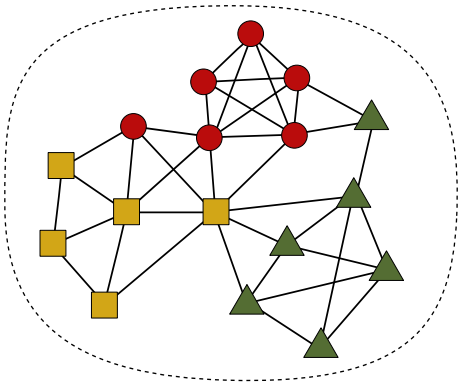
ADAM W. HACKETT

CASCADE DYNAMICS ON COMPLEX NETWORKS

CASCADE DYNAMICS ON COMPLEX NETWORKS

ADAM W. HACKETT



A thesis submitted for the degree of Doctor of Philosophy
in the

Department of Mathematics and Statistics
Faculty of Science and Engineering
University of Limerick

SUPERVISOR: Prof. James P. Gleeson
HEAD OF DEPARTMENT: Dr. Mark Burke

October 2011

To my parents.

The mind is not a vessel to be filled,
but a fire to be kindled.

— after Plutarch [78]

CASCADE DYNAMICS ON COMPLEX NETWORKS

ADAM W. HACKETT

ABSTRACT

The network topologies on which many natural and synthetic systems are built provide ideal settings for the emergence of complex phenomena. One well-studied manifestation of this, called a *cascade* or *avalanche*, is observed when interactions between the components of a system allow an initially localized effect to propagate globally. For example, the malfunction of technological systems like email networks or electrical power grids is often attributable to a cascade of failures triggered by some isolated event. Similarly, the transmission of infectious diseases and the adoption of innovations or cultural fads may induce cascades among people in society.

In recent years, it has been extensively demonstrated that the dynamics of cascades depends sensitively on the patterns of interaction laid out in the underlying network. One of the goals of *network theory* is to provide a solid theoretical basis for this dependence. In order to do this it is necessary, first, to construct network models that are both mathematically sound and capture the salient features of their real-world counterparts. So far, there has been limited success in this direction. The primary shortcoming of most existing network models in this regard is their lack of realistic structural motifs, in particular the absence of significant levels of *clustering*, which refers to the propensity of triples of connected vertices to form triangles, and is a prominent feature of networked systems across multiple settings.

In this thesis we investigate the interplay between network structure and cascade dynamics. Beginning with dynamics, we consider an analytically tractable technique to determine the expected cascade size in a broad range of dynamical models on locally tree-like networks of arbitrary degree distribution. We validate this approach by demonstrating its excellent agreement with the results of extensive numerical simulations, and closely examine its applicability to real socio-technological systems. Here we focus particularly on problems relating to social influence and opinion formation, and we develop a number of important modifications of the basic theory.

Following this, we turn our attention to the structural characterization of networks. We investigate the properties of a new generation of network models that incorporate clustering by embedding cliques of fully connected vertices within a locally tree-like topology, and that thus directly extend the classical *configuration model* construction. In one such model, devised by a member of our group, the sizes of these cliques may vary, allowing one to prescribe a clustering spectrum to match empirically measured values.

Finally, we significantly extend the theory of dynamics on tree-like networks to these new, more structurally realistic ones. From this we uncover answers to some important questions, which have earned considerable recent attention, concerning the effects of increased clustering on cascades.

PUBLICATIONS

Some of the ideas and results presented in this thesis have appeared previously in the following publications:

[76] Gleeson, J. P., Melnik, S., and Hackett, A. How clustering affects the bond percolation threshold in complex networks. *Phys. Rev. E* ***81***, 066114 (2010).

[84] Hackett, A., Gleeson, J. P., and Melnik, S. Site percolation in clustered random networks. *Int. J. Comp. Syst. Sci.* ***1***, 25-30 (2011).

[85] Hackett, A., Melnik, S., and Gleeson, J. P. Cascades on a class of clustered random networks. *Phys. Rev. E* ***83***, 056107 (2011).

[99] Melnik, S., Hackett, A., Porter, M. A., Mucha J. P., and Gleeson, J. P. The unreasonable effectiveness of tree-based theory for networks with clustering. *Phys. Rev. E* ***83***, 036112 (2011).

Chapter 5 contains work from

[83] Hackett, A., and Gleeson, J. P. Cascades on graphs with embedded cliques. In preparation.

The increment of meaning corresponds to the increased perception of the connections and continuities of the activities in which we are engaged.

— John Dewey [40]

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. James Gleeson for his patient support and guidance throughout the course of this project. His unyielding work ethic and enthusiasm for tackling important research challenges by forging new collaborations has been a great source of inspiration to me as an aspiring academic, and I am grateful to have conducted the work presented here as a member of his Stochastic Dynamics and Complex Systems (SDCS) research group. Special thanks also go to Dr. Sergey Melnik[1] for his helpful advice, comments, and suggestions during the writing of this thesis and previous publications.

I have benefited greatly from thought-provoking discussions with and encouragement from colleagues in the Department of Mathematics and Statistics at the University of Limerick, and it has been a pleasure to watch the university, and this department in particular, go from strength to strength during my time here. Thanks to the late Prof. Frank Hodnett, the founding head of the department, for teaching me calculus in my first year as an undergraduate. May he rest in peace.

Finally, it has been a privilege for me to contribute to the tireless efforts of all those associated with the Mathematics Applications Consortium for Science and Industry (MACSI) to reinforce collaborative ties between the Irish mathematics research community and industrial partners, and, more broadly, to foster a nationwide educational culture in which the application of mathematical methods of problem solving is appreciated as an imperative and intensely creative endeavour, long may it continue to do so.

This work was funded by Science Foundation Ireland under programmes 06/IN.1/I366 and MACSI 06/MI/005.

1 Member of SDCS.

CONTENTS

LIST OF FIGURES

CODE LISTINGS

ACRONYMS

CDF Cumulative Distribution Function

GCC Giant Connected Component

MACSI Mathematics Applications Consortium for Science and Industry

PAP Permanently Active Property

PMF Probability Mass Function

PRG Poisson Random Graph

RFIM Random Field Ising Model

SDCS Stochastic Dynamics and Complex Systems

SIR Susceptible-Infective-Recovered

SIS Susceptible-Infective-Susceptible

SFN Scale-Free Network

SWN Small-World Network

VoIP Voice over Internet Protocol

1 INTRODUCTION

1.1 CASCADES AND COMPLEXITY

We have all heard it said that we live in a connected age, an age in which our fate, for better or worse, is becoming ever more bound up in the contingencies of accumulated individual actions. Owing in part to the proliferation of interactive technologies such as Web-based social media much of our contemporary popular discourse is driven by the notion that each of us has an important part to play in shaping not just our immediate environments but the world at large; however, while few deny the moral imperative to confront the criticality of our times, a cynic may note the convenience of the dictum *we are all in this together* as means for the truly powerful and culpable to reapportion blame.

Wherever the truth may lie in respect to the extent to which each of us can effect change, the narrative of connectivity and accumulated action does have a genuine basis in reality. From mass political uprisings and global financial crises, to infectious disease epidemics and ecological catastrophes, the potentiality of localized phenomena to very quickly accrue global significance is perhaps greater today than ever before. We call events of this type *cascades* or *avalanches*. The task of providing a scientific explanation as to why they seem to occur so often nowadays is far from a trivial one; however, there is a broad appreciation that they are a symptom of the increasing *complexity* of our world.

What is complexity? It can be roughly defined as the degree of difficulty in predicting the global behaviours of a system, given that the properties of each of its constituent parts are known. Evidently, any system that scores highly in this regard may be termed complex. Notwithstanding its current vogue, when expressed in this simplified manner we see that it is not such a novel concept after all but rather one that has been with us from the earliest days of Western intellectual thought: recall Aristotle's famous line, "... the totality is not as it were a mere heap but the whole is something besides the parts" [79]. In fact, throughout history many of the most eminent thinkers have addressed various manifestations of complexity, though they lacked our modern vocabulary: Adam Smith's *invisible hand* [131]; and Charles Darwin's *natural selection* [36] are just two prominent examples.

If this is true, then what need of a so-called *new* science of complexity, of which so much is written? The conventional argument goes that ever

since its germination in the philosophy of Descartes and the mechanics of Galileo the modern scientific method has been engaged in a sort of brute reductionism that does well at explaining the workings of the fundamental cogs of the universe but does not tell us how these all fit together to form not a unified, but rather a diversified, evolving, and chaotic whole. This picture somewhat misses the point. Despite the inflated rhetoric, when confronted with the legacy of the past four hundred years of scientific enquiry no one is seriously prepared for a wholesale abandonment of the basic reductionist principles. Furthermore, it is hard to conceive of a framework that could stand up to the measure of this legacy that has not been built from the bottom up.

Nevertheless, there is an important sense in which the new paradigm is timely and correct. Though in reality we may have been studying complex systems all along, traditional scientific taxonomies have served to conceal this fact from view, like the proverbial *ghost in the machine* [126]. The (supposed) hierarchy of knowledge that leads from the ivory towers of mathematics and physics, through the teeming citadels of chemistry and biology, right down to the dank mire of psychology and sociology is no longer adequate. In order to describe our connected age we are compelled to knit these fields together in a variegated tapestry of axioms, observations and metaphors, and although the final picture that will emerge from this endeavour is not yet clear it is well understood that the study of *networks* will be a unifying theme.[1]

1.2 THE RISE OF NETWORK THEORY

There are many systems of interest to scientists that are composed of individually functioning parts connected together in networks. Examples range across multiple disciplines and scales, from the neural circuitry of nematode worms [148] to the hyperlinks and webpages of the World Wide Web [89]. Over the past decade or so, with the advent of cheap and powerful personal computers, the measurement and analysis of these networked systems has revealed that many of them share unifying structural traits. In turn, it has been shown that these traits can strongly determine the various complex phenomena that these systems exhibit, including cascades.

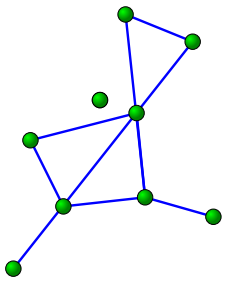
This synthesis of structure and process is the cornerstone of the burgeoning field of network theory. In its broadest terms this new field is an interdisciplinary framework for the study of complex systems, whose practitioners utilize a diverse array of tools and techniques from condensed

1 In the vanguard of this kind of thinking is the Santa Fe Institute in New Mexico, USA [http://santafe.edu/]. For a lively account of its founding vision see [66].

matter physics and discrete mathematics to behavioural psychology and beyond. In essence, however, it is a form of advanced applied graph theory, the historical development of which, as we will briefly sketch, can be traced through the augmentation of the basic graph theoretic modelling technique.

1.2.1 *The Origin of Graphs*

Recall, first, that a graph is simply a collection of points connected by lines. More formally, we call the points of a graph *vertices*, or *nodes*, and the lines *edges*. To model a complex system as a graph is to filter out the functional details of each of its components, and the idiosyncrasies of their interactions with each other, and to focus instead on the underlying structure (topology) as an inert mathematical construct. Although this technique is central also to network theory, the word network, in contrast, usually carries with it connotations of the context in which the overarching system exists, particularly when that system displays any sort of nonlinear dynamics. For example, when investigating the spread of infectious disease on a human sexual contact network it makes sense to consider the relevant sociological parameters as well as the abstract topology, and it is in such settings that the interdisciplinary aspect that distinguishes network theory comes to the fore.



A simple graph of nine vertices and ten edges.

It is vitally important, however, to not underestimate the value of the graphical abstraction. In many instances it provides the key to understanding the emergence of global system behaviours. In this regard we often find that ostensibly simple structural characteristics, such as the density of edges present and the way these edges are distributed between vertices, can play a non-trivial role. The first person to realize this was the progenitor of the theory of graphs, the Swiss mathematician, Leonhard Euler (1707–1783). His famous solution to the *Seven Bridges of Königsberg* problem, presented to the St. Petersburg Academy on August 26 1735, and later published in 1741 [60], is commonly cited as the earliest example of the use of this technique in a mathematical proof.

The city of Königsberg (now called Kaliningrad) was built on four land masses connected by seven bridges. The puzzle, which had allegedly stumped all of the townsfolk, asked for a walking route through the city that would cross each bridge exactly once.[2] When presented with the problem, Euler, in his genius, saw that all of the information necessary to solve it was contained within the abstract topology of the city's network of bridges. He, therefore, recast this network as a graph of four vertices

2 In graph theory such a route is now called an Eulerian *walk* or *path*. An alternative form of the problem asked for a route that would cross each bridge once, starting and finishing on the same land mass. This is an example of what is now called an Eulerian *circuit* or *tour*.

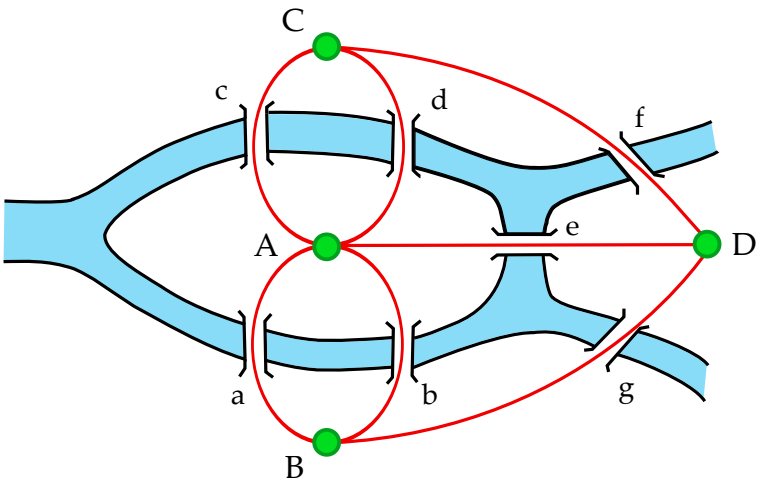


Figure 1.1: The river Pregel (blue) divides Königsberg into four land masses, labelled A to D. In Euler's time, the city was connected by a network of seven bridges, labelled a to g. Euler thought of these bridges as edges (red) connecting the vertices (green) of a graph.

connected by seven edges (see Fig. 1.1). This then enabled him to make the key observation that the walk being sought for would require a traveller to enter and leave each non-terminal vertex an equal number of times. In addition, if every bridge was to be traversed exactly once, it followed that the number of edges attached to each vertex, except possibly for the start and end vertices, must be even. Thus, a necessary condition for such a route to exist is that at most two vertices are attached to an odd number of edges. In Königsberg, however, all four land masses were touched by an odd number of bridges, rendering the walk impossible. It turns out that this condition is also sufficient; a result stated by Euler but not proven until much later, c. 1871, by Carl Hierholzer [88].

From this grounding the theory of graphs developed steadily over the next 220 years or so with mathematicians asking, and uncovering answers to, more and more intricate questions as their knowledge increased (see [14] and references therein). Throughout this time, however, graph theory became increasingly embedded in pure mathematics; its practical underpinnings falling victim to the scholastic disdain for applications that has so often characterized the latter. In fact, one might say that the appeal of the theory, as it progressed, was purely esoteric since only those mathematicians with a strong taste for rigour and abstraction tended to succeed in proving, or for that matter even deciphering, the outstanding conjectures of the day. The questions that tended to be asked about graphs typically concerned route finding; graph colouring; vertex covering; and graph enumeration [86, 117]. All of these problems were strictly deterministic; however, a watershed moment would occur around the middle

of 20th century when the prolific Hungarian mathematician Paul Erdös (1913–1996) would focus his attention on graph theory and, in a series of papers co-authored with Alfréd Rényi (1921–1970), develop the concept of a *random graph*.

1.2.2 *From Order to Randomness*

In the late 1950s interest in the statistical aspects of graphs reached critical mass; to the extent that a flood of publications all broadly related to this topic appeared in a very short period of time. In the space of only four years several authors, namely Gilbert [67], Ford and Uhlenbeck [63], Austin *et al.* [9] and Erdös and Rényi [56], all offered seminal contributions to what we now call the theory of random graphs.

