge that this is simplistic, and that there are many other ways to quantify how important vertices are in the network, we believe that these weights provide an educated guess at the importance of each vertex, in the absence of further information. The vertex with unknown role was assigned a weight of 0. The weights of each vertex are shown in Table 6.2.

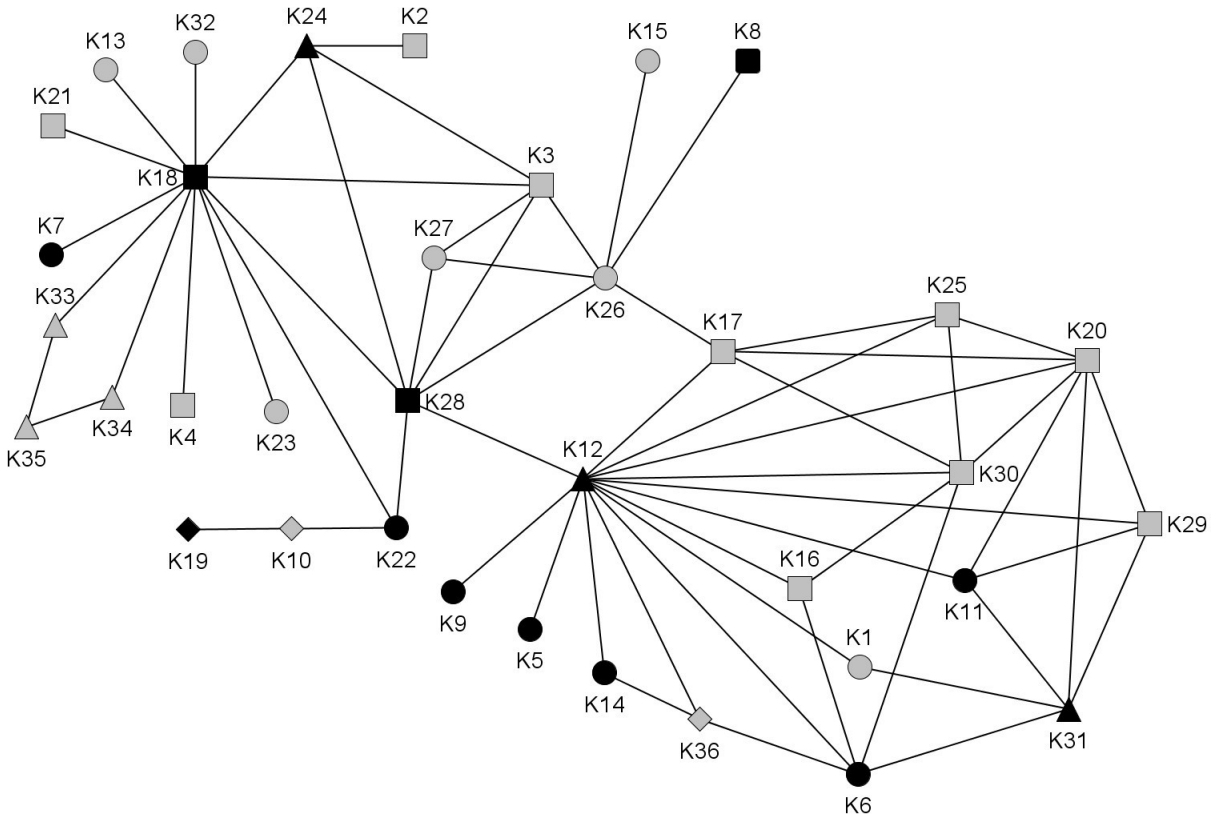


Figure 6.1: Network map of the methamphetamine trafficking network showing roles.
Roles:
Manager/Assistant manager ■
Clan lab branch manager ▲
Resource provider ●
Unknown ◆
Possession of specialist skills ◆
Worker "labourer" ■
Corrupt officials ▲
Wholesale dealer ●

We used four different law enforcement simulations to examine the differential impact of targeting criminal networks based on degrees, roles, a mix of degree/role information, and a random strategy. The four different rules we used in the simulations were as follows:

(1) Random attack. Vertices are targeted in a random order. Although it could

Role	Vertices with that role	Weight of each vertex
Managers/Assistant managers	$K18, K28$	$\frac{1}{2}$
Possession of specialist skills	$K10, K36$	$\frac{1}{2}$
Clan lab "branch manager"	$K12, K24, K31$	$\frac{1}{3}$
Corrupt official	$K33, K34, K35$	$\frac{1}{3}$
Wholesale dealer	$K1, K13, K15, K23,$ $K26, K27, K32$	$\frac{1}{7}$
Resource provider	$K5, K6, K7, K8,$ $K9, K11, K14, K22$	$\frac{1}{8}$
Worker/"labourer"	$K2, K3, K4, K16, K17,$ $K20, K21, K25, K29, K30$	$\frac{1}{10}$
Unknown role	$K19$	0

Table 6.2: Roles and associated weights assigned to vertices in the network

be argued that random removal of vertices might simulate "random" law enforcement interventions (e.g., stop and search; border detection), it is better conceived as a baseline comparison for targeted intervention. Random removal is relatively easy as no knowledge of the network structure is required. If some individuals are hard to locate, with a random strategy, any person will suffice as a target [25].

(2) Degree attack. The vertex of highest remaining degree was selected for removal;

(3) Weight attack. The vertex with the highest remaining weight was removed from the network; and

(4) Mixed strategy. The degree attack and weight attack can be combined to produce

a family of mixed strategies. For a given constant c between 0 and 1, we define the score of a vertex v in the current network to be

$$S(v) = (1 - c)\, d(v) + c\, Bw(v),$$

where $d(v)$ denotes the degree of vertex v in the current network, $w(v)$ denotes the weight of vertex v, and B is a constant chosen so that, over the initial network, the average contribution of two terms is equal. (For the methamphetamine network we used $B = 124/7$). When $c = 0$ we obtain degree attack, and when $c = 1$ we obtain weight attack. For intermediate values of c we have a combination of the two. For example, when $c = 0.5$, role and degree contribute equally to the selection of vertices to be removed. We investigated many possible combinations of weight and degree attack (c) to find the most effective combination. For the given network, setting $c = 0.1$ gave the best results. (This is explained and quantified in the next section.) So we only report on the mixed strategy with $c = 0.1$. For this value of c, ten percent of vertex selection is based on role, and ninety percent is based on degree.

In each simulation, at each time step a vertex of the current network is chosen according to some rule and deleted from the network. In the case of a tie (e.g. for maximum degree, or for maximum weight) a vertex with the maximal value was chosen randomly. We performed 100 runs of each simulation, always starting the methamphetamine network as the initial network.

The vulnerability of dark networks to being dismantled by law enforcement can be measured in various ways. In earlier work [23] we investigated the following measures of fragmentation for the degree targeting and random targeting simulations: the number of vertices in the largest connected component, the number of isolated vertices, maximum

degree, and the number of connected components. For each of these measures, degree targeting significantly outperformed random targeting.

In the current study we use two outcome measures for the connectivity of the network: one which ignores role information, and one which tries to take role information into account. A connected component in a network is a maximal set of vertices such that all pairs of vertices in the set are joined by a path in the network. Let $n(G)$ denote the number of vertices in the largest connected component of the current network G. This measure (and the other fragmentation measures mentioned above) only considers the topological structure of the network, and ignores individual attributes of the vertices, such as role. But we wish to investigate how individual attributes, such as roles, can affect the choice and performance of intervention strategies. Therefore it is desirable, and arguably appropriate, to include role information in our method for measuring the success of the intervention strategies. This led us to investigate another measure, which we call the disruption function. For the current network G the disruption function is given by

$$n(G) + Kw(G),$$

where $w(G)$ is the maximum, over all connected components, of the sum of the weights of the vertices in that component. The disruption function is given by the number of vertices in the largest connected component, plus a constant multiplied by the largest total weight among all connected components. Note that the heaviest component need not be the largest connected component. In this way, the disruption function is a composite measure of fragmentation and the "irreplaceability" of the remaining connected vertices, and hence takes both topological and individual-level data into account. The constant K is chosen so that when G is the initial network, the contribution from both terms is

equal. For the methamphetamine network, this is achieved by setting $K = 36/7$.

For each of our four simulations, both of these measures were calculated after every vertex deletion in each run, and then averaged over the 100 runs. The values for the four simulations were then plotted together in a graph, for each of the two measures.

6.2 Results

6.2.1 Network structure

One way to examine the mathematical properties of real-world networks is to compare them with their random graph equivalents. We use the Erdős-Rényi model $\mathcal{G}(n, p)$ discussed in Section 3.1.1. Here, random graphs are constructed by making connections between vertices using a pre-set probability p. For example, if $p = 0.5$ then, for each pair of vertices in the network, there is a 50% probability that they will be connected. To produce equivalent random graphs against which to compare our real world networks, the probability is calibrated to give the same average degree in the random graph as for the real world network. In the methamphetamine trafficking network, there were 36 vertices, 62 edges, and an average degree of 3.444. We used probability $p = 0.09841$ to give the same average degree in the corresponding random graph. The resulting graph is also called an "exponential network" in [3]. In this random graph, the probability that there is a vertex with degree of 14 is less than 7.6×10^{-5} (less than an 8 in 100 thousand chance), and the probability that there exists a vertex with degree 14 and a vertex with degree 12 (as in the methamphetamine network) is less than 3.09×10^{-7} (an approximately 3 in 10 million chance). So the methamphetamine trafficking network is very far from being like an Erdős-Rényi random graph.