Historically the consensus has been that each of these authors should be equally credited with the foundation of the theory; however, this view is overly conciliatory. The reality is that Erdös had considered the statistical properties of graphs at least as early as 1947 [54], when he first demonstrated the application of probabilistic methods to extremal problems [28]. Béla Bollobás, in his essential book [15], insists that the honour of having founded the theory should belong exclusively to Erdös and Rényi [56]. He argues that while each of the aforementioned authors helped shift the focus towards statistical considerations in general, it is only in [56] that we find the probabilistic treatment of graphs which is the true core of the theory as we know it today. According to Bollobás, "the other authors were all concerned with enumeration problems and their techniques were essentially deterministic" [15]. On the other hand, the approach of Erdös and Rényi "has only the slightest connection with enumeration" [15]. They were "not interested in exact formulae but rather in approximating a variety of exact values by appropriate probability distributions and using probabilistic ideas, whenever possible" [15].

In their probabilistic approach — introduced in [56] and extended further in [57, 58, 59] — Erdös and Rényi considered the properties of a *typical* graph in a probability space, or *ensemble*, $G_{n,M}$, which consists of all graphs with a given set of n labelled vertices and M edges. Each random graph realization is drawn from this ensemble with equal probability $1/\binom{N}{M}$, where $N = \binom{n}{2}$; and, if we allow n to vary we have ensembles of graphs corresponding to each natural number n.

By extending this latter idea to the limit $n \to \infty$ Erdös and Rényi found that many interesting graphical properties, such as the presence of subgraphs of particular sizes and configurations, are dependent on the density of edges between vertices. Specifically, if we choose a function

$M = M(n)$, and we are interested in the existence of some property Q as $n \to \infty$, then in many cases a critical point exists in the evolution of $M(n)$ at which a swift transition from Q being very unlikely to it being very likely occurs. Following the conventional nomenclature, if $Q \to 1$ as $n \to \infty$, we say that *almost every* graph has property Q, otherwise, almost every graph fails to have it.

Armed with this notion of probability spaces of graphs, succeeding theorists were able to provide answers to numerous questions that would have remained inaccessible to a strictly deterministic approach. Erdös and Rényi themselves extended the main results of [56] to prove other important theorems relating to the presence of *cycles* of connected vertices and connected subgraphs containing no short cycles called *trees* [55, 57, 59].

From the point of view of the emergence of modern network theory the introduction of probabilistic methods opened the doors to the mathematical treatment of problems of a more applied nature, concerning the complex networks found in the real world, many of which are extremely large and evolving. However, Erdös and Rényi were not interested in the practical application of such methods (they were the purest of mathematicians). Instead, that work was left to others in more grounded, though perhaps less rigorous, fields of study such as biology, sociology and even psychology. In many cases these other researchers touched upon ideas that today we recognize as fundamental network theoretic concepts; however, as discussed earlier, the traditional compartmentalization of scientific disciplines, has meant that the significance of these studies has not been fully appreciated (by mathematicians at least) until very recently.

1.2.3 *Networks in the Real World*

Perhaps the most innovative research in this first, and for many years overlooked, wave of network oriented investigations was conducted by the trailblazing Russian-born mathematical biologist Anatol Rapoport (1911-2007). In a work predating the publications of Erdös and Rényi, Rapoport, with his collaborator Ray Solomonoff (1926-2009), [133] considered a type of bond percolation process in which edges (or *axons* in their neurological vernacular) are added at random between pairs of vertices in an infinitely large set ($n \to \infty$). They were particularly interested in the size and frequency of connected components in the resulting random graphs (or *nets*) as the average number of edges attached to each vertex, the mean degree, z, is increased, and predicted that as the value of z increases beyond $z = 1$, a single *giant connected component* (GCC) would emerge wherein a finite fraction of the vertices in the graph would be connected. Such a component

is now known in the parlance of percolation theory as a *percolating cluster*. Below that critical value, we can expect to see only isolated small components and no component that spans the graph.[3] The most extraordinary aspect of this paper, however, is that the authors then proceeded to discuss what implications these findings might have for (*i*) a network of neurons, (*ii*) the spread of an epidemic disease through a society, and (*iii*) a problem relating to genetic diversity. Such concern with real-world applications was remarkably prescient of far more recent studies.

Inspired by Rapoport's pioneering work the political scientist Ithiel de Sola Pool (1917-1984) and the mathematician Manfred Kochen (1928-1989) began, in the mid to late 1950s, to consider the application of probabilistic methods to interpersonal contact networks. In this endeavour they were perhaps the first to formally model individual people and their acquaintanceships with one another (sans any particular dynamics) as vertices and edges in a graph. It is widely accepted (see [12, 110, 145]) that de Sola Pool and Kochen's major paper on this topic [37] was distributed amongst their peers in the social sciences in a preprint form as early as 1958 (around the time when Erdös and Rényi were first formulating random graphs); however, the authors, apparently dissatisfied with the treatment they had given, did not consent to the publication of this work until 1978. Despite their misgivings, the questions raised in this paper are some of the most profoundly influential ever considered in the field of social networks and continue to be of interest to researchers to this day. For example, they were the first to ask:

i) How much contact is there between people in different social strata and community groupings?

ii) Does having a greater number of direct acquaintanceships; i.e., a higher *degree*, imply greater personal influence?

iii) What is the expected length of the shortest chain of intermediaries between two people chosen at random?

They were also the first to recognise the importance of the distribution of degrees, and the shape and mean of this distribution in addressing such questions.

Of all the conclusions de Sola Pool and Kochen drew from their extensive analysis the most surprising was that there are on average only two intermediaries between pairs of randomly chosen individuals anywhere on Earth, and furthermore that social stratification does not significantly affect this result. At the time this must have seemed highly implausible; and this may

3 This result is sometimes misattributed to Erdös and Rényi who independently discovered it almost a decade later [56].

be the reason why the authors delayed publication for so long. However, while they were deliberating over the validity of their assumptions one of the people they had shown their preprint to went ahead and carried out a series of experiments to see if they were right.

The iconoclastic American social psychologist Stanley Milgram (1933-1984) was the first person to provide strong empirical evidence for what is now known as the *small-world* phenomenon. In slightly more general terms than those in which it is expressed in [37], this is the hypothesis that any two randomly chosen individuals, from any corner of the earth, can be connected by a short chain of acquaintances — the average length of this chain is still a matter of contention.

In the most famous of his experiments Milgram, with the aid of his collaborator Travers, [100, 139] sent letters to 296 people chosen at random from the joint populations of Nebraska and Boston. Those who received letters were asked to pass them on to another randomly chosen target individual located in Massachusetts. The crux of the experiment was the condition imposed on the carriers that they could only pass their letter on to a first name acquaintance who, for whatever reason, they felt was closer to the target than they themselves. Only 64 out of the original batch of 296 letters reached the target individual. Travers and Milgram [139] calculated that those letters that did arrive at the target took an average of 5.2 steps.[4] Although this is slightly higher than the value predicted in [37], this result still seemed to provide some minor evidence at least that we do indeed live in a small world. Certainly it propelled the idea of social connectivity into the popular consciousness, and more importantly provided the first tangible demonstration of the potentially fascinating role played by network topology in real world complex systems. The flaws in the experiment, not least of which being the bias that was almost certainly introduced in Travers and Milgram's calculations, are well documented [94, 112]; however, its landmark status remains as it continues to stimulate important research [45, 93].

Milgram's work also foreshadowed how technological barriers would hamper progress in the direction of empirical verification. For much of the 20th century data sets remained difficult to compile and were generally too small to be the subjects of any meaningful statistical analysis. As alluded to earlier, the availability of relatively cheap personal computers capable of storing and quickly analysing vast amounts of data has been crucial in facilitating the recent explosion of interest in networks in general. The limitations imposed on the investigations of earlier researchers, like

4 The phrase *six degrees of separation* was coined in reference to this many years later in the play of that name by John Guare [82].

Milgram and Rapoport, no longer exist for today's network scientists. Systems as diverse as the co-appearance network of Hollywood actors [8, 147]; scientific citation networks [81, 106, 123]; the electrical power grids of the United States [8, 42]; metabolic [90, 143] and genetic [103, 132] regulatory networks; and email logs [53, 140] and telephone call records [2, 4], amongst others, have all been measured and analysed, to a greater or lesser extent, over the past decade or so.[5]

Of course, the World Wide Web [6, 22, 95] and its underlying hardware, the Internet [62, 98, 120], have also been the subjects of intense scrutiny in this regard.[6] Since their respective inceptions both of these networks have grown at astonishingly fast rates, and in largely unplanned and unregulated manners [1, 68]. These features have made them both unfathomably complex and, therefore, the ideal objects of study for network scientists looking to test their theories. However, the significance of these particular systems extends far beyond academia. Nothing so characterizes our socio-technical era as the increasing pervasiveness of the Web and related technologies, such as search engines, social networking websites, and VoIP communication services, in our everyday lives. Undoubtedly, our awareness of such developments and our concerns for where they might lead contribute to our renewed interest in networks and our eager reappraisal of the once outlandish hypotheses of the social scientists of the 1950s.[7]

Finally, it is not unfair to say that the novelty and long-term value (if there is to be any) of the new wave of network based analyses hinges on the ability and willingness of its practitioners to combine the rigour of graph theory and other branches of mathematics and physics with the perceptiveness of studies in the so-called soft sciences. As we have attempted to convey in this historical sketch, for many years there has been an artificial disjunction imposed upon the study of networks by the confrontation of different academic cultures. The highest goal of the new endeavour, in our view, is to provide fundamental theoretical arguments to explain the structural and dynamical commonalities, or (slightly more optimistically) unifying traits, observed in the networked systems all around us, whatever their categorization. It is in this spirit that the research presented in this thesis concerning the phenomenon of cascades has been conducted.

5 See the review articles [5, 48, 110] and books [18, 23, 46, 49, 112, 113] for comprehensive lists of references.

6 Ibid.

7 For instance, when one considers that the 44th President of the United States, Barack Obama, has both *Facebook* and *Twitter* accounts the idea of a small world does not seem so implausible after all.

1.3 THESIS ORGANIZATION

This thesis will provide an account of the research carried out by me over the past three years as a member of Prof. James Gleeson's SDCS research group. The focus of this work has been to model cascading processes on complex networks by extending, in a number of different ways, the analytical framework of Gleeson and Cahalane [73]. For much of our presentation it will be helpful for the reader to think of these cascades as taking place on social networks, and the terms in which we express our analyses will often reflect this. However, all of our results are based solely on mathematical arguments, and their applicability is not limited to any one domain: sociology, epidemiology, technology, and finance all present network-oriented problems for which our results may be relevant.

There are two broad themes in this thesis. The first concerns the modelling of cascade dynamics on locally tree-like random graphs (Chapter 3). The second concerns the modelling of cascade dynamics on random graphs with realistically high levels of *clustering* (Chapters 4 and 5).

Having established, in our opening review, the historical context against which to gauge our contribution to the study of networks, in Chapter 2 we provide an account of the definitions and concepts fundamental to contemporary work in the field. This includes a partial glossary of important mathematical terms and reviews of some of the major theoretical developments in modelling both network structure and dynamics.

In Chapter 3 we introduce the tree-based approach of [73] for calculating the expected cascade size and the position of the cascade threshold. We offer an extension of this approach as part of a detailed investigation into the role of so-called *influentials* in the spread of information in society. We also discuss the broader effectiveness of the tree-like approximation of network structure for modelling real-world processes. This includes a review of some of our recently published work on the subject [99].

In Chapter 4 we discuss at length the phenomenon of clustering. We review two recent structural models which have presented methods of creating ensembles of highly clustered random graphs [71, 111]. Our comparison of these two models contains details of a publication of ours concerning the effects of clustering on cascades [76]. This sets the scene for Chapter 5, where we will demonstrate how the framework of [73] can be extended once again to create analytical models of cascade dynamics on each of the clustered graph ensembles of [71] and [111]. These models contain within their scope a range of processes including Watts's model, SIR contagion dynamics, and site and bond percolation. One of these models has appeared previously in publication [85]. The other is currently in preparation [83].

Finally, Chapter 6 offers a summary of our work, and overall conclusions.

2

DEFINITIONS AND CONCEPTS

We begin this review chapter with a glossary of some important terminology in the contemporary study of networks. These terms will be part of our vocabulary throughout our presentation, and we will refer back to this section in order to clarify our meaning, whenever it is deemed necessary.

2.1 SOME NETWORK TERMINOLOGY

Vertex: The fundamental unit of a graph. An abstract mathematical entity which we use to represent some distinct part of a networked system; e.g., a person in society, or a page on the Web. Each vertex is given a label $i \in \{1, \ldots, n\}$, where n is the total number of vertices in the graph. We refer to n as the *size* of the graph. Other field-specific terms for vertex are *node* (computer science), *site* (physics), and *agent* (sociology).

Edge: A line between two vertices used to represent a connection between the corresponding parts of the networked system; e.g., an acquaintanceship between two people in society, or a hyperlink leading from one webpage to another. The terms *link* (computer science), *bond* (physics), and *tie* (sociology) are also sometimes used.

Adjacency matrix: A matrix representation of a graph indicating which vertices are adjacent to which others. The elements of the matrix are denoted a_{ij}, where $i, j \in \{1, \ldots, n\}$. For the most part in this thesis, $a_{ij} \in \{0, 1\}$, and signifies the simple presence or absence of an edge between vertices i and j. Furthermore, the link is usually symmetric: $a_{ij} = a_{ji}$. This means that the relationship that j bears to i is the same as that which i bears to j. However, there are some more intricate possibilities; for example, weighted edges, where bonds of different strength are represented by letting a_{ij} take values on the real number line, and/or directed edges, which allow for non-reciprocal relationships between vertices. Unless stated otherwise, we ignore the possibility of self-edges (or *loops*).

Degree: The degree k_i of a vertex i is the number of edges incident to i. For directed graphs (or *digraphs*); i.e., graphs with directed edges, we define the out-degree, k_i^{out}, as the number of edges rooted at i, pointing away

from i, and the in-degree, k_i^{in}, as the number of edges pointing towards i.

Mean degree: The average degree of the vertices in the graph: $z = \frac{1}{n}\sum_{i=1}^{n} k_i$. That is, the average number of connections per vertex. Also, the first moment of the degree distribution.

Degree distribution: For a given network topology, the degree distribution prescribes the probability, p_k, that a vertex, chosen at random, has degree k. In- and out-degree distributions are similarly defined for directed graphs. Power-law degree distributions are often observed in empirical data.

Degree-degree correlation: A measure of the correlation in the degrees of the vertices at either end of a randomly selected edge. This gives us an insight into the extent to which two vertices of certain degrees are related to each other. It is usually measured by a version of the Pearson correlation coefficient for vertex degrees, defined as

$$r = \frac{\langle kk' \rangle_e - \langle (k+k')/2 \rangle_e^2}{\langle (k^2+k'^2)/2 \rangle_e - \langle (k+k')/2 \rangle_e^2}, \tag{2.1}$$

where $\langle \cdot \rangle_e$ denotes the average over all edges and (k, k') denotes the degrees of the two vertices at either end of an edge [107].

Geodesic path: This quantity, usually denoted L_{ij}, is the minimum number of edges one must traverse in reaching a specified target vertex j from a given starting vertex i. The average geodesic path length, or mean intervertex distance, taken over all pairs $\{i, j\}$ is given by

$$L = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geqslant j} L_{ij}. \tag{2.2}$$

Average geodesic path lengths tend to be much longer on very regular graphs, like lattices, than they are on more random topologies, in which it is often possible to find shortcuts from one vertex to another.

By convention, the term *small-world network* (SWN) designates those networks in which the average geodesic distance scales as $L \sim \log(n)$ or slower as the size of the network, n, diverges ($n \to \infty$). This logarithmic scaling can be proved for a variety of real and model networks (see [16]).

Connected component: A connected subgraph; that is, a subset of the vertices of a graph such that there exists some path through the graph connecting any two vertices in this subset. If it exists, the component that contains

the majority of the vertices is called the giant connected component (GCC). In practice we tend to think of graphs as dynamic entities, evolving over time by the addition of new vertices and edges. Hence, the GCC is more commonly defined as a connected component that spans a finite fraction of the vertices in an infinitely large graph ($n \to \infty$). This definition is closer to the idea of a *percolating cluster*.

Clique: A fully connected subgraph. That is, a subset of the vertices of a graph in which each vertex is connected to all of the others. A clique of c vertices is called a c-clique; triangles are 3-cliques. Cliques are, by definition, maximally clustered.

Clustering: This refers to the propensity for vertex triples to be fully connected. Real-world networks tend to have very high levels of clustering in comparison to classical random graphs. This is often one of the most distinguishing features of real network topologies. For example, it is a well known feature of social networks, in general, that if agent A is connected to agent B, by a bond of friendship for instance, and agent B is in turn connected to agent C, then it is highly probable that agents A and C are also connected. Sociologists call this phenomenon *transitivity* or *triadic closure*.