To further test whether the network could be said to be scale-free, a goodness-of-fit test was applied to the data, following the method described in [26]. The outcome of this test suggests that the degree distribution of the methamphetamine network (for degrees 2 and above) is consistent with a power-law distribution with cut-off. Confidence in this conclusion is measured by the p-value, which for our network was around 0.2. (The hypothesis of a power-law distribution should be rejected if the p-value is less than 0.1)

6.2.2 Law enforcement simulations

First we investigated the dependence of the disruption function on the parameter c, in order to find a near-optimal value of c (at least for the given network). Note that the disruption function takes high values when the network is still well-connected and contains hard-to-replace vertices. On the other hand, a low value of the disruption function indicates that the network does not contain large connected components or smaller components containing high-weight vertices. Therefore, a low area under the disruption curve corresponds to efficient dismantling of the network, as measured by the disruption function. The area under the curve of the disruption function was calculated for many different values of c and averaged over 100 runs for each value of c (see Figure 6.2). We see that the area under the disruption function was lowest for values of c around 0.1. Therefore we chose to work with $c = 0.1$ as a convenient value.

Next we performed 100 runs of each of our simulations, and measured fragmentation of the network at each step by calculating the size of the largest connected component (see Figure 6.3). The X axis shows the number of time steps which have occurred, which equals the number of vertices which have been deleted from the network. The Y axis shows the average size of the largest connected component after each deletion, averaged

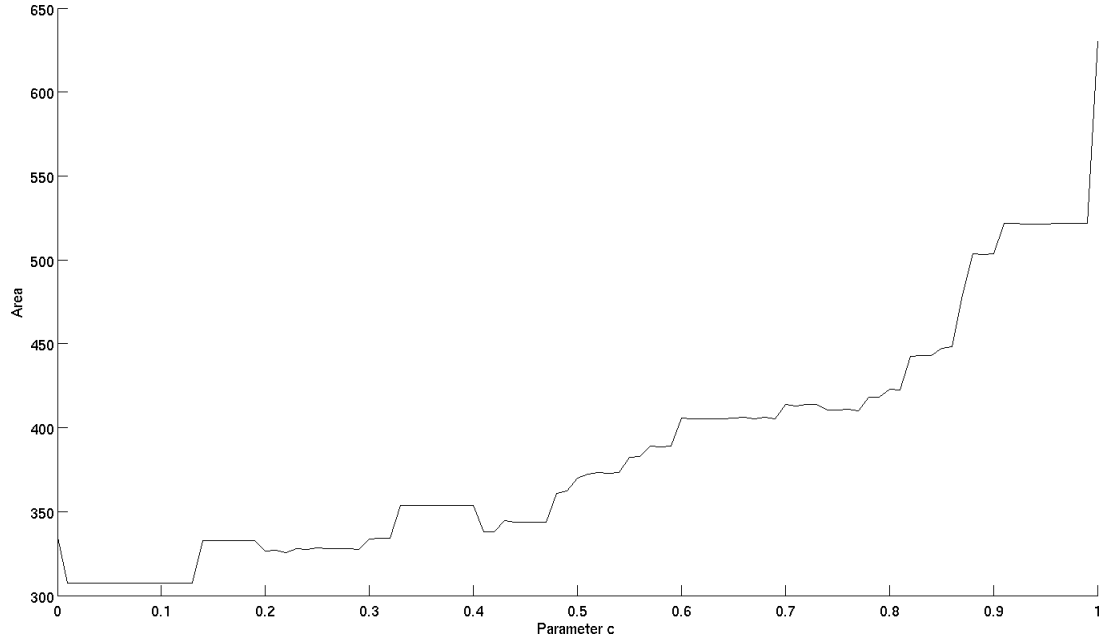


Figure 6.2: Average area under the disruption function, for various values of c

over 100 runs of the simulation. There are four plots shown, corresponding to the four strategies.

Just by viewing this plot, we see that random targeting is ineffective compared with the three other strategies. The best strategies (as measured by the size of the largest connected component) appear to be degree targeting and the mixed strategy. This can be quantified by comparing the area under each curve. These areas are 470.31 for random targeting, 373.43 for weight targeting, 160.60 for degree targeting and 158.99 for the mixed strategy. This confirms the visual impression that the degree targeting and mixed strategy have a similar performance, and that both outperform the other two strategies (targeting based on role only, and random targeting).

Finally we performed 100 runs of each of our simulations, calculating the disruption function of the network at each step. Figure 6.4 shows a plot of the average value of