Clustering coefficient: A natural way to measure clustering is as the probability that a randomly chosen connected triple of vertices, $\{i, j, k\}$, form a triangle. This probability is called the clustering coefficient and is given by

$$C_1 = \frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triples of vertices}}. \tag{2.3}$$

This is a global measure of the clustering in a graph [114].

Local clustering coefficient: An alternative definition introduced by Watts and Strogatz [147]. This is a measure of clustering on a local level. We define this as the probability that a triple connected to a randomly chosen vertex, i, form a triangle:

$$c_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}. \tag{2.4}$$

The average of the local clustering coefficient over the entire graph provides a second global measure: $C_2 = \frac{1}{n}\sum_i c_i$.

Degree-dependent clustering coefficient: The average of the local clustering coefficient over the class of vertices of degree k [129, 142]:

$$c_k = \frac{1}{n_k} \sum_{i \in \Upsilon(k)} c_i, \tag{2.5}$$

where n_k is the number of vertices of degree k in our graph and $\Upsilon(k)$ is the set of such vertices. If we calculate c_k for each degree k we can construct a *clustering spectrum*.

Generating function: A formal power series, whose coefficients correspond to a specific sequence of numbers [149]. For example, the ordinary generating function of the sequence a_n, where $n \in \mathbb{N}$, is

$$G(a_n; x) = \sum_{n=0}^{\infty} a_n x^n. \tag{2.6}$$

Similarly, the ordinary generating function of a two dimensional array of numbers $a_{m,n}$, where $m, n \in \mathbb{N}$, is $G(a_{m,n}; x, y) = \sum_{m,n=0}^{\infty} a_{m,n} x^m y^n$.

If the coefficients in Eq. (2.6) are a sequence of normalized probabilities q_k, such that $\sum_{k=0}^{\infty} q_k = 1$, then $G(q_k; x)$ is the probability generating function of the *probability mass function* (PMF) $q_k = \Pr(K = k)$, where K is a discrete random variable. Similarly, $G(q_{k,j}; x, y)$ can encode the joint PMF $q_{k,j} = \Pr(K = k \text{ and } J = j)$, for discrete random variables K and J.

The probability generating function of the degree distribution of a graph, $G(p_k; x)$, has been used extensively in network-based analyses [24, 114, 144]. Some useful features of this representation are the following [114]:

Derivatives. The probability p_k of a random vertex having degree k, is given by the kth derivative of $G(p_k; x)$ evaluated at $x = 0$:

$$p_k = \frac{1}{k!} \frac{d^k G(p_k; x)}{dx^k} \bigg|_{x=0}. \tag{2.7}$$

Moments. The moments of the degree distribution, $\langle k^n \rangle$, are given by

$$\langle k^n \rangle = \left[\left(x \frac{d}{dx} \right)^n G(p_k; x) \right]_{x=1}. \tag{2.8}$$

For example, the first moment $z = \langle k \rangle = G'(1)$.

Powers. The distribution of the sum of the degrees of the vertices in a randomly selected subset of size m is generated by $\left[G(p_k; x) \right]^m$. For example, the coefficients of $\left[G(p_k; x) \right]^2$ are the probabilities that the degrees of two vertices sum to $0, 1, 2$, etc.

2.2 NETWORK MODELS

In network theory, a *model* is a prescription for the creation of an ensemble of graphs, usually with a view to capturing the structure of some networked system. In this section we review, in order of publication what are, from our perspective, the most important models to date. We begin with the one that pointed the direction towards the study of real-world networks in the first place: the Poisson random graph model of Erdös and Rényi.

2.2.1 *Poisson Random Graphs*

We have seen previously how, in a series of papers published in the late 1950s and early 1960s, Erdös and Rényi introduced probabilistic methods to graph theory, thereby creating the theory of random graphs. As part of this major contribution they also created one of the earliest, and most comprehensively studied, random graph models [57], the basic formulation of which can be given as follows: Starting with n disconnected vertices, add edges between each possible pairing of these vertices with independent probability, p.[1] The ensemble of graphs created in this way is called $G_{n,p}$.[2] Since there are $n-1$ choices for the set of edges incident to any vertex, the average degree of a $G_{n,p}$ graph is $z = p(n-1)$, and it can be shown that as the size of such a graph diverges $(n \to \infty)$ its degree distribution converges to a Poisson distribution with mean z (hence the name):

$$p_k = \frac{z^k e^{-z}}{k!}. \tag{2.9}$$

Average geodesic path lengths in $G_{n,p}$ graphs scale as $L = \log(n)/\log(z)$ as $n \to \infty$ [110]. This logarithmic, or *small-world*, scaling is a property that they share in common with many real networks (see Section 2.2.3 below). However, the similarities end here.

Poisson random graphs are, in fact, poor models of networked systems. This should come as no surprise since, as we have seen, despite the significance of their work in bridging the gap between traditional graph theory and the empirical study of networks, Erdös and Rényi had not intended to create a realistic network model. Instead, their graphs represent completely random structures in which the presence of connections between

1 While a graph constructed in this way shares similarities with those investigated by Solomonoff and Rapoport in [133]. The analytical approach adopted by those authors was by no means as rigorous as that of Erdös and Rényi, and they are, for the most part, not credited with the creation of this model.

2 This ensemble is not the same as the $G_{n,M}$ ensemble, discussed earlier, which contains all graphs that can be created having n vertices and exactly M edges. For further clarification of the precise distinctions between the two see [15].

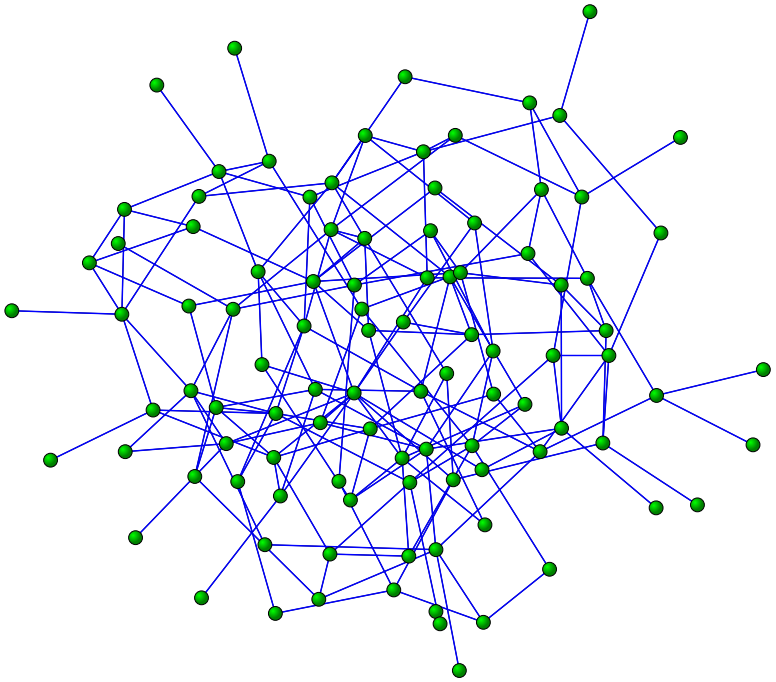


Figure 2.1: A Poisson random graph with $n = 100$ and $z = 4$.

components is entirely down to chance. In principle, such graphs should serve us no better as models of the structure of real-world systems than simple lattices, especially since we are interested in systems that exhibit the hallmarks of complexity. Concerning this point, it is important to recognize that the complexity of a system is commensurate to the degree of difficulty in uncovering and understanding the fundamental principles that govern its behaviours, and not to how complicated those behaviours may appear.[3] Therefore, while the graph presented in Fig. 2.1 certainly looks complicated, considering the simplicity of its construction, we cannot call it complex.

In the sociological domain, our everyday experience indicates that beneath the apparent chaos of social interactions there must be rules that determine the formation and dissolution of ties. While, of course, most of these rules are as yet to be discovered; nevertheless, we can all appreciate that things like bonds of friendship, for example, do not come about entirely by chance. Setting aside for a moment the implications of the small-world hypothesis, it seems intuitively obvious that one is much more likely to be acquainted with those from one's own socio-economic background, workplace, geographical locality, or race. Related to this is the phenomenon of high transitivity, or clustering (see Section 2.1); a feature conspicuously absent from the Poisson random graph model since the

3 In this sense complexity is as distinct from complete randomness as it is from strictly determined order.

probability of an edge connecting two of the neighbours of a randomly chosen vertex, i, in a $G_{n,p}$ graph is p, regardless of the fact that they share a mutual neighbour in i. Thus, the clustering coefficient of these graphs is simply $C_1 = p = z/(n-1)$ which, for fixed z, falls off as n^{-1} as $n \to \infty$.

2.2.2 *The Configuration Model*

In addition to producing graphs with vanishing clustering coefficients, another obvious shortcoming of Erdös-Rényi's model is the inflexibility of the Poisson degree distributions that it creates. Evidently, there is no such *a priori* restriction on the structure of a real network. A realistic model ought to be able to produce a broader range of distributions, or at least something with greater variability about its mean. In 1995 Molloy and Reed offered a very elegant solution to this problem, with the publication of their *configuration model* [104] for the creation of ensembles of random graphs of arbitrary degree distribution.

To construct a random graph using this model our first step is to actually pick a desired degree distribution, $\widehat{p_k}$. From this, we then draw a degree sequence $\{k_i\}$, prescribing the degree of each individual vertex, $i \in \{1 \dots n\}$. Using this sequence we create another list in which the label, i, of each vertex appears exactly k_i times. Finally, to construct a realisation of a random graph we pair up the elements of this list uniformly at random and place the number 1 (for an undirected unweighted edge) in an adjacency matrix in the positions indexed by these pairs.[4] The simplest way to visualise this process is as connecting together half-edges, or *stubs* of edges, to form complete edges between pairs of vertices. (Hence, the list of labels which we have just referred to is more commonly called a *stubslist*.) The actual degree distribution of the resulting random graph, p_k, will not be precisely the same as $\widehat{p_k}$. However, the match between the two improves as n increases, and they become indistinguishable as $n \to \infty$.

In theory, the configuration model can produce graphs fitting any well-defined degree distribution including those drawn from real networks. This makes it very powerful. Many of the most successful network models seen in recent years have been built around its basic framework. In fact, it is difficult to imagine much of the recent progress in this area having occurred without it. It is rather unfortunate, then, that it suffers from precisely the same drawback as the Poisson random graph model in that the level of clustering vanishes as $n \to \infty$. This is true regardless of the choice of degree distribution and, therefore, this significantly diminishes its usefulness as a model of real networks. That is, of course, unless some modification can be

4 Any entries that are not set to 1 at this step are set to 0, indicating the absence of an edge.

devised to bring clustering into the model. Some progress has been made on this front in recent years by ourselves and others, and we will discuss this topic in great detail later in Chapters 4 and 5 when we present our work concerning highly clustered networks.

Despite the lack of clustering, the configuration model has continued to be of interest to network scientists and many of its properties, including, for example, the criterion for the appearance of the giant connected component, have been identified and analysed. Some of the results pertaining to these properties were derived rigorously like in [29], while others were found using heuristics and approximations. Newman *et al.* [114], for example, exploited the lack of clustering in configuration model graphs to derive a generating function formalism from which they obtained quite a number of fundamental insights.

2.2.3 *Small-World Networks*

The fundamental shortcoming of both Erdös-Rényi and configuration model random graphs is their lack of structure. As we have already pointed out, complex networks are not entirely random. On the other hand, nor are they completely ordered. By their very nature, they contain elements of both regimes. One of the most conspicuous ways in which order manifests itself is through the phenomenon of high clustering. Unfortunately, both of the aforementioned models lack this important feature.

To address this problem, in 1998 Watts and Strogatz introduced the *small-world network* (actually, a new class random graphs) [147]. This has proven to be one of the most influential publications of modern network theory, inspiring much of the current wave of interest in the area.[5] An appealing feature of this model is the fact that it allows one to interpolate continuously between ordered and random topologies simply by tuning a single parameter. The original algorithm is remarkably simple:

i) Arrange n vertices in a ring; i.e., a 1-dimensional lattice with periodic boundary conditions.

ii) Join each vertex to its nearest neighbours within a specified range $d \in \mathbb{N}$ such that all vertices have the same degree $z = k = 2d$.

iii) Rewire a fraction p of the edges. Rewiring is achieved by disconnecting one end of an edge and reconnecting it to a different vertex, chosen at random (see Fig. 2.2).[6]

5 As of October 3 2011, the Science Citation Index counted 6,298 citations of this paper.

6 Multiple edges between pairs of vertices, and edges with both ends connected to the same vertex (*loops*), are not allowed.

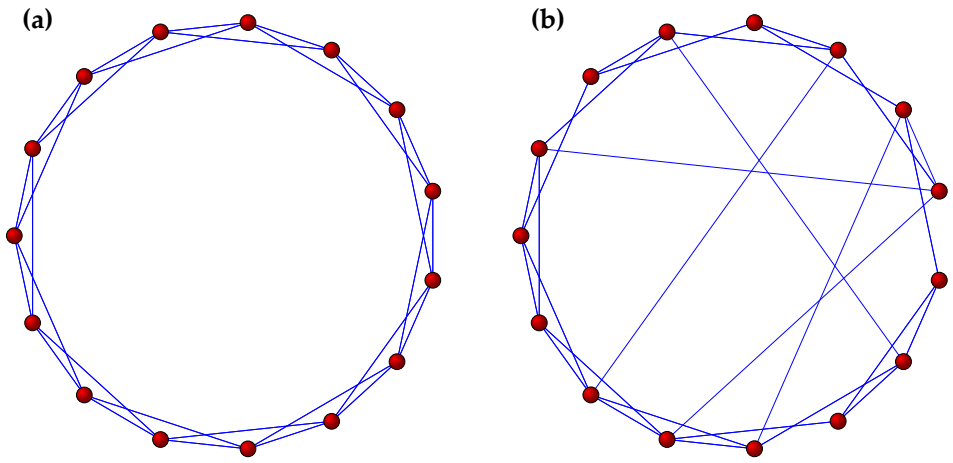


Figure 2.2: In **(a)** we have a ring of 15 vertices, each connected to its nearest and second nearest neighbours ($z = 4$). In **(b)** some edges have been randomly rewired to create a small-world network.

The value assigned to the parameter p determines the complexity of the resulting graph topology. At $p = 0$ we have a perfectly ordered ring lattice. This type of structure is essentially the polar opposite of a Poisson random graph: it has both a long average geodesic path length and a high clustering coefficient. A straightforward analysis [110] shows that $C_1 = (3k-3)/(4k-2)$, which scales as $C_1 \simeq \frac{3}{4}$ as k gets large, and also that $L \sim n/4k$ as $n \to \infty$. Increasing the value of p introduces edges between vertices outside of the original range, d. Doing this quickly destroys the lattice structure, dramatically reducing the values of C_1 and L. At $p = 1$ we achieve a fully randomised graph with a degree distribution similar, but not identical, to a Poisson distribution with mean z. (We will discuss this issue further below.)

Naturally, Watts and Strogatz were particularly interested in the structure of their graphs in the intermediate region between $p = 0$ and $p = 1$. Relying primarily on numerical simulations, they discovered an entire range of p values for which both high clustering and short average geodesic path lengths could coexist. Specifically, they found that rewiring between $p = 0.01$ and $p = 0.05$ of the edges decimated the value of L, while maintaining close to maximal C_1.

They then presented an argument suggesting that these are the two fundamental ingredients underlying the complexity of real-world networks. In other words, that most networks can be accurately modelled by graphs constructed in or around this parameter regime. This claim was bolstered by empirical evidence derived from (*i*) a co-appearance network of Hollywood film actors (actors who appeared in the same film were directly linked); (*ii*) the power-grid of the Western U.S.; and (*iii*) the neuronal network of the

nematode *C. elegans*. All three were found to have measurably short L and high C_1. That is, they were all highly clustered *small-worlds*. The authors predicted (not unreasonably) that many other networks would be found to share the same characteristics. And so, for a fleeting moment it looked as if the problem of creating a realistic network model had been solved.

Of course, the reality of the situation would turn out to be much more complicated. The Watts-Strogatz model, like any other, suffers from a number of significant drawbacks. Firstly, the rewiring process used is not particularly amenable to a mathematical analysis; most of the results given in [147] were found through numerical simulations. The difficulty arises primarily because removing existing edges can result in the graph fragmenting into disconnected components. To combat this a number of variants of the original algorithm have been suggested. The first, and most popular, solution was offered by Newman [115]. In his version of the model instead of rewiring existing edges one simply adds extra ones between randomly chosen vertices. Graphs produced in this way have similar properties to those produced by Watts and Strogatz but are also guaranteed to consist of a single connected component.

Secondly, as mentioned above, the degree distributions obtained at $p = 1$ are not quite right. The problem is that they are too narrow; i.e., they lack variance. Given that we have a delta spike $p = 0$, and that the variance increases as p is increased, until it reaches its maximum at $p = 1$, it is not too difficult to see that the degree distributions generated by this model will never be broader than a Poisson distribution for any $p \in [0, 1]$ [110]. One may justifiably ask: Is this really such a problem? Certainly, at the time few could have predicted the next dramatic twist in the tale of network theory: the discovery that many networks (perhaps even most) have broadly heterogeneous degree distributions, which are more accurately fitted by various types of power law than by any kind of Poisson-like distribution. Thus, despite possessing the realistic features of high clustering and short path lengths, the Watts-Strogatz model would soon be outmoded.

2.2.4 *The Barabási-Albert Model*

Around the turn of the millennium strong empirical evidence began to emerge [6, 8, 123] that real degree distributions do not adhere to the simple centrally-peaked form which had hitherto been the archetype. Instead, it was revealed that they are often heavily right-skewed, with many vertices of very low degree and a few of very high degree. More than that, in between these two extremes they quite often lack any easily identifiable peak by which to characterise the majority of vertices. In other words, the

distribution of degrees can be so heterogeneous that the mean value z is virtually useless as a descriptor of the graph topology. This heterogeneity is typically of such an extent that a large portion of the data, usually located towards the tail of the distribution, can be accurately fitted by a power law of the general form

$$p_k \sim L(k)k^{-\gamma}, \tag{2.10}$$

where $\gamma > 1$, and the coefficient $L(k)$ is a slowly varying function satisfying $\lim_{k\to\infty} L(tk)/L(k) = 1$, with t constant. (In many cases $L(k)$ is defined as some simple normalization factor C [30].)

In 1998 Redner [123] carried out a study of scientific citation networks which revealed that the probability of a journal article being cited by k others decayed as $k^{-\gamma}$ with an exponent $\gamma \approx 3$. In graph theoretical terms, treating articles as vertices and drawing directed edges from each article to those it cites, we would say that the in-degree of vertices in this network is distributed according to a power law with said exponent and $L(k) = C$. One year later a study appeared which suggested that the World Wide Web is also a power law distributed network [6]. In this case webpages are vertices and hyperlinks are edges. The network is directed since there may be a link from one page to another but no backlink. The authors found that both the in-degree and out-degree of the Web decayed as power laws with exponents $\gamma \approx 2.1$ and $\gamma \approx 2.45$, respectively. Of course, these results were not derived from an analysis of the Web in its entirety. Rather, they were extrapolated from measurements made on the the portion of the Web hosted on the `nd.edu` domain, which at that time numbered 325,729 documents and had 1,469,680 links. This was assumed to be a representative sample of the Web as a whole. Similar results were obtained from the co-appearance network of actors ($\gamma \approx 2.3$), electrical power grids ($\gamma \approx 4$), the Internet ($\gamma \approx 2.5$), and telephone calling graphs ($\gamma \approx 2.1$). Yet further examples were unearthed the following year in an extensive survey carried out by Amaral *et al.* [8].

In view of the substantial evidence provided by these (and many other) empirical studies it soon became apparent that power law degree distributions are ubiquitous in both man-made and naturally occurring networks (see Table 3.1 of [110]). This was particularly exciting because of the interesting physical properties that power laws often imply. For one, they are *scale invariant*. This means that a rescaling of the function's argument will cause only a proportional scaling in the function itself. For example, considering the simple form presented in Eq. (2.10) we see that multiplication by a constant factor will not change the overall shape of the function. Because of this, power law distributed networks are often termed *scale-free*. In many contexts, this property can signify the presence of deep structural

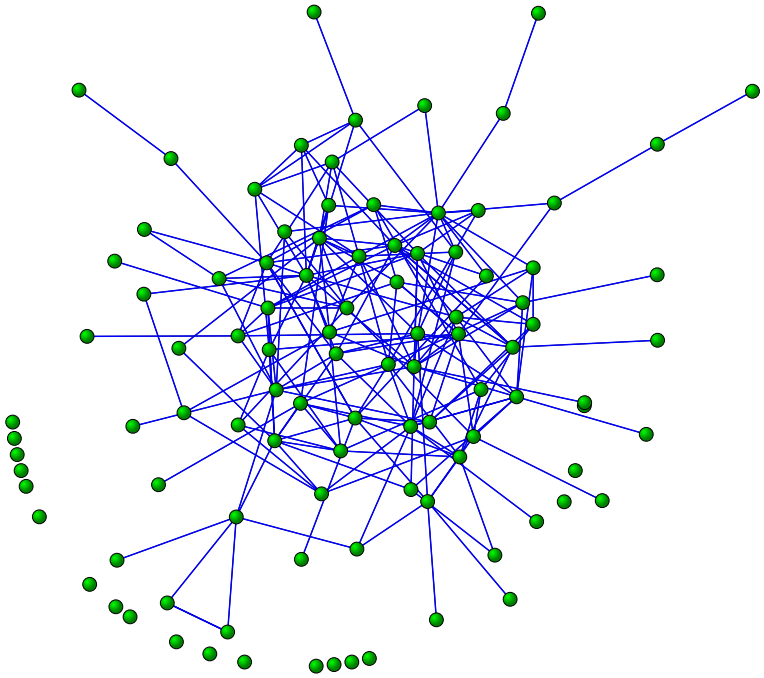


Figure 2.3: A scale-free (i.e. power law distributed) network of 100 vertices with 200 edges. While the average degree is $z = 4$, many vertices have fewer than 4 neighbours and a few have very many more neighbours.

symmetries. From a complex systems point of view it is often seen as the hallmark of an underlying organizational hierarchy or stochastic process. This boded well for the creation, finally, of a simple unifying model which would capture all of the salient features of real networks.

First off the mark, in this respect, were Barabási and Albert, with the publication in 1999 of their model for the creation of a class of scale-free random graphs [13].[7] Two rather simple observations laid the foundation for this model. Firstly, real networks are not static, instead they are constantly growing by the addition of new vertices and edges; and secondly, new vertices tend to attach preferentially to already well connected ones.[8] The first point seems fairly intuitive; however, the second takes some justification, especially given the generality of the claim. The authors argued that the World Wide Web and other socio-technological systems naturally evolve in this way, and that the same mechanism can explain power-law degree distributions in other domains. In hindsight, whether or not this is actually the case is debatable, and in some quarters the model has been criticized as lacking in sufficient rigour [16].

7 As with much else in contemporary network theory this model has closely related precedents in the sociometric studies of the mid 20th century. In this case, the work of Herbert Simon [130] and Derek de Solla Price [38, 39] is the most directly relevant.

8 This phenomenon was called *cumulative advantage* by de Solla Price [39] in relation to citations in scientific papers.

Nonetheless, using these two ingredients the authors derived a simple growth process that was capable of producing scale-free graphs with power-law exponents $\gamma \in [2, 3]$. Significantly, not only was the end product of this process scale-free, but the statistics of the power-law degree distribution became stationary as the network evolved. This was considered an important requirement for modelling purposes since real scale-free networks are clearly not finished in any meaningful sense of the word, and yet are often observed to possess constant power-law exponents in the above range.

In spite of all this progress, the Barbási-Albert model is not much closer to a universal network model than any earlier attempt. The scale-free graphs that it produces lack significant levels of clustering. This is just one of its major shortcomings. Numerous refinements and modifications of the original algorithm exist. The most significant of these are discussed in detail in the review article [5].

The list of network models presented in this section is by no means exhaustive. Rather, it serves to highlight the four most fundamental contributions to the subject so far. As we have seen, each of these four models capture some of the features of real-world networks, while, on the other hand, none capture all. The phenomenon of high clustering, in particular, has proven to be remarkably elusive. We will discuss this issue further in Chapter 4. For now, however, we move away from the topic of network structure and introduce the other major strand of network theory: the study of the processes that take place on networks. The mathematical description of these processes will be expounded upon in the next chapter.

2.3 PROCESSES ON NETWORKS

While modelling efforts, such as those discussed in the previous section, may help us to understand the topological characteristics of empirically observed networks, these constitute merely the first steps towards our ultimate goal of developing a comprehensive understanding of the systems built on those networks. Unfortunately, to form a complete picture of a system it is not enough to simply uncover the properties of its underlying architecture; we must also provide a detailed description of how this architecture affects the system's functional particulars. Within network theory, studies of the latter are much less well developed than those of the former. As we have shown in the introduction, the reasons for this can be found in the historical development of the theory. Many of the tools and techniques currently used by network scientists to investigate functional dynamics have been appropriated from the long-standing methodologies of other fields of study. Depending on one's point of view this may be

seen as a weakness of the theory or as one of its primary virtues. In support of the latter opinion, there is one area in particular where this synthesizing approach has borne ample fruit, leading to quite a number of important new results, and that is in the study of how things propagate over networks.[9]

Broadly speaking, most of the attention thus far concerning the dynamical properties of networks has focused on the propagation of either information or disease through society, or structural failure in synthetic systems, like the Internet. In respect to these spreading processes researchers have been overwhelmingly interested in determining whether or not small localised uniformities of state or behaviour in a given system are likely to evolve into a large scale uniformity observable across the majority of its components, an event we call a cascade. For example, those interested in the spread of information through society [144, 146] are often concerned with such phenomena as fashions or rumours, which grab the attention of vast swathes of the population in a very short period of time and often die out just as quickly. Similarly, many epidemiological studies [96, 119] have been preoccupied with determining how small isolated outbreaks of disease can quickly develop into pandemics. And, those concerned with the robustness of the Internet, or other such man-made systems, to either random breakdowns [31] or intentional attacks [32], often worry about the extent to which structural failure may propagate throughout these systems.

Investigations of this sort are often complicated by the fact that each of these processes is very different to what one might call traditional diffusive propagation. The spread of things like information, disease, or failure is not conservative. If you have an idea which your friend also adopts, you may still hold on to it. Similarly, if you transmit the flu to someone, that does not mean you are therefore cured. Hence, the diffusion equation or some other mass-conserving equation is, usually, not an appropriate tool for modelling such processes. More often than not, we are forced to apply probabilistic methods in our analyses. Thankfully, as we shall see, networks provide a natural framework on which to do so.

2.3.1 *Failure and Resilience*

The question of resilience has been addressed by network scientists primarily by adapting ideas from the branch of condensed matter physics called percolation theory. According to Bollobás and Riordan [17], this theory was initiated over half a century ago by Broadbent and Hammersley [21] "in order to model the flow of fluid in a porous medium with randomly blocked

9 For a review of the major advances made in this direction see [113].

channels" [17]. Recalling the more familiar language of Erdös and Rényi, when viewed from an abstract perspective, percolation theory can be seen as the study of the component structure of the random graphs obtained by selecting either vertices (site percolation) or edges (bond percolation) independently of each other, and with uniform occupation probability ϕ_s or ϕ_b, respectively. Indeed, a close reading of the literature on the subject reveals that much of it concerns problems similar to those considered by Erdös and Rényi, and Rapoport before them, albeit framed differently. It includes, for example, questions relating to the distribution of small connected components (or *clusters*), and the criterion for the emergence of the giant connected component (or *percolating cluster*).[10]

Since, however, percolation has traditionally been studied by mathematical physicists, many of its classical results have been derived from a mean-field approach where the medium in which the percolating process occurs is represented by some infinite-dimensional structure of minimal complexity. Typically, either one of the following three types of graph have been considered [46]: (i) an infinite dimensional lattice, (ii) a fully connected graph, or (iii) a Bethe lattice.[11] As network scientists, these types of trivial topologies are of limited value to us; rather, our interest in percolation stems from the relatively recent work of Molloy and Reed [104, 105], who investigated the component structure of the infinite undirected graphs of arbitrary degree distribution, p_k, generated from their configuration model.

The most celebrated result presented in [104] relates the mean number of first (z_1) and second (z_2) nearest neighbours in configuration model graphs to the birth of the GCC. By way of some particularly dense mathematical arguments the authors proved that the GCC exists if and only if,

$$\sum_k k(k-2)p_k > 0. \tag{2.11}$$

Extracting the first and second moments of the degree distribution from this expression, it can be written in the form $z_2 > z_1$, where $z_1 = \langle k \rangle$ and $z_2 = \langle k^2 \rangle - \langle k \rangle$. Although this result is only trivially related to percolation (as in this case $\phi_s = \phi_b = 1$), it represents an important precedent for those in the networks community interested in the subject, and some of the most significant work of recent years has been based (though perhaps not always consciously) on various modifications and extensions of Molloy and Reed's approach. We will illustrate this point further in a moment.

10 Two solid introductions to percolation are to be found in [17] and [134]. While the treatment given in [17] is decidedly mathematical, [134] is written from the physicists' point of view.

11 This is essentially an infinite dimensional random regular graph, where regular means all vertices have the same degree. We shall return to this object in next chapter.

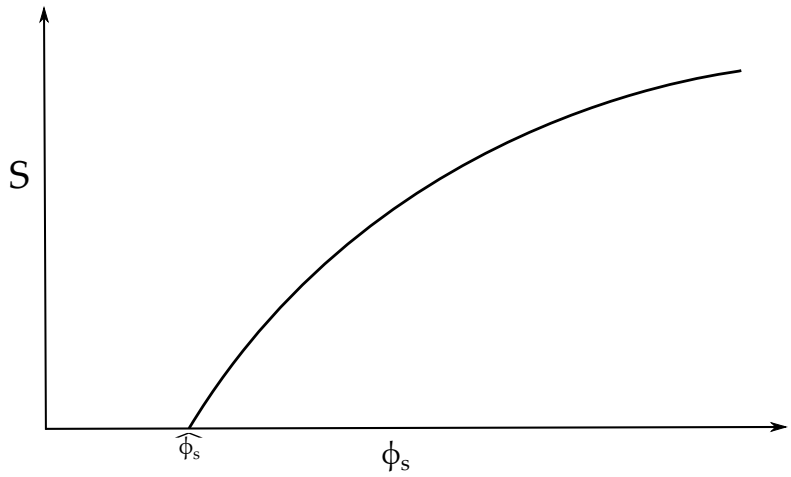


Figure 2.4: The giant connected component appears once some critical occupation probability, $\widehat{\phi}$, has been reached. For uniform site percolation $\widehat{\phi_s}$ usually marks the beginning of a continuous phase transition. S: the relative size of the GCC; ϕ_s: site occupation probability.

First, in case it is not entirely clear to the reader at this stage of our discussion how percolation relates to network resilience, let us state the connection explicitly. Notice that the way we have presented percolation thus far — as a process by which connected components are formed from a set of initially isolated vertices, leading eventually to the appearance of a giant connected component (see Fig. 2.4) — is just one of two valid interpretations. If, instead, we read this picture backwards, so to speak, percolation concerns the breakdown of a giant connected component into a set of isolated vertices, where either sites or bonds are broken with independent probability $1 - \phi_s$ or $1 - \phi_b$. In fact, this is precisely how it is interpreted in most contemporary network-oriented investigations.

Cohen *et al.* [31], for example, modelled the random breakdown of parts in the Internet as a site percolation process with uniform probability of removal ψ_s $(= 1 - \phi_s)$, and thereby derived the following result for the percolation transition on a graph of arbitrary degree distribution:

$$\phi_s z_2 > z_1, \tag{2.12}$$

where z_1 and z_2 are the mean numbers of first and second nearest neighbours in the undamaged version of the graph, respectively. Clearly, this result is a generalization of Ineq. (2.11) above (a point which, in fairness to the authors of [31], they readily acknowledge).

The most original aspect of [31] concerns the resilience of scale-free networks. In support of earlier observations made by Albert *et al.* [7], the authors verified mathematically that, depending on the value of the exponent, networks with power-law degree distributions of the general

form $p_k \sim k^{-\gamma}$ (like the Internet) may be extremely robust to random failures. Specifically, they showed that if $\gamma > 3$, then there will be a phase transition at some point $\psi_s < 1$ at which the GCC disappears; if however, $\gamma \leqslant 3$, no such transition exists. In this case, in order to destroy the GCC one would have to remove almost every vertex from the graph.