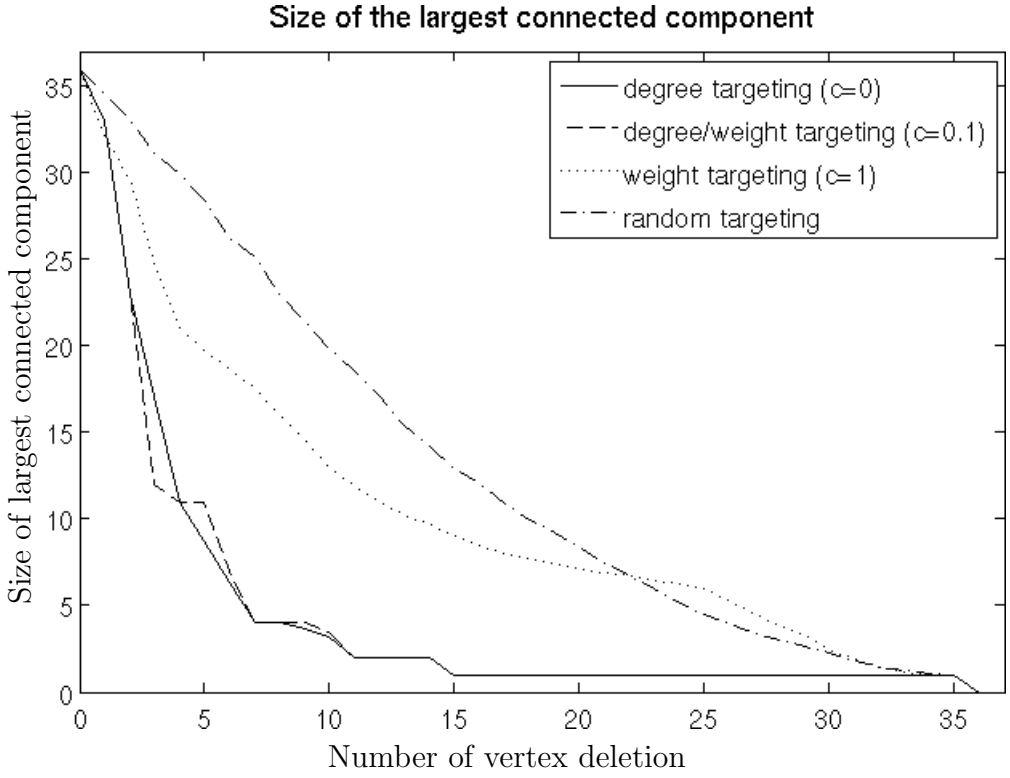


Figure 6.3: The four simulations, measured by the size of the largest connected component

this function over the four simulations. Again, the X axis shows the number of time steps which have passed, which equals the number of vertices which have been deleted. The Y axis shows the average value of the disruption function after each deletion, where the average is taken over 100 runs in each simulation. Four plots are again shown, corresponding to our four strategies.

Viewing the plot, we see that the best performance (as measured by the disruption function) is obtained using the degree targeting and mixed strategies, with these being very similar. In fact, the curve for the mixed strategy lies below the curve for the degree strategy most of the time, which suggests that the mixed strategy performed the best overall. The random strategy is far worse than all the others. The effectiveness

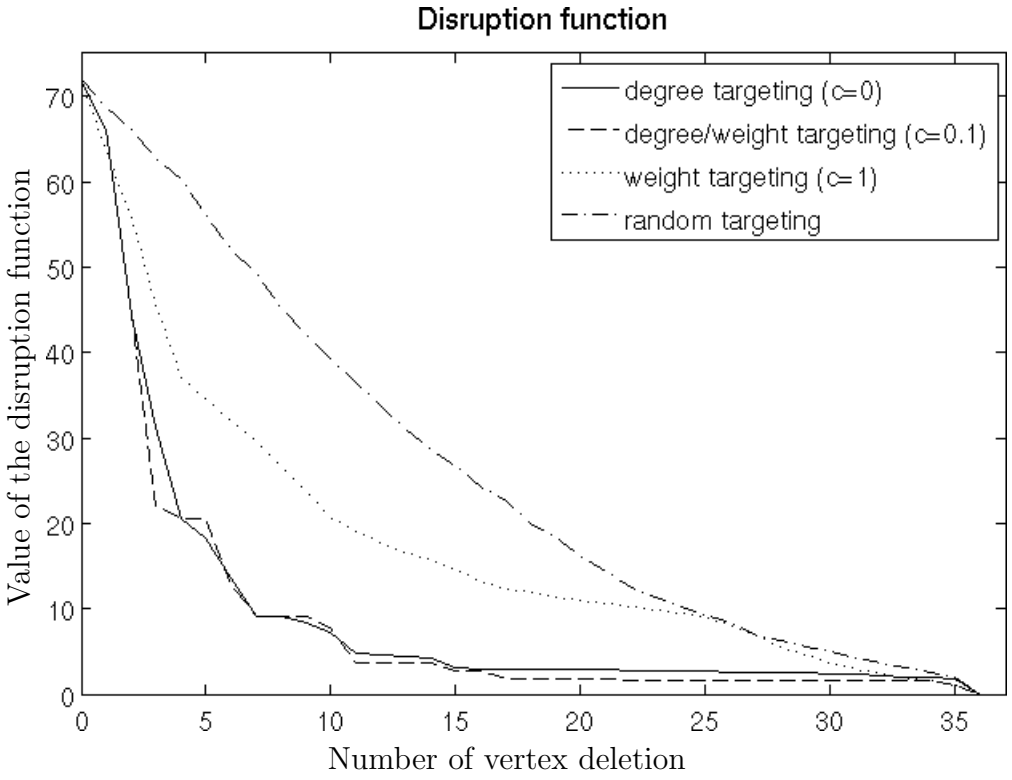


Figure 6.4: The four simulations, measured using the disruption function

of targeting based on role only falls between random targeting and the degree/mixed interventions. We quantify this by calculating the area under each curve, giving the following areas: 939.20 for random targeting, 635.19 for weight targeting, 335.25 for degree targeting and 307.58 for the mixed strategy. This confirms that the mixed strategy was the best in this case.