Naturally, the next question that follows from this result is: What will happen if, instead of random removals, we target vertices or edges of certain types? Of particular interest is the potential impact of targeting only the highest degree vertices for removal. This problem was addressed early on by Albert *et al.* [7], but only through the use of numerical simulations. The first, and most elegant, analytical treatment was given by Callaway *et al.* [24]. The authors of this paper presented a number of remarkable results that can be applied equally to the problems of random breakdown and targeted attack. Using a generating function approach in which the probability of deletion of a vertex is assumed to be some arbitrary function of vertex degree, $\psi_s = \psi_s(k)$, they showed that it is possible to calculate exactly not only the position of the percolation threshold and the expected size of the GCC[12], but also the size distribution of non-critical components below the threshold. In relation to intentional attack they provided theoretical backing for the simulations of [7] by showing that while scale-free networks may be resilient to random breakdowns, they are extremely vulnerable to the removal of their highest degree vertices. In some cases removing only 1% of these top vertices is enough to destroy the GCC [24].

Targeted removal has subsequently been studied in various guises by other authors, including Cohen *et al.* [32], and occasionally in rather complicated and interesting ways [47, 118]. Note also that the application of ideas from percolation to problems of network robustness (or growth) currently extends far beyond the foundational work presented in this section. The concept of *explosive* percolation, for example, which lays outside scope of this thesis, is a particularly hot topic [3, 124].

2.3.2 *Epidemics and Rumours*

Although it is an interesting process in its own right, when compared to the types of dynamics which we observe from day to day on social networks percolation (whether random or otherwise) can appear rather mundane. In view of the fact that the world consists of living persons who (unlike technological systems) each hold at least some degree of autonomy, removing with predetermined probability an undifferentiated set of vertices or edges from a graph becomes an artificially simplistic modelling device.

12 A result equivalent to this was derived by Molloy and Reed in [105].

To illustrate, consider the spread of disease within a population. If we were to apply percolation directly in this case, we would be assuming that individuals are in either of two acquiescent states: diseased or not diseased. However, thanks to the work of epidemiologists we know that real contagion dynamics are much more subtle than this, and depend upon an array of different factors (mostly human) including, for example, the mobility of individuals; their community groupings; and various immunities.

In recognition of these limitations, attempts to model outbreaks of infectious disease by network scientists usually rely on the adaptation of classical models borrowed from the epidemiological literature. The two that appear most often are named, respectively, *susceptible- infective-recovered* (SIR) and *susceptible-infective-susceptible* (SIS), after the different iterations of state considered in each case. We shall not delve into the details of work in this area since it deals in concepts analogous to those discussed in relation to resilience above. For example, the fraction of infected individuals (or *prevalence*) in SIR corresponds closely to the relative size of the GCC. Similarly, the idea of an epidemic threshold [80] parallels that of a critical occupation probability. In fact, it has been shown by Newman [108] that a generalization of the SIR model applied to random graphs of arbitrary degree distribution can be mapped directly onto bond percolation.

Instead of disease we will focus our discussion of human-driven dynamics on the spread of information, and the various cascading phenomena which it engenders in the political, economic, and cultural domains. Though also closely related to percolation, this topic has a much broader scope than network-based epidemiology as each of the foregoing areas presents its own distinct set of idiosyncrasies. There is a long-standing interest in information cascades in sociology and related fields; however, in the interest of expediency we will postpone until the next chapter consideration of the historical context, and begin here with the paper of 2002 by Duncan Watts [144] which has inspired much of the current wave of research on the subject, including our own work.

2.3.2.1 *Watts's model*

In [144] Watts provided a simple yet rich framework for investigating cascade dynamics on complex networks. In principle his approach (called Watts's model) lends itself to a number of different applications; however, it has found most success as a model of how information, or, as the case may be, misinformation, propagates through society. Utilizing once again the graphical abstraction, let us think of a random graph as representing some social group: vertices are people and edges are bonds of acquaintanceship. As Watts conceives it, the decision of a person to partake in the cascade,

of some fashion or opinion for example, depends only on the states of his or her nearest neighbours (adjacent vertices). Accordingly, each vertex i is assigned a unique threshold of resilience $r_i \in \mathbb{R}$ drawn from some probability distribution $q(r)$, which may be interpreted as representing their independence of mind; i.e., their tolerance against herd-like behaviour. The state of each vertex as a function of time is a binary variable $v_i(t) \in \{0, 1\}$, where $v_i(t) = 1$ means i participates and $v_i(t) = 0$ means he does not. Hence, the model may be interpreted as a particular instance drawn from the more general class of models of interaction dynamics known as *binary decisions with externalities* [128]. If the fraction of a person's nearest neighbours that are actively participating in the cascade is lower than his threshold he will remain independent, refusing to participate ($v_i(t) = 0$); however, if this fraction exceeds his threshold he too will participate ($v_i(t) = 1$). Given a network topology representative of some arbitrary subset of the population, we can simulate the propagation of a cascade through this population using its adjacency matrix as follows:

i) Assign each vertex a unique threshold r_i drawn uniformly at random from $q(r)$.

ii) Starting with all vertices inactive ($v_i(0) = 0$, $\forall\, i$), initiate the cascade dynamics by manually activating a small number of randomly selected *seed* vertices.

iii) Update the state of each vertex, $v_i(t)$, according to the following decision rule:

$$v_i(t) = \begin{cases} 1, & \text{if} \quad \frac{1}{k_i} \sum_j a_{ij} v_j(t) > r_i, \\ \text{unchanged} & \text{otherwise.} \end{cases} \tag{2.13}$$

iv) Repeat step (*iii*) until until no further changes of state are possible.

Note, once a vertex has been assigned the state $v_i(t) = 1$ (active) it cannot return to state $v_i(t) = 0$ (inactive). This crucial feature is referred to as the *permanently active property* (PAP). It guarantees that the dynamics described by steps (*i*)-(*iv*) will achieve a state of completion in which a final, steady-state density of vertices in the graph are active. Evidently, this density will correspond to the cascade size on that particular run of the model: $\frac{1}{n} \sum_{i=1}^{n} v_i(t)$. By averaging this value over many such runs we can compute an expected cascade size, which we shall call ρ. An implementation of this algorithm in MATLAB® code is given in Appendix C.

Applying this methodology enabled Watts to present a number of important numerical results for the size and frequency of cascades in networks

with various degree and threshold distributions. His main analytical result, derived by the generating function formalism of [114], states that the necessary condition for a seed consisting of a single vertex to cause a global cascade is

$$G_0''(1) > z, \tag{2.14}$$

where $G_0''(1)$ is the second moment of the generating function (see Section 2.1) for the degree distribution of *vulnerable* vertices, and z is the mean degree of all vertices in the network. Vulnerable vertices are those which require only one of their neighbours to be active in order to become activated themselves. Thus they mimic, in a very rough sense, the behaviour of *early adopters* [125].[13] When Ineq. (2.14) holds, the average size of connected components of vulnerable vertices diverges, in which case a small initial perturbation (the seed) may trigger a global cascade.[14]

In order to justify the use of generating functions for this result it is necessary to assume that the local edge topology around any randomly chosen vertex is *tree-like*; that is, a branching structure with no clustering. For a random graph constructed using the configuration model, the likelihood that this assumption is valid improves as the size of the graph, n, increases, and has been shown to be valid almost surely as $n \to \infty$ (see Section 2.2.2). Hence, this assumption will amount to quite an accurate approximation, provided the graphs we create are very large.[15] On the other hand, as we have seen, observable networks tend to have very high levels of clustering, and, therefore, we would not necessarily expect such a *tree-based* approach to work as well in real-world applications. (We will discuss this point further in Section 3.3.)

Finally, note that Watts's model is purely prescriptive; in the sense that it contains no governing equations. Thus, generally speaking, statistics such as the expected cascade size, ρ, can be found only by numerical simulations.[16] The generating functions used to obtain Ineq. (2.14) are certainly elegant but they by no means constitute a full, analytically tractable treatment. Thus, the model as presented in [144] lacks a robust method of predicting ρ. As we will now show, addressing this problem has been an important motivating factor in our own research on cascade dynamics.

13 This term has been popularized more recently by Malcolm Gladwell [69].

14 We address the issue of single seed activation in greater detail in Appendix A.2 after our examination of the *influentials* hypothesis in Section 3.2.

15 The random graphs that we analyse usually lie somewhere in the range $n \sim 10^4$ to $n \sim 10^6$, but can be much larger in some cases.

16 On this point note that throughout the succeeding chapters whenever we refer to running numerical simulations on a *network* what is implied by this statement is running a script similar to that shown in Appendix C on the *adjacency matrix* representation of the network.

3

MODELLING CASCADES

Inspired by Watts's work, and also by the need to address its limitations, in their 2007 paper [73] Gleeson and Cahalane introduced a comprehensive analytical framework for investigating cascade dynamics on complex networks. Like Watts's model itself their analysis hinges on the assumption that the topology under consideration is locally tree-like. This technique is by no means a novelty in physics. The Bethe lattice, which as we have noted is a type of connected acyclic graph where each vertex is connected to z others, has for many years been used to simplify problems related to the Ising model, thereby allowing exact solutions to be found.

The Ising model is a mathematical model in statistical mechanics named after the German physicist Ernst Ising (1900-1998). It is used to derive statistics about the global behaviours of large collections of interacting particles based on local information. Ising's original motivation was the phenomenon of ferromagnetism, for which he offered an explanation in terms of the statistical behaviour of iron atoms. As a consequence, terms such as spin, ferromagnetic and anti-ferromagnetic are still widely used to denote certain variables of the model; however, it is not limited to this conceptualization and has been put to use in various different settings, from the study of gases to the neural network of the brain.

In [41] Dhar *et al.* investigated the single-spin flip dynamics of the *random field Ising model* (RFIM) on a Bethe lattice at zero temperature. This problem is closely related to Watts's model. The single-spin flip condition is analogous to the permanently active property; we can think of the random field as the threshold distribution $q(r)$; the Bethe lattice approximation is equivalent to the locally tree-like assumption; and *at zero temperature* may be interpreted as meaning that there are no exogenous forces driving the cascade dynamics. In Watts's model this last condition amounts to saying that there are no global factors affecting propagation; such as, for example, the wide scale media coverage of some product or opinion. Only local interpersonal influence can determine a person's state. (We will examine this aspect further in Section 3.2.) Interestingly, Dhar *et al*. showed in [41] that it is possible to provide an analytically tractable treatment of this version of the RFIM. This suggested that perhaps something similar could be done for Watts's model. In [73] Gleeson and Cahalane did just that when, by

exploiting the similarities between these two models, they derived their analytical model of cascade dynamics.

3.1 A TREE-BASED ANALYTICAL APPROACH

The first step of Glesson and Cahalane's derivation was for them to approximate the topology of an infinite random graph of arbitrary degree distribution, p_k, by a non-clustered tree-like structure with a randomly chosen root vertex (see Fig. 3.1, left panel). They then defined the variable $q_n(k)$ to be the probability that a vertex of degree k at level n of the tree is active, conditional on its *parent* in the tree, the vertex on the next highest level with which it shares a link, being inactive (see Fig. 3.1, right panel). From this basis they were able to write the following iterative equation for the conditional probability that a vertex of degree k is active on each level of the tree:

$$q_{n+1}(k) = \rho_0(k) + (1 - \rho_0(k))G(k, q_n), \tag{3.1}$$

where

$$G(k, q_n) = \sum_{m=0}^{k-1} \binom{k-1}{m} q_n^m (1 - q_n)^{k-1-m} F(m, k). \tag{3.2}$$

Let us take a moment to parse these expressions. The fraction of vertices of degree k that are initially active, by being chosen among our random set of seed vertices, is $\rho_0(k)$. (This is necessarily equal to $q_0(k)$, the probability that a vertex of degree k is initially active.) The probability that a vertex of degree k that is not in the seed will subsequently become activated is given by the function $G(k, q_n)$. This consists of two parts: the binomial probability that at least m of this vertex's $k-1$ *children* on the next lowest level are active, and the *neighbourhood influence response function* $F(m, k)$, which is analogous to the decision rule of Eq. (2.13). If the threshold distribution $q(r)$ is a Dirac delta function, $q(r) = \delta(r - R)$, whereby each vertex has the same threshold R, then $F(m, k)$ may be expressed as

$$F(m, k) = \begin{cases} 1 & \text{if} \quad m > Rk, \\ 0 & \text{if} \quad m \leqslant Rk. \end{cases} \tag{3.3}$$

Thus, combining these pieces of information, Eq. (3.1) tells us that the conditional probability of a vertex of degree k on the next level up (generically called $n+1$) being active is equal to the probability that it was initially active plus the probability that it was not initially active multiplied by the probability that it subsequently became activated by copying the majority behaviour of the neighbours directly below it on the current level (n).

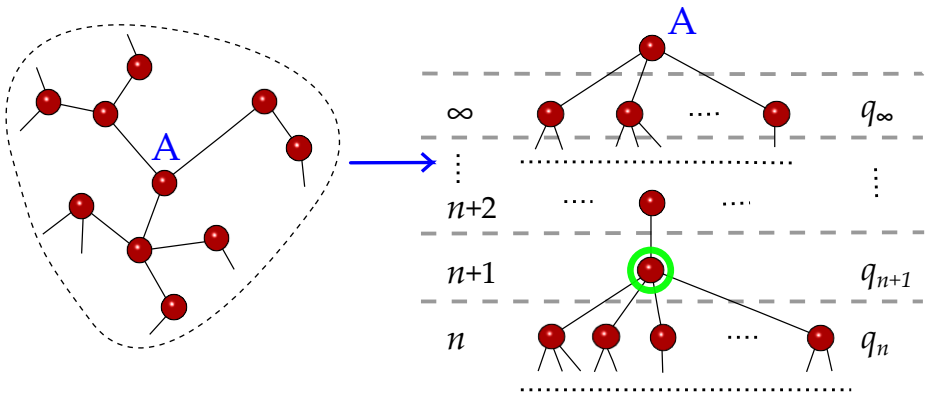


Figure 3.1: Left panel: assume the graph topology is locally tree-like, and let a randomly chosen vertex, A, be the root. Right panel: consider level by level propagation towards A.

Continuing in this vein, to find the degree-independent conditional probability that a vertex picked at random from level $n+1$ is active, q_{n+1}, one takes the average over all values of k as follows:

$$q_{n+1} = \sum_{k=0}^{\infty} \frac{k}{z} p_k \left[\rho_0(k) + (1-\rho_0(k)) G(k, q_n)\right]. \tag{3.4}$$

Note, $(k/z)p_k$ (where z is the mean degree) is the probability of reaching a child of degree p_k by travelling along a randomly chosen edge from its parent. Since every vertex bar the root has a parent this is the correct term to use for averaging on all non-terminal levels (see [110]).

Iterating Eq. (3.4) to the steady state gives q_∞: the probability that a vertex at the penultimate level of the tree, directly beneath the root, is active. Once q_∞ is found it can be used to calculate the the probability of activation of the root itself. This probability corresponds exactly to the steady state density of active vertices[1]; i.e., the expected cascade size, and is given by

$$\rho = \sum_{k=0}^{\infty} p_k \left[\rho_0(k) + (1-\rho_0(k)) H(k, q_\infty)\right], \tag{3.5}$$

where

$$H(k, q_\infty) = \sum_{m=0}^{k} \binom{k}{m} q_\infty^m (1-q_\infty)^{k-m} F(m, k). \tag{3.6}$$

Comparing these expressions to Eqs. (3.4) and (3.2) we see that the main difference here is that the root has no parent, and thus with probability p_k it has k children.

1 To understand why this is so consider an infinite graph in the steady state with a certain fraction of its vertices permanently active. If we pick a vertex at random from this graph the probability that we will find it active is equal to the relative size of the active fraction.

By letting the seed be chosen uniformly at random over all degree classes ($\rho_0(k) = \rho_0$, independent of k), in [73] Gleeson and Cahalane presented Eqs. (3.4) and (3.5) in the following simplified forms:

$$q_{n+1} = \rho_0 + (1 - \rho_0) \sum_{k=0}^{\infty} \frac{k}{z} p_k G(k, q_n), \tag{3.7}$$

$$\rho = \rho_0 + (1 - \rho_0) \sum_{k=0}^{\infty} p_k H(k, q_\infty). \tag{3.8}$$

Finally, analysing these two equations, they derived their own first-order cascade condition, Ineq. (3.9), which as $\rho_0 \to 0$ reduces to Watts's original condition Ineq. (2.14), provided the probability of automatic activation of a vertex $F(0, k)$ is 0:

$$\sum_{k=1}^{\infty} \frac{k(k-1)}{z} p_k \left[F(1, k) - F(0, k)\right] > \frac{1}{1 - \rho_0}. \tag{3.9}$$

The similarities of this tree-based approach to the zero-temperature RFIM on a Bethe lattice are more than merely conceptual. The framework of [73] may, in fact, be viewed as a generalization of the latter model since it reduces to it when we have a random regular graph, and no manually activated seed vertices ($\rho_0 = 0$). However, as we shall see momentarily, this last condition is feasible only for certain types of threshold distribution.

More broadly, as was demonstrated by Gleeson in [70], the results presented above are generalisable to a wider range of dynamics on random networks than those described by Watts's model. Different processes can be modelled by choosing the appropriate form for the response function $F(m, k)$. Examples include site and bond percolation, and *k*-core decomposition [47, 77]. We do not discuss this work here since we shall deal with response functions later on in Chapter 5, when we introduce our extension of the basic theory to highly clustered graphs.

3.1.1 *Theory Versus Simulations*

In this subsection we present some of the quantitative results obtained by applying the tree-based theory to Watts's model. The purpose of this is to give a flavour of how we typically go about verifying our analytical expressions, and to show how well Eq. (3.8) matches the output of numerical simulations, at least on non-clustered random graphs. In Figs. 3.2 and 3.3 we plot the expected cascade size ρ against the mean degree z on Poisson random graphs (written PRGs for short) of 10^5 vertices (see captions for

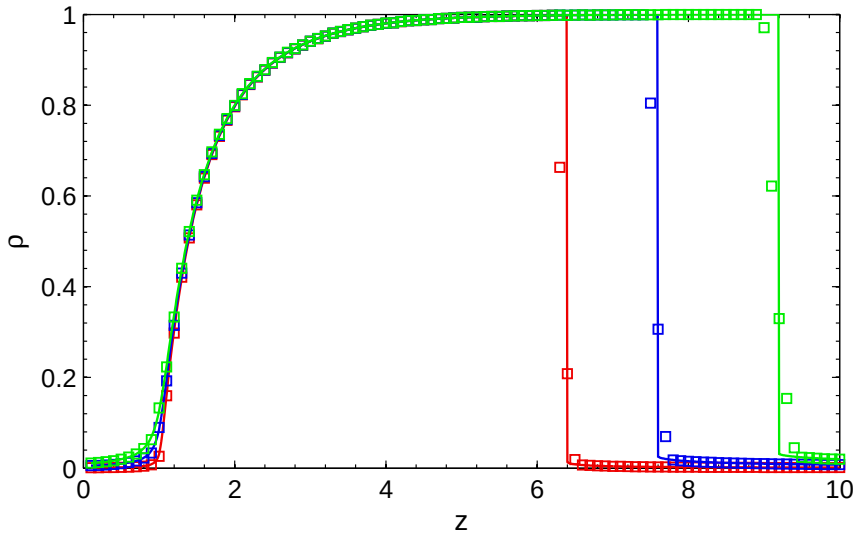


Figure 3.2: Cascade dynamics of Watts's model on PRGs with $n = 10^5$ and uniform thresholds, $R = 0.18$. Numerical simulations (squares) averaged over 100 realisations and tree-based theory (lines). Final active density ρ vs. mean degree z. Colour indicates seed fraction: $\rho_0 = 10^{-3}$ red; $\rho_0 = 5 \times 10^{-3}$ blue; and $\rho_0 = 10^{-2}$ green.

details). These figures are, respectively, reproductions of Figs. 1(b) and 2(b) of [73], and were created using my own code.

In Fig. 3.2 the different colours correspond to different seed fractions ρ_0 (see caption). The threshold distribution is uniform with $q(r) = \delta(r - 0.18)$; meaning every vertex requires a fraction $R = 0.18$ of its neighbours to be active before it will join in the cascade. The match between theory and numerics is clearly excellent. Note also the somewhat curious sequence of *tipping points* where ρ drops discontinuously, from close to 1 (global cascade) to almost 0 (no cascade). This feature is readily explained, however, by considering the behaviour of the response function of Eq. (3.3). Because it depends on the *relative* number, m/k, of active neighbours (as opposed to, say, the absolute number, m), as the average degree, k, of vertices increases we reach a point where it becomes extremely difficult for the cascade to attract new adherents. Thus, here there is a window of z values in which global cascades may occur; but, the same is not necessarily true of other choices of threshold distribution. Gleeson and Cahalane [73] have given bounds for this window which depend on the interplay of R and z (see Figs. 1(a) and 2(a) of [73]), and we can calculate quite accurately where these tipping points will occur by applying the second-order cascade condition expressed by Ineq. (6) of [73].[2]

2 Obviously, this second- order condition is more accurate than the first-order one, Ineq. (3.9), shown above; however, we choose not to reproduce it here as its derivation concerns technical aspects of [73] which will not be pertinent to our discussion as we continue.

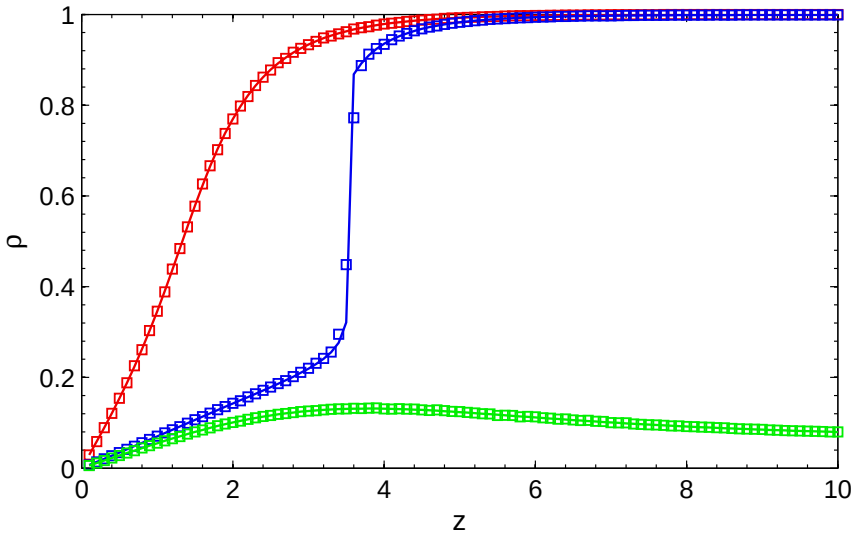


Figure 3.3: Cascade dynamics of Watts's model on PRGs with $n = 10^5$ and Gaussian thresholds, mean R and standard deviation $\sigma = 0.2$. Numerical simulations (squares) averaged over 100 realisations and tree-based theory (lines). Final active density ρ vs. mean degree z. Seed fraction $\rho_0 = 0$. Colour indicates mean threshold: $R = 0.2$ red; $R = 0.362$ blue; and $R = 0.38$ green.

In Fig. 3.3 the threshold distribution is Gaussian, $q(r) = N(R, 0.04)$. In this case, the different colours correspond to different choices of R (see caption). Once again, the agreement between theory and numerics is excellent; however, here we observe a cascade dynamics strikingly dissimilar to that in Fig. 3.2. Some R values result in a discontinuous transition between the global and localized cascade regimes while others do not. Furthermore, it appears that as R decreases, the range of z values for which global cascades can occur broadens steadily, and that the upper bound, or tipping point, is eliminated entirely. This, however, is an artefact of our particular parameter settings. Generally, one may also find well bounded windows for global cascades on networks with Gaussian thresholds, though they are perhaps less prevalent than when uniform thresholds are applied. A full description, including a bifurcation analysis, of the idiosyncrasies engendered by either choice of threshold distribution has been given in [73], and we leave it to the reader to examine this earlier work at his own discretion. In terms of our discussion, what interests us most about Fig. 3.3, besides the accuracy of the theory, is the fact that here we were able to instigate cascades without a manually activated seed; i.e., $\rho_0 = 0$. This was possible because the bell shape of the Gaussian distribution at low R and $\sigma = 0.2$ meant that some vertices were assigned negative thresholds, which in Watts's model translates as automatic activation. Thus, there was no need to assign a nonzero value to ρ_0. Evidently, this would not have worked with uniform thresholds, distributed according to a Dirac delta function.

What have we learned from Figs. 3.2 and 3.3? Well, clearly Gleeson and Cahalane's original tree-based analytical approach is very accurate on non-clustered Poisson random graphs, and this accuracy is robust to the choice of $q(r)$. Other than that, however, these figures raise more questions than they answer. Bearing in mind the definitions and concepts of Chapter 2, there are quite a number of important problems which we have yet to address, and to which the theory, as presented here, may or may not be directly applicable. This sets the scene for the various extensions and modifications which we will consider from now until the end of this thesis.

The next section concerns the idea of targeted activation of seed vertices (see Section 2.3.1), which we apply to Watts's model in order to investigate the so-called *influentials* hypothesis of information dynamics on social networks. This requires a straightforward but, as we shall demonstrate, powerful extension of the basic theory.

3.2 THE INFLUENTIALS HYPOTHESIS

Watts's model has been widely accepted as a reasonable, if somewhat simplistic, description of how information propagates through society. There was a time, however, when it would not have been as willingly accepted as such. Historically, the two-step flow model [92, 97] has been the most popular, and the most successful, theory concerning this topic. Its advocates claimed that it provided an accurate account of the roles played by interpersonal influence and media exposure in the formation of public opinion. According to this model the flow of information in society occurs between three distinct categories of people. At the highest level of influence we have those who work in the mass media; at the lowest level we have the common herd; and, intermediate between these two we have a small group of *opinion leaders* or *influentials*, who are the arbiters of all things trendy.

Recent progress in the theory of networks, however, seriously challenges the validity of this view. In particular, analyses of the development of the World Wide Web over the past ten to fifteen years have made this outlook feel rather naive and dated. It has become blatantly apparent (at least to the generation who grew up in this era) that public opinion formation is much more complex than the hierarchical structure of the two-step flow model. Local, interpersonal influence is, now at least, a much more significant factor than originally assumed — for every boisterous opinion we can find myriad others to contrast against it, at the click of a mouse.

Accordingly, in the picture now offered by Watts and others, society looks more like a random graph, where vertices are people, links are social bonds, and the degree of each vertex corresponds to that person's

influence. There is no *a priori* hierarchy determining the direction in which influence is exerted. Significantly, however, opinion is still strongly divided over the importance of influential individuals. Some still maintain that the occurrence of epidemic-like phenomena in society depends crucially on these so-called trendsetters [69]. Watts himself has been quite outspoken in his disagreement with this view [135].

In [146] Watts and Dodds attempted to remedy the alleged misconception of the influentials hypothesis. Their central argument was that, "... large-scale changes in public opinion are not driven by highly influential people who influence everyone else but rather by easily influenced people influencing other easily influenced people" [146]. This claim was backed up by the results of numerical simulations in which the mean size of cascades initiated exclusively by influential individuals was compared to the size of those initiated by average individuals on random graphs of various degree distributions. An influential was defined here as any vertex with a degree greater than that of 90% of the population. An average vertex was one chosen at random from the entire distribution of degrees.[3] Both average and influential vertices were found to have similar effects on the spread of information in a Poisson random graph. Even in the case of a highly right-skewed power-law degree distribution influentials were found to be of less importance than had previously been assumed [146].

However, similar to Watts's earlier work [144], the numerical results presented [146] were not given a corresponding analytical description. We will now show how the tree-based theory of the previous section can be used to provide such a match. Thus, we develop an analytically tractable method for quantifying the effects of influentials that may be usefully compared to the simulations of [146].

3.2.1 *Extension of Theory*

First, generalising the interpretation given in [146], we define an influential to be any vertex of our graph with a degree located in the top $100\tau\%$ of the degree distribution, where $\tau \in (0, 1]$. Ergo, at least $100(1-\tau)\%$ of all the other vertices are of a lesser degree. Next, by defining

$$k^\star = \min\{k : F_k \geqslant (1-\tau)\}, \tag{3.10}$$

where F_k is the *cumulative distribution function* (CDF) of degrees k, we find

3 The average degree in an infinitely large seed of vertices chosen at random in this way should adhere to the mean degree z, hence the name. We too use the epithet *average* in this sense. This is not to be confused with a seed in which every vertex has degree z.

that τ can be expressed as

$$\tau = 1 - F_{k^\star} + \alpha p_{k^\star}, \tag{3.11}$$

where $\alpha \in [0, 1)$. The idea here is essentially the following. Because the degree distributions of our graphs are necessarily discrete, more often than not we will not be able to find a degree $k^\star$ that cuts off exactly $100\tau\%$ of the vertices. Hence, we find the value of $k^\star$ that gets us as close as possible and then add on an extra piece of probability $\alpha p_{k^\star}$ such that exactly $100\tau\%$ is singled out. Solving for α, we have,

$$\alpha = \frac{\tau + F_{k^\star} - 1}{p_{k^\star}}. \tag{3.12}$$

Using Eqs. (3.11) and (3.12) we determine the effect of initially activating only influentials on the expression for degree dependent seed fraction, $\rho_0(k)$. In summary, $\rho_0(k)$ can now be written as the following function

$$\rho_0(k) = \begin{cases} 0, & \text{if} \quad k < k^\star, \\ \alpha\rho_0/\tau, & \text{if} \quad k = k^\star, \\ \rho_0/\tau, & \text{if} \quad k > k^\star. \end{cases} \tag{3.13}$$

As a quick verification of this we can calculate ρ_0, the mean value of the seed fraction over all degrees k, to show that

$$\begin{aligned} &\sum_{k=0}^{\infty} p_k \rho_0(k), \\ &= \frac{\rho_0}{\tau}\Big(\alpha p_{k^\star} + \sum_{k=k^\star+1}^{\infty} p_k\Big), \\ &= \rho_0. \end{aligned}$$

From Eq. (3.13) the starting probability, q_0, of our tree-based iteration is now given by

$$\begin{aligned} q_0 &= \sum_{k=0}^{\infty} \frac{k}{z} p_k q_0(k), \\ &= \sum_{k=0}^{\infty} \frac{k}{z} p_k \rho_0(k), \\ &= \frac{\alpha\rho_0}{\tau}\frac{k^\star}{z} p_{k^\star} + \frac{\rho_0}{\tau}\sum_{k=k^\star+1}^{\infty} \frac{k}{z} p_k, \\ &= \frac{\hat{z}}{z}\rho_0, \end{aligned} \tag{3.14}$$

where

$$\hat{z} = \alpha k^\star \frac{p_{k^\star}}{\tau} + \sum_{k=k^\star+1}^{\infty} k \frac{p_k}{\tau}, \tag{3.15}$$

is the mean degree of an influential vertex.

Having established these new initial conditions we can now proceed to describe the resultingly modified cascade dynamics. Substituting Eq. (3.13) into Eq. (3.4) we have

$$\begin{aligned} q_{n+1} = &\sum_{k=0}^{k^\star-1} \frac{k}{z} p_k \left[0 + (1-0)G(q_n, k)\right] \\ &+ \frac{k^\star}{z} p_{k^\star} \left[\alpha\rho_0/\tau + (1-\alpha\rho_0/\tau)G(q_n, k^\star)\right] \\ &+ \sum_{k=k^\star+1}^{\infty} \frac{k}{z} p_k \left[\rho_0/\tau + (1-\rho_0/\tau)G(q_n, k)\right]. \end{aligned} \tag{3.16}$$

Similarly, substituting Eq. (3.13) into Eq. (3.5) we find that

$$\begin{aligned} \rho = &\sum_{k=0}^{k^\star-1} p_k \left[0 + (1-0)H(q_\infty, k)\right] \\ &+ p_{k^\star} \left[\alpha\rho_0/\tau + (1-\alpha\rho_0/\tau)H(q_\infty, k^\star)\right] \\ &+ \sum_{k=k^\star+1}^{\infty} p_k \left[\rho_0/\tau + (1-\rho_0/\tau)H(q_\infty, k)\right]. \end{aligned} \tag{3.17}$$

Equations (3.16) and (3.17) simplify to

$$\begin{aligned} q_{n+1} = q_0 &+ \sum_{k=0}^{\infty} \frac{k}{z} p_k G(q_n, k) \\ &- \frac{\rho_0}{\tau} \left[\alpha \frac{k^\star}{z} p_{k^\star} G(q_n, k^\star) + \sum_{k=k^\star+1}^{\infty} \frac{k}{z} p_k G(q_n, k)\right], \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} \rho = \rho_0 &+ \sum_{k=0}^{\infty} p_k H(q_\infty, k) \\ &- \frac{\rho_0}{\tau} \left[\alpha\, p_{k^\star} H(q_\infty, k^\star) + \sum_{k=k^\star+1}^{\infty} p_k H(q_\infty, k)\right], \end{aligned} \tag{3.19}$$

respectively.