6.3 Discussion

The current study aims to examine whether criminal networks show evidence of being scale-free in structure, and to estimate the differential effectiveness of different law enforcement strategies aimed at dismantling criminal networks. We used a case study of

an Australian methamphetamine trafficking network.

We performed some calculations to compare the degree distribution of the methamphetamine network with the expected degree distribution of a random graph with the same density of links. Our calculations show that the methamphetamine network is not an "exponential network", as the vertices of high degree are extremely unlikely to exist in a random graph with the same average degree. Many real-world networks seem to be either exponential or scale-free [4]. By applying a goodness of fit test, we confirmed that a power-law distribution is a plausible hypothesis for the degree distribution of the metamphetamine network.

Scale-free networks are vulnerable to being dismantled by deleting vertices with many connections (i.e., hubs). Our simulations examine whether law enforcement should target hubs, or whether including information about vertex attributes is a more effective strategy. In our simulations, the degree targeting strategy proved very effective at fragmenting the network, with respect to both measures (size of largest component, and the disruption function). This is consistent with the finding that the network may be scale-free in structure and with previous research [40]. It is somewhat surprising that the mixed strategy achieved an improvement over degree targeting when measured using the maximum component size, since this is a purely topological measure. We believe that this improvement is an artefact of the properties of the methamphetamine network. Careful study of Figure 6.3 reveals that the curve for the mixed strategy dips below the curve for the degree targeting at the third step (after the third deletion) By the fourth deletion the two curves meet and thereafter the curve for degree targeting is always below or meeting the mixed strategy curve. So the improvement obtained by the mixed strategy is entirely due to the choices made in the third step of the simulation. We investigate this further below.

Under both strategies, the first two vertices which are deleted are always $K12$ and $K18$ (in that order). After these deletions, there are 22 vertices in the maximum connected component of the remaining network. Then the degree targeting strategy will randomly choose a vertex of highest remaining degree, namely $K20$ or $K28$ (both of degree 7). Each of these vertices is chosen with probability 0.5 and is deleted, leading to a network with maximum connected component of size 21 or 18. The average over 100 runs of the maximum component size at this step will be close to 19.5 (which equals the average of 21 and 18). However, at the third step the vertex with the highest score in the mixed strategy is $K26$. Hence this vertex will be deleted at the third step, leading to a network with a maximum connected component size of 11. This fairly dramatic improvement (reducing the size of the largest component from 22 to 11 in one step) is due to an important structural property of this particular criminal network. As can be seen from Figure 6.1, deleting the links from $K17$ to $K26$ and from $K12$ to $K28$ will disconnect the network into two pieces of roughly equal size. This can also be achieved by deleting one vertex from each of these two links, such as $K12$ and $K26$. In graph theory terminology, the set containing $K12$ and $K26$ is a cut set, meaning that deleting these vertices creates a disconnected graph. It is a very useful cut set for fragmenting the network because each remaining component is much smaller than the original network (around half the size). By chance, the mixed strategy manages to delete such a cut set in the first three steps, thereby producing a steep drop in the size of the largest connected component.

It is clear from Figures 6.3 and 6.4 that random targeting is much less effective than degree targeting, irrespective of whether effectiveness is measured using the size of the largest connected component or the disruption function. To quantify this, we calculated the ratios of the areas under the curve for random targeting versus degree targeting. This ratio equals 3.22 with respect to the maximum component size measure (Figure 6.3), and

equals 3.25 with respect to the disruption function (Figure 6.4). For this network, and using these two measures, degree targeting was three times more effective than random targeting.

As mentioned previously, the mixed strategy performed (slightly) better relative to the degree strategy. The ratio of the areas under the curve for the degree targeting versus mixed strategy was 1.006 with respect to the maximum component size measure (Figure 6.3) and was 1.092 with respect to the disruption function (Figure 6.4). It is not surprising that this small improvement is larger when measured by the disruption function, given that the disruption function takes role information into account when measuring fragmentation.

It should be noted that we used the value $c = 0.1$ since this gave optimal performance for the given network, as shown in Figure 6.2. It is noteworthy that this value of c is quite small, meaning that the mixed strategy only takes role information into account to a limited extent. This could be an artefact of the particular network studied.

Overall, strategies which targeted vertices based on degrees and on a combination of degrees and roles of individuals were most effective. Law enforcement interventions which use role information only and which do not consider degrees were relatively ineffective at dismantling the network. Interestingly, adding role information to degrees increased the effectiveness of law enforcement interventions in dismantling the network when the outcome measure incorporated the roles or ease with which individuals could be replaced (the disruption function).

The results of the study suggest that to effectively target and dismantle criminal networks, law enforcement should consider vertex level features (such as the roles played by individuals in the network) in addition to vertex topography features such as degrees. Indeed, the results suggest that only using role information to determine vertices to target

is less effective than using either degrees or degrees in concert with individual attributes (such as roles). The results underscore the utility of measures of degree in choosing individuals to target when the aim is to dismantle criminal networks. Scale-free network structure suggests that targeting informed by degrees is important. Law enforcement targeting which does not use degrees to guide arrests is likely to be ineffective (when the aim is to dismantle criminal networks). However, the study provides some early evidence that including information on vertex attributes may enhance law enforcement effectiveness.

There are a number of important limitations of this study, and for the generalisability of the results to real-world criminal networks:

(1) Networks are multimodal (include people, events, locations, resources). However, for the current study we used only static connections between people, and did not collect information on the type or strength of links;

(2) Criminal justice/law enforcement data can include intentional misinformation (e.g., aliases) and inaccuracies (e.g., typographic errors);

(3) Law enforcement and criminal justice data, such as that used in this study, are often incomplete. The network used in this study may be only a part of a larger network which remains hidden;

(4) The degrees of a particular vertices may be artificially inflated by the amount of information gathered on particular vertices during the investigation.

(5) The network that we studied is very small, with only 36 vertices. The mathematical techniques used in our analyses are limited by the small size of the network;

(6) There are many ways to quantify role information and to measure fragmentation of a network: we feel that the choices that we made were logical, however other choices may also be valid and may lead to different results;

(7) The simulations assumed that no new connections are made in response to vertex removal;

(8) The conclusions are limited to the network we examined: the methamphetamine trafficking network. Further research is required to determine the generalisability of our results;

(9) The aims of law enforcement are diverse. In this project we evaluated the extent to which law enforcement interventions (arrests) can dismantle a criminal network. However, law enforcement may seek to accomplish goals other than dismantling a network. For example, the aim may be to incapacitate the network so that the groups can no longer act illicitly, or to breach trust within the network such that the network disintegrates via internal distrust and conflict.

Despite the limitations, the simulation methodology we employed here can provide insights into the structural and functional damage that can be done to dark networks by targeted removal of vertices. It has the potential to demonstrate the utility of targeted vertex removal as a law enforcement intervention, especially compared with the removal of vertices such as drug couriers, wholesale dealers or other individuals who are easily replaceable and are not well connected with other vertices in the network, but may simply be more "visible". In addition, it provides a potential method for measuring the effectiveness of law enforcement interventions which aim to dismantle criminal networks.

Chapter 7

Conclusion

In this thesis we introduced two applications of graph theory which arose in two different areas.

The first application, described in Chapters 2–5, is at intersection of epidemiology and graph theory. The main aim of our research was filling a gap in modelling of contact networks which can be used for investigation of epidemics propagation. We found that graph theory is a good tool for network modelling and investigating properties of obtained simulated networks.

In the second project we worked with an existing real-world network describing a methamphetamine network. We used computer simulation to investigate strategies for dismantling the network.

In spite of the distinctions both applications have features in common. Both explored whether the simulated or real-world networks possess scale-free behaviour. Also we investigated whether it is possible to create a network similar to real networks and to explore properties of existing networks using mathematical and programming tools.

Future work may include simulation of disease propagation on the SEEN model and rigorous analysis of some aspects of the model.

Bibliography

[1] W. Aiello, F. Chung, L. Lu. A random graph model for power law graphs, *Experimental Mathematics* **10(1)** (2001), 53–66.

[2] R. Albert, A.-L. Barabási. Statistical mechanics of complex networks, *Reviews of Modern Physics* **74** (2002), 47–97.

[3] R. Albert, H. Jeong, A.L. Barabási. Diameter of the World-Wide Web. *Nature* **401** (1999), 130–131.

[4] R. Albert, H. Jeong, A.L. Barabási. Letters to Nature: Error and attack tolerance of complex networks. *Nature* **406** (2000) 378–382.

[5] N. Alon, J.H. Spencer. *The Probabilistic Method (2nd edn.)*, Wiley, New York, 2001.

[6] R.M. Anderson. Mathematical and statistical studies of the epidemiology of HIV, *AIDS* **3(6)** (1989), 333–346.

[7] R.M. Anderson, S. Gupta, W. Ng. The significance of sexual partner contact network for the transmission dynamics of HIV, *Journal of Acquired Immune Deficiency Syndromes* **3** (1990), 417–429.

[8] R.M. Anderson, R.M. May. *Infectious Diseases of Humans*, Oxford University Press, Oxford New York Tokyo, 1991.