Taken together, these last two equations are our main result of this section. We refer to them as an extension of the tree-based theory since by varying τ we can investigate both the dynamics that take place when any vertex may be initially active ($\tau = 1$) and those that take place when only influentials are initially active ($\tau < 1$). Note, if we set the parameter

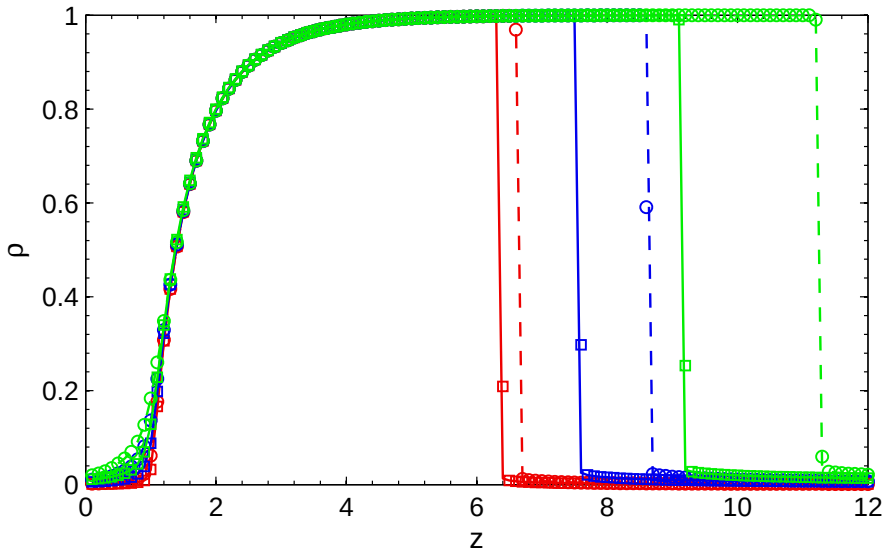


Figure 3.4: Cascade dynamics of Watts's model on PRGs with $n = 10^6$ and uniform thresholds, $R = 0.18$. Numerical simulations (symbols) averaged over 100 realisations and extended tree-based theory (lines). Final active density ρ vs. mean degree z. Solid lines: $\tau = 1$, average seed. Dashed lines: $\tau = 0.1$, influential seed. Colour indicates seed fraction: $\rho_0 = 10^{-3}$ red; $\rho_0 = 5 \times 10^{-3}$ blue; $\rho_0 = 10^{-2}$ green.

$\tau = 1$, Eqs. (3.18) and (3.19) reduce to the original equations of [73] (Eqs. (3.7) and (3.8)). In a similar manner to Section 3.1.1, by applying these extended governing equations we produced the agreement between theory and numerical simulations of Watts's model shown in Figs. 3.4 and 3.5.

The first of these, Fig. 3.4, shows cascade dynamics for random and targeted seeds on Poisson random graphs of 10^6 vertices (see caption for details). First and foremost, this figure illustrates how well our extended theory works in this important test case. It also illustrates, more clearly perhaps than Fig. 3.2, the discontinuous nature of the transition that takes place at the tipping points. This was one of the advantages of stepping up to 10^6 vertices. As one might have expected, the effect of targeting influentials is to extend the range of z values for which global cascades can occur. Significantly, however, we find that when global cascades do occur for both types of seed, the size of those instigated by a seed of influentials in the top 10% of the degree distribution is never considerably greater than the size of those instigated by randomly chosen average degree vertices.

For our second test, the results of which are shown in Fig. 3.5, we have investigated the role of influentials in cascade dynamics on scale-free networks (SFNs) at $n = 10^6$. These networks (or more properly random

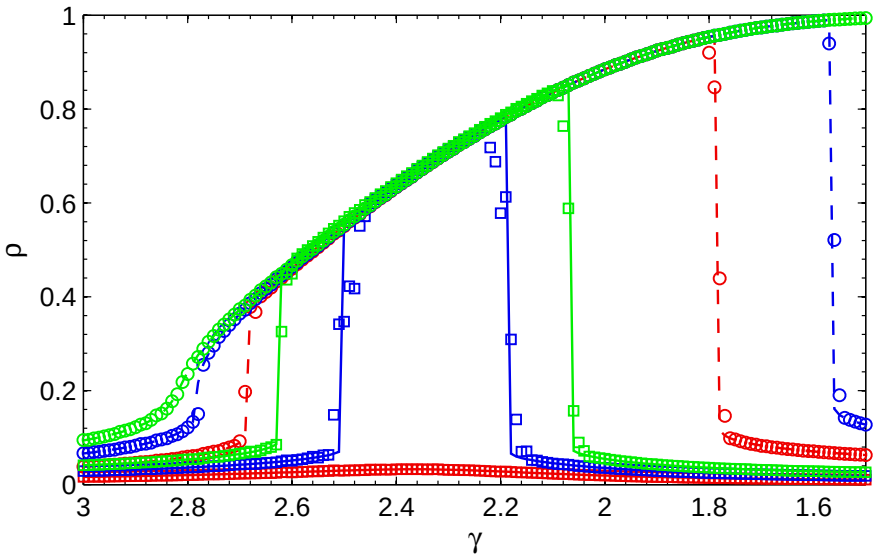


Figure 3.5: Cascade dynamics of Watts's model on SFNs with $n = 10^6$ and uniform thresholds, $R = 0.16$. Numerical simulations (symbols) averaged over 100 realisations and extended tree-based theory (lines). Final active density ρ vs. slope γ. Solid lines: $\tau = 1$, average seed. Dashed lines: $\tau = 0.1$, influential seed. Colour indicates seed fraction: $\rho_0 = 4 \times 10^{-3}$ red; $\rho_0 = 7 \times 10^{-3}$ blue; $\rho_0 = 10^{-2}$ green.

graphs) were generated from a truncated power-law degree distribution of the form

$$p_k = \frac{k^{-\gamma}}{C}, \tag{3.20}$$

where $C = \sum_{k=1}^{k_{max}} p_k$. The truncation is applied such that $p_{k_{max}} \geqslant 10^{-5}$, with k_{max} the maximum degree. That is, given a slope γ, we generate probabilities according to $k^{-\gamma}$ until $p_{k_{max}}$ drops below 10^{-5}, we then discard this value, letting the previous number in the distribution be our $p_{k_{max}}$. The entire distribution is then normalised by dividing each p_k value by the total sum of probabilities, C.

In Fig. 3.5 we plot the mean cascade size ρ against the slope of the degree distribution γ (see caption for details). Our γ values range from $\gamma = 1.5$ to $\gamma = 3$. We have chosen to arrange these values in decreasing order because in power-law functions like Eq. (3.20) the slope (exponent) is inversely proportional to the mean; therefore, when read from left to right our mean degrees increase from $z = 1.3499$ at $\gamma = 3$ to $z = 26.2861$ at $\gamma = 1.5$, making it easier to compare this figure to those already shown. As discussed previously in Chapter 2, this range of γ appears to be the most relevant to real-world applications.

While Fig. 3.5 is qualitatively quite different from Fig. 3.4, many of the same inferences drawn from that figure may also be drawn from this one.

The only significant change is in the effect influentials have on the cascade window. We know from Fig. 3.4 that using influentials exclusively as our seed tends to broaden the range of z values for which global cascades can occur. The same applies here, only in a much more pronounced way. However, this was to be expected since for scale-free networks the heavily right-skewed tails of their power-law degree distributions mean that the difference in magnitude between the highest degree vertices and those of average degree is typically far greater than in Poisson random graphs. This last statement posits on behalf of the reader the following interpretation of the role played by influentials in driving cascades.

3.2.2 *Approximation*

Having paid close attention to the analysis and results described thus far one cannot have failed to have noticed that targeting high degree vertices for our seed seems to affect the cascade dynamics of Watts's model in a manner essentially similar to what one might expect to find after having increased the number of vertices in a random seeding. For example, looking again at Fig. 3.4 we see that for each seed fraction (colour) the line representing a seed of influentials (dashed) is more or less the same as that representing a seed of average vertices (solid) only with its tipping point shifted to the right. Furthermore looking at each color, this shifting phenomenon appears to be the only significant difference between different seed sizes. These observations suggest that it may be possible to approximate the behaviour produced by targeting influentials merely by increasing the relative size of a randomly chosen seed. If this intuition is correct it would corroborate Watts and Dodds's [146] claim that cascades are driven by the great number of easily influenced individuals in a population rather than by the comparatively small number of influencers, by showing that the effect that either group has on cascades can be approximately replicated by the other. It would also simplify our theoretical analysis as instead of using the awkward looking Eqs. (3.18) and (3.19) we would simply use Gleeson and Cahalane's original Eqs. (3.7) and (3.8) with a larger ρ_0.

Of course, the increase applied to ρ_0 cannot be entirely arbitrary; it must have some sort of theoretical underpinning. Taking what we have learned from our derivation in the previous subsection we know from Eq. (3.14) that targeting influentials changes the probability that a vertex is initially active from $q_0 = \rho_0$ to $q_0 = (\hat{z}/z)\rho_0$. That is, it increases this probability by a multiplicative factor equal to the ratio between $\hat{z}$, the mean degree of influential vertices, and z, the mean degree of all vertices. For non-targeted seed activation we know from the original theory of [73] that $q_0 = \rho_0$. Thus,

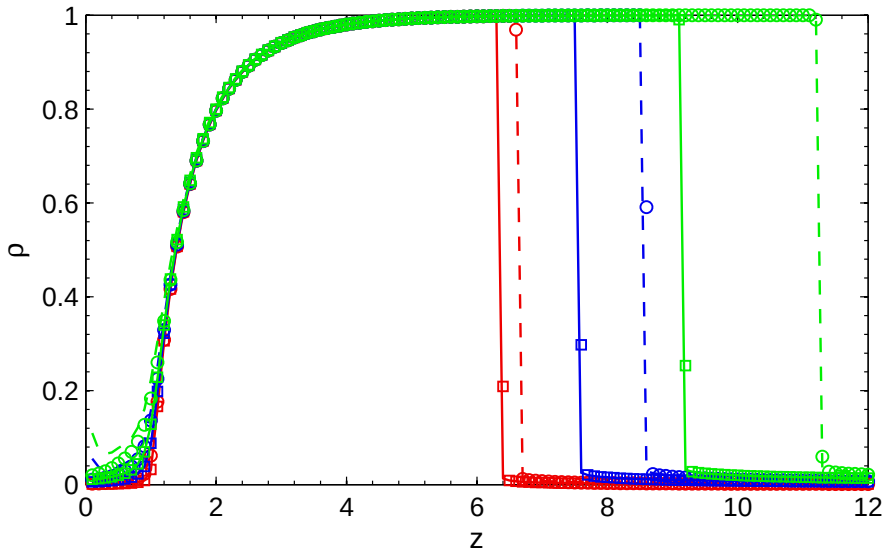


Figure 3.6: Cascade dynamics of Watts's model on PRGs with $n = 10^6$ and uniform thresholds, $R = 0.18$. Numerical simulations (symbols) averaged over 100 realisations and original tree-based theory (lines). Final active density ρ vs. mean degree z. Solid lines: average seed. Dashed lines: approximation of influential seed. Colour indicates seed fraction: $\rho_0 = 10^{-3}$ red; $\rho_0 = 5 \times 10^{-3}$ blue; $\rho_0 = 10^{-2}$ green.

in our effort to approximate the effect induced by targeting high degree vertices, from within the original framework, we have the following natural choice for our updated seed fraction

$$\rho_0 \to \frac{\hat{z}}{z}\rho_0. \tag{3.21}$$

In other words, instead of picking influentials and applying our extended theory we will now attempt to replicate their effect on cascade dynamics, which is at least in part attributable to the renormalization of the initial probability q_0 defined by Eq. (3.14), by picking a correspondingly renormalized number of average vertices and then applying Eqs. (3.7) and (3.8). The results obtained from this approximation are given by the dashed lines in Figs. 3.6 and 3.7 (see captions).

Regarding Fig. 3.6 we see that the approximation appears to be accurate, particularly at the tipping points; although, it does not work quite so well at low z values. On this point, however, we would remind the reader that what we are primarily interested in here is the qualitative match of the approximation to numerical simulations with $\tau = 0.1$, in order that we might find out whether or not our increased average seed captures well the extent (height and width) of the cascade window produced by influentials. Clearly, our conjecture that it does is supported by this figure. It is also supported by the dashed lines in Fig. 3.7 which again, though

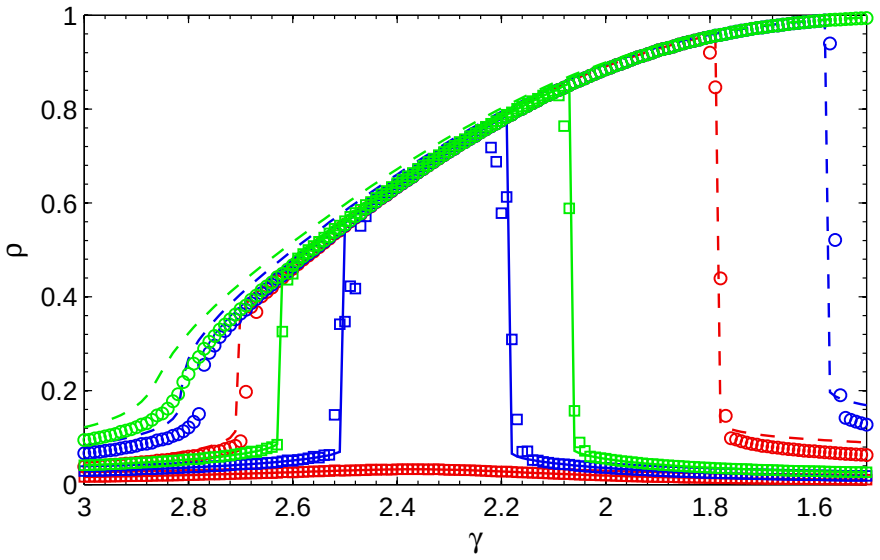


Figure 3.7: Cascade dynamics of Watts's model on SFNs with $n = 10^6$ and uniform thresholds, $R = 0.16$. Numerical simulations (symbols) averaged over 100 realisations and original tree-based theory (lines). Final active density ρ vs. slope γ. Solid lines: average seed. Dashed lines: approximation of influential seed. Colour indicates seed fraction: $\rho_0 = 4 \times 10^{-3}$ red; $\rho_0 = 7 \times 10^{-3}$ blue; $\rho_0 = 10^{-2}$ green.

not quantitatively accurate at lower z (higher γ) values, still show that the extent of cascades from influentials can be comparatively matched by a sufficiently increased number of random seed vertices.[4] Thus, while this approximation is by no means an adequate replacement for our full theory, it does provide further theoretical justification for the view of information dynamics and opinion formation espoused in [146].

Looking beyond the theoretical domain, the insight provided by this analysis may have important practical implications. Consider, for example, the business of mass marketing. Many companies spend huge sums of money on advertising their products and services. It is the marketing executive's job to determine the most cost effective strategy to bring these products and services to the attention of as many people as possible. The conventional wisdom on which the influentials hypothesis is founded says that some specific subgroup of the population must be more important than others in helping information to spread. Interestingly, regardless of whether or not a marketing campaign is successful, real-life influentials are never physically identified. If a cascade does occur it is more or less taken for granted that this amorphous group (or some equivalent) must have been activated. Conversely, unsuccessful campaigns are those that,

4 We stress that Eq. (3.21) is not the only way to increase ρ_0, and that we chose it merely because it carries with it an intuitive justification. Further analysis of the renormalization of ρ_0 may reveal a more accurate approximation.

presumably, have failed to activate influentials.[5] However, Figs. 3.6 and 3.7 demonstrate that rather than company executives spending lavishly on elaborate schemes to track down and activate these elusive, super effectual members of the population, it may be better worth their while to focus on directly activating as many "average" members as possible.[6]

This concludes for now our investigation of the influentials hypothesis and our analytical treatment of problems directly related to dynamics on social networks. Needless to say, there are many technical aspects of this study that we have not yet considered. Some of these we leave to the reader to discover by referring to the extensive literature that has been built up in recent years, both through the rigorous framework of network-based analyses [64, 121, 122] and through the more accessible medium of popular science [11, 27, 145]. Other aspects, however, we cannot fail to address ourselves since the theory that we have provided here is so apt to be applied to them. In particular our extended theory can be used to derive analytical expressions for the critical seed fraction for global cascades, and the expected cascade size in the case of single seed activation. We present these results in Appendix A.

3.3 THE EFFECTIVENESS OF THE TREE ANALOGY

The remarkable ease with which the theoretical results of the previous section were derived suggests that the framework of Section 3.1, and appropriate modifications thereof, may serve as powerful tools in tackling numerous questions of practical significance. We have said that Gleeson [70] has extended the basic tree-based approach to address different types of percolation processes; and one may recall the congruity, noted in Section 2.3.1, between the properties of random graphs under percolation and the robustness of technological systems. One may also recall the noted similarities between bond percolation and the SIR contagion model. Combined with our extension into the domain of opinion dynamics, then, it appears that quite a number of the problems of interest to us in the social and technological spheres (those that admit of a mathematical interpretation at least) should be treatable. It would seem that the only question remaining to us is: How much further do we wish take this type of analysis?

5 One might consider it somewhat unsettling that such a multi-billion dollar industry as mass marketing, which has become so invasive in our lives, appears to be run in large part on a logical fallacy equivalent to *the ground outside is wet, therefore it must have rained*.

6 It is important to recognise that this conclusion does not diminish the significance of influential individuals as spreaders of information but rather reinforces it. This is evident from the fact that we require a large group of average vertices to match the outcome produced by a smaller group influentials. This will be further drawn out in Appendix A.2.

Lest we get carried away in our own hubris, however, it is worth reminding ourselves of the very pertinent fact that the approach of [73], and therefore every result subsequently derived from it, is predicated on the accuracy of the locally tree-like approximation of the network topology under consideration. For infinitely large random graphs constructed using the configuration model (like all of those examined thus far) we need have no concern about this approximation affecting our results since these graphs are almost surely non-clustered in the $n \to \infty$ limit. Therefore, there is a real sense in which the matches obtained between theory and numerical simulations in Figs. 3.2 to 3.5 constitute merely an elementary form of verification for the framework put forth. In other words, to further test our approach we should investigate more complicated graphical structures. If we stop to think for a moment what these might be, it immediately becomes evident that there are a multitude of quite straightforward variations of the basic configuration-type graph which we simply have not considered yet; this is before we even approach the task of introducing clustering.

For example, what can we say about cascade dynamics on digraphs; i.e, graphs with directed edges, or for that matter graphs with degree- degree correlations? Well, as it turns out the generalization of the tree-based theory to directed graphs of arbitrary in- and out-degree distributions was given in the master's dissertation of Gleeson's student Alan Dunne [51]. Degree-correlated graphs were dealt with in [70], where the approach of [73] was first extended to account for correlations between the degrees k and k' of the end vertices of a randomly chosen edge, defined by the joint probability distribution $P(k, k')$, and then used to provide analytical results for k-core sizes. For this reason, our exposition is not overly concerned with these two particular aspects of higher order network approximations.[7] As already stated, our main contribution in this respect will be our extension of the theory of [73] to highly clustered graphs.

Short of dismissing these features of network structure entirely, however, we note briefly that there is one particular field of application where an analytical description of cascades on directed graphs has proven to be quite fruitful for us. In very recent, and as yet unpublished, work we (Gleeson, Melnik, and Hackett) have collaborated with financial mathematician Tom Hurd in an investigation of the very topical phenomenon of credit default contagions on banking networks [74]. There are too many specifically finance-oriented components to this study for a detailed discussion of it to be of relevance here; however, the main thrust of the modelling technique adopted in this work can be summarized as follows.

7 In relation to opinion dynamics, the aspects that we have not addressed ourselves, which in any case have already been amply dealt with in the papers [121, 122] referred to above, are precisely cascades on directed and/or degree correlated graphs.

First, we modelled an interbank network as a directed graph where for each institution its incoming edges (pointing towards it) represent the number of debtor banks by which it is owed money and its outgoing edges (pointing away from it) represent the number of creditor banks to which it owes money. The method of construction of these graphs was similar to the approach of [51] and was based on a simple modification of the configuration model. For the distributions of in- and out-degrees, respectively j and k, we used a product of independent Poisson distributions:

$$p_{jk} = \frac{z^j}{j!}e^{-z}\frac{z^k}{k!}e^{-z}. \tag{3.22}$$

We then assigned artificial balance sheets of appropriate assets and liabilities to each vertex in order to replicate in a crude sense those of real financial institutions. Finally, the spreading dynamics were modelled in essentially the same manner as usual by *shocking* seed banks chosen either at random or from a specific (j, k)-class. In the latter case targeting high degree classes, gave us an insight into the infamous *too big to fail* concept, which, we note, is not too dissimilar from the influentials hypothesis. Shocking in this context referred to setting equal to zero the external assets of a bank. The mechanism of contagion was a specialized version of child to parent activation on a directed tree, where an activation symbolized an irrecoverable default on debts. Thus our analytical expressions were suitably modified, leading ultimately to new results for the extent and frequency of default cascades.

This work on banking networks, though it does not fit quite so easily into our generalized framework, still illustrates yet another successful application of the tree analogy for the purpose of modelling cascade dynamics on networks. However, given that the directed graphs used were constructed from little more than a modification of the configuration model, the limitations of these graphs as a testing ground for tree- based approaches are not significantly less than those associated with the standard undirected graphs discussed above.

What we are really driving at in the entire discussion of this section, and in fact in this thesis as whole, is the applicability of an analytical model of cascades, which relies on a locally tree-like assumption, to real complex networks. Everything we have learned so far about real-world structures indicates that we are going to have to account somehow for high levels of clustering, and indeed the next two chapters will be devoted entirely to this problem. Before diving headlong into this task, however, we might ask ourselves first, following the best empiricist tradition: Do we have any hard evidence (beyond received opinion) to believe that tree-based

theories do not work at all on clustered networks? The answer to this question may appear self-evident and we could justifiably disregard it completely; however, we may be better served by taking it seriously. The answers discovered will at least indicate what degree of discrepancy we will be seeking to remedy in succeeding chapters.

In another quite recent paper [99], therefore, we (Melnik, Hackett, Porter, Mucha, and Gleeson) have taken this latter question and investigated in some detail the application of the standard first-order approach of [73] which we called p_k-theory and its extension to degree-correlated graphs given in [70], called $P(k, k')$-theory, both of which are of course tree-based, to modelling dynamics on various real-world networks. The networks considered in this study broached the realms of technology, biology and sociology, and included the power grid of the western United States[8] [147]; the autonomous systems level Internet[9]; a network of the 500 most congested U.S. airports[10] [35]; the protein interaction network of the yeast *S. Cerevisae*[11] [33, 34]; the metabolic[12] [50] and neural[13] [147] networks of the nematode *C. Elegans*; a scientific coauthorship network[14] [106]; and a set of 100 different networks each representing for a single major U.S. university the *friendships* of students of the university on the social networking website Facebook® [137, 138], among others.

The dynamics considered in this study were bond percolation, *k*-core decomposition, SIS disease dynamics, and Watts's model with both uniform and Gaussian distributed thresholds. Figures 1 to 4 and Fig. 7 of [99] illustrate the matches obtained between theory and numerical simulations on a number of the above networks in calculations of a key quantity of interest for each dynamics.[15] For example, the bond percolation results in Fig. 1 of [99] show the relative size of the percolating cluster S as a function of uniform bond occupation probability ϕ_b. For the purpose of illustration, parts (a) and (d) of this figure, corresponding respectively to the University of Oklahoma Facebook network and the western U.S. power grid, are reproduced here in Fig. 3.8(a) and 3.8(b). Before discussing the conclusions drawn from the analysis of this and other figures in [99], let us first clarify the methods by which it was produced.

8 Download data at [http://www-personal.umich.edu/~mejn/netdata/power.zip].
9 Based on CAIDA measurements from 30-Jun-2008. See [http://www.caida.org/data/active/as-relationships/].
10 See [http://sites.google.com/site/cxnets/US_largest500_airportnetwork.txt].
11 See [http://sites.google.com/site/cxnets/DIP.dat].
12 See [http://deim.urv.cat/~aarenas/data/xarxes/celegans_metabolic.zip].
13 See [http://www-personal.umich.edu/~mejn/netdata/celegansneural.zip].
14 Specifically, coauthorships in preprints posted under condensed matter on [http://arxiv.org/] between 1-Jan-1995 and 31-Mar-2005. Download at [http://www-personal.umich.edu/~mejn/netdata/cond-mat-2005.zip].
15 For detailed definitions of these quantities see [99].

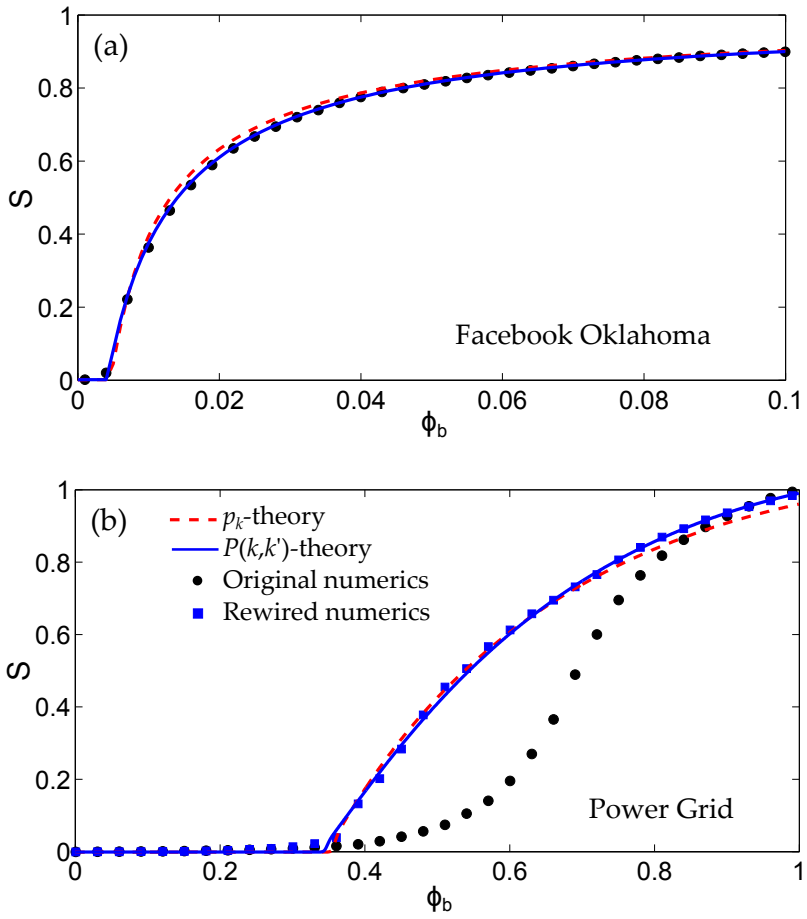


Figure 3.8: Bond percolation on two real-world networks. GCC size S vs. bond occupation probability ϕ_b for (a) the Facebook network of the university of Oklahoma, (b) the western United States power grid. Source [99].

The creation of the theory lines in either window of Fig. 3.8 was quite straightforward. For the p_k-theory (red dashed) we extracted the degree distribution, p_k, from the adjacency matrix representation of the network in question, and then using the equations of Section 3.1 with the appropriately defined response function (see Eq. (6) of [70]) we calculated S ($\equiv \rho$).[16] Similarly, for the $P(k,k')$-theory (blue solid) we extracted the degree-degree correlation matrix $P(k,k')$ of the network and calculated S using the equations of Sec. V of [70].[17] On the other hand, the numerical results shown in Fig. 3.8 have a somewhat more interesting background. First, for both networks we ran bond percolation processes on their measured (downloaded) adjacency matrices. Similar to Watts's model, this means that in each case we fed the adjacency matrix of the network into a MATLAB script that simulated bond percolation and returned the relevant statistics. The specific algorithm employed was that defined by Newman and Ziff in [116]. Our

16 Alternatively, one may use the original analytical approach of Callaway *et al.* [24] to find S.
17 In this case the original analytical approach was determined by Vázquez and Moreno [141].

implementation of this algorithm is given in Appendix C; and the values of S returned by our script are illustrated in both windows of Fig. 3.8 as black circles. Second, for the power grid network we also ran bond percolation on a *rewired* version of its adjacency matrix, where rewiring refers to running this matrix through an algorithm that removes clustering but preserves degree-degree correlations (see Appendix C). The values of S returned by our script in this case are plotted as blue squares.

The first question raised by Fig. 3.8 is obviously the following: Why do the theory lines match the numerical result for the unrewired network so much better in Fig. 3.8(a) than they do in Fig. 3.8(b)? The answer that immediately springs to mind is that the error in each window must be commensurate to the level of clustering in the respective networks, and therefore the severity of the discrepancy in Fig. 3.8(b) indicates the presence of extremely high clustering in the power grid network. However, the measured values of the global clustering coefficient C_2 (see Section 2.1) for the Facebook Oklahoma and power grid datasets completely confound this intuition. For Facebook Oklahoma $C_2 = 0.23$, while for the power grid $C_2 = 0.08$. Hence, the error in Fig. 3.8(b) cannot be directly attributed to high clustering. Nor for that matter can the accuracy in Fig. 3.8(a) be attributed to low clustering. Furthermore, the rewired numerics on the power grid show that the error for the original network does not arise from finite size effects. The $P(k, k')$-theory is accurate for the rewired version because the ensemble of fully rewired networks is the same as the ensemble of random networks defined by the $P(k, k')$ matrix (up to finite size effects).

If it is not a result of high clustering nor an artefact of the finite size of the power grid, then to what can we ascribe this lack of accuracy of the theory? To answer this question we constructed an error measure E (see Eq. (1) of [99]) that determines the mean vertical distance between the $P(k, k')$-theory line and the original (unrewired) numerics over the interval $\phi_b = [0, 1]$. By applying an analogous definition for other dynamics we have observed that throughout the entire range of aforementioned real-world networks the value of E bears a much more significant relationship to the mean intervertex distance than it does to clustering. To be precise, scatter plots of E for every network against a variety of possible correlates (see Fig. 9 of [99]) have revealed that C_2 is a very poor predictor of this error; its coefficient of determination being $R^2 = 0.08680$. C_1 is also poor, with $R^2 = 0.20134$. The best match we could find was $(L - L_1)/z$, which gave $R^2 = 0.93581$. Thus we have found that the size of E is correlated with the difference between the mean intervertex distance (or *average geodesic path length*, see Section 2.1) in the original network, L, and that in the rewired version, L_1, divided by the mean degree z.

In [99] we interpreted these results as follows: A tree-based approach, such as $P(k, k')$-theory, may be expected to work quite well on any given complex network provided the network's edges are sufficiently well mixed between its vertices for the empirical value of L to be comparably close to the randomized L_1, with a high z value obviously improving the accuracy of the theory even further. Remarkably, the level of clustering in the network simply does not appear to be a significant factor in determining the effectiveness of the tree analogy. It would appear, therefore, that any attempt to broaden the applicability of the theory of Gleeson and Cahalane [73] to cascade dynamics on real-world networks, by extending their approach to highly clustered graphs, will be futile. However, in the next chapter we will offer some arguments to counter this interpretation.

4

NETWORKS WITH CLUSTERING

In this chapter and the next we will consider the problem of modelling cascade dynamics on networks with clustering. We will begin here by reviewing recent developments in the construction of ensembles of random graphs with tunable clustering coefficients, before then proceeding in Chapter 5 to seek extensions of the tree-based approach of [73] to account for these more complex topologies. First, however, in acknowledgement that the insight provided at the end of the previous chapter — that the accuracy of non-clustered theoretical approaches seems to depend on mean intervertex distance rather than clustering — may appear to the reader to undermine the task we have just set ourselves, we now take a moment to consider carefully the precise meaning of the results of [99].

There are a number of issues that could be raised concerning these results; however, for our purposes it is sufficient to highlight no more than two closely related points. First, the error measure, E, used in [99] was of a very specific form in that it simply measured the discrepancy between theory and numerics in terms of the vertical distance between data points. Mean intervertex distance may well be a better indicator of this discrepancy than clustering, but that does not imply that clustering has no effect whatsoever on dynamics. Perhaps, similar to an influential seed, the effects of clustering are only detectable by observing closely the transition from the localized to global cascade regimes.[1] Recall that in our analysis of Fig. 3.4 we noted that within the window of z values for which both types of seed, average and influential, produced global cascades, the size of those produced by influential seeds was never considerably greater than the size of those produced by average degree seeds (both gave $\rho \approx 1$). Thus in this range of z it appeared