[9] D. ben-Avraham, A.F. Rosenfeld, R. Cohen, S. Havlin. Geographical embedding of scale-free networks, *Physica A* **330** (2003), 107–116.

[10] J. Bang-Jensen, G. Gutin. *Digraphs Theory: Algorithms and Applications*, Springer-Verlag, London, 2007.

[11] S. Bansal, B.T. Grenfell, L.A. Meyers. When individual behaviour matters: homogeneous and network models in epidemiology, *Journal of the Royal Society Interface* **4(16)** (2007), 879–891.

[12] A.-L. Barabási, R. Albert. Emergence of scaling in random networks, *Science* **286** (1999), 509–512.

[13] A.-L. Barabási. Scale-free networks: a decade and beyond, *Science* **325** (2009), 412–413.

[14] L. Barnett, E.A. Di Paolo, S. Bullock. Spatially embedded random networks. *Physical Review E* **76(5)** (2007), 0586115

[15] J. van der Berg, R. Meester, D.G. White. Dynamic boolean model, *Stochastis Processes and their Applications* **69** (1997), 247–257.

[16] B. F. de Blasio, Å. Svensson, F. Liljeros. Preferential attachment in sexual networks, *Proceedings of the National Academy of Sciences* **104(26)** (2007), 10762–10767.

[17] B. Bollobás. *Random Graphs (2nd edn.)*, Cambridge University Press, Cambridge, 2001.

[18] B. Bollobás, O. Riordan. Mathematical results on scale-free random graphs in *Handbook of Graphs and Networks* Wiley-VCH, Berlin, 2003.

[19] B. Bollobás, O. Riordan. The diameter of a scale-free random graph, *Combinatorica* **24(1)** (2004), 5–34.

[20] B. Bollobás, C. Borgs, T. Chayes, O. Riordan. Directed scale-free graphs, *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '03)* (2003), 132–139.

[21] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems, *Proceedings of the National Academy of Sciences* **99(3)** (2002), 7280–7287.

[22] A. Bonato. *A Course on the Web Graph*, American Mathematical Society, Providence, Rhode Island, USA, 2008

[23] D.A. Bright, C. Greenhill, N. Levenkova. Attack of the nodes: scale-free criminal networks and vulnerability to targeted law enforcement interventions. Paper presented at the 2nd Illicit Networks Workshop (2010), Wollongong, Australia.

[24] S. Bullock, L. Barnett, E.A. Di Paolo. Spatial embedding and the structure of complex networks. *Complexity* **16(2)** (2010), 20–28.

[25] K.M. Carley. Destabalization of covert networks. *Computational and Mathematical Organization Theory* **12** (2006), 51–66.

[26] A. Clauset, C.R. Shalizi, M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review* **51(4)** (2009), 661–703.

[27] J. Dall, M. Christensen. Random geometric graphs. *Physical Review E* **66** (2002), 1-16.

[28] R. Diestel. *Graph Theory (3rd edn.)*, Springer-Verlag, Heidelberg, New York, 2005.

[29] K. Dietz, K.P. Hadeler. Epidemiological models for sexually transmitted diseases, *Journal of Mathematical Biology* **26** (1988), 1–25.

[30] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin. Structure of growing networks with preferential linking, *Physical Review Letters* **85(21)** (2000), 4633–4636.

[31] P. Erdös, A. Renýi. On random graphs I, *Publicationes Mathematicae (Debrecen)* **6** (1959), 290–297.

[32] P. Erdös, A. Renýi. On the evolution of random graphs, *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5** (1960), 17–61.

[33] B. Erens, S. McManus, J. Field, C. Korovessis, A. Johnson, K. Fenton, K. Wellings. National Survey of Sexual Attitudes and Lifestyles II: Technical Report, (2001) Available from http://www.natsal.ac.uk/media/821885/technical_report.pdf

[34] M. Faloutsos, P. Faloutsos, C. Faloutsos. On power-law relationships of the Internet topology, *Computer Communication Review* **29(4)** (1999), 251–262.

[35] B. Foxman, M. Newman, B. Percha, K.K. Holmes, S.O. Aral. Measures of sexual partnerships: Lengths, gaps, overlaps, and sexually transmitted infection, *Sexually Transmitted Diseases* **33(4)** (2006), 209–214.

[36] E.N. Gilbert. Random graphs, *The Annals of Mathematical Statistics* **30(4)** (1959), 1141–1144.

[37] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation, *Mathematical Biosciences* **63** (1983), 157–172.

[38] G. Grimmett, D. Stirzaker. *Probability and Random Processes*, Oxford University Press, New York, 2001.

[39] S. Janson, T. Łuczak, A. Rusiński. *Random Graphs*, Wiley, New York, 2000.

[40] B. Keegan, M.A. Ahmed, D. Williams, J. Srivastava, N. Contractor. Dark gold: Statistical properties of clandestine networks in massively multiplayer online games. Paper presented at the 2010 IEEE Second International Conference on Social Computing, IEEE Computer Society, Los Alamitos, CA, USA.

[41] M.J. Keeling, K.T.D. Eames. Networks and epidemic models, *Journal of the Royal Society Interface* **2(4)** (2005), 295–307.

[42] P.L. Krapivsky, S. Redner. Organization of growing random networks, *Physical Review E* **63** (2001), 066123.

[43] J.R. Kraut-Becher, S.O. Aral. Gap length: an important factor in sexually transmitted disease transmission, *Sexually Transmitted Diseases* **30(3)** (2003), 221–225.

[44] M. Kretzschmar. Sexual network structure and sexually transmitted disease prevention, *Sexually Transmitted Diseases* **27** (2000), 627–635.

[45] M. Kretzschmar, M. Morris. Measures of concurrency in networks and the spread of infectious disease, *Mathematical Biosciences* **133** (1996), 165–195.

[46] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, Y. Åberg. The web of human sexual contacts, *Nature* **411** (2001), 907–908.

[47] A.J. Lotka. The frequency distribution of scientific productivity, *Journal of the Washington Academy of Sciences* **16(12)** (1926), 317–324.

[48] C. McDiarmid, A. Steger, D.J.A. Welsh. Random planar graphs, *Journal of Combinatorial Theory, Series B* **93** (2005), 187–205.

[49] S. Milgram. The small-world problem, *Psychology Today* **1(1)** (1967), 61–67.

[50] M. Morris, M. Kretzschmar. Concurrent partnership and the spread of HIV, *AIDS* **11(5)** (1997), 641–648.

[51] C. Morselli. *Crime and Networks*, Taylor & Francis Ltd, New York and London, 2013.

[52] M.E.J. Newman. The spread of epidemic disease on networks, *Physical Review E* **66** (2002), 016128.

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

[54] M.E.J. Newman, J. Park Why social networks are different from other types of networks, *Physical Review E* **68** (2003), 036122.

[55] R. Pastor-Satorras, A. Vespignani. Epidemic dynamics and endemic states in complex networks, *Physical Review E* **63** (2001), 066117.

[56] M. Penrose. *Random Geometric Graphs*, Oxford University Press, Oxford, 2003.

[57] Y. Peres, A. Sinclair, P. Sousi, A. Stauffler. Mobile geometric Graphs: Detection, coverage and percolation, *Probability Theory and Related Fields* **156** (2013), 273–305.

[58] K.A. Powers, I.F. Hoffman, A.C. Ghani, M.C. Hosseinipour, C.D. Pilcher, M.A. Price, A.E. Pettifor, D.A. Chilongozi, F.E.A. Martinson, M.S. Cohen, W.C. Miller. Sexual partnership patterns in Malawi: Implications for HIV/STI transmission, *Sexually Transmitted Diseases* **38** (2011), 657–666.

[59] D.J. de S. Price. Networks of scientific papers, *Science* **149** (1965), 510–515.

[60] B. Roche, J.M. Drake, P. Rohani. An agent-based model to study the epidemiological and evolutionary dynamics of influenza viruses, *BMC Bioinformatics* **12** (2011), 87.

[61] A.F. Rosenfeld, R. Cohen, D. ben-Avraham, S. Havlin. Scale-free networks on lattices, *Physical Review Letters* **89(21)** (2002), 218701.

[62] A. Schneeberger, C.H. Mercer, S.A.J. Gregson, N.M. Ferguson, C.A. Nyamukapa, R.M. Anderson, A.M. Johnson, G.P. Garnett. Scale-free networks and sexually transmitted diseases: A description of observed patterns of sexual contacts in Britian and Zimbabwe, *Sexually Transmitted Diseases* **31** (2004), 380–387.

[63] K.J. Sharkey. Deterministic epidemiological models at the individual level, *Journal of Mathematical Biology* **57** (2008), 311–331.

[64] R. Solomonoff, A. Rapoport. Connectivity of random nets, *Bulletin of Mathematical Biophysics* **13** (1951), 107–117.

[65] E.M. Volz, J.C. Miller, A. Galvani, L.A. Meyers. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics, *PLoS Computational Biology* **7** (2011), e1002042.

[66] M. Walters. Random geometric graphs in *Surveys in Combinatorics 2011*, Cambridge University Press, **392** (2011), 365–401.

[67] C.P. Warren, L.M. Sander, I.M. Sokolov. Geography in a scale-free network model, *Physical Review E* **66** (2002), 056105.

[68] D.J. Watts, S.H. Strogatz. Collective dynamics of "small-world" networks, *Nature* **393** (1998), 440–442.

[69] N.C. Wormald. Models of random regular graphs, *Surveys in Combinatorics*, Cambridge University Press, **276** (1999), 239–298.

[70] R. Xulvi-Brunet, I.M. Sokolov. Evolving networks with disadvantaged long-range connections, *Physical Review E* **66** (2002), 026118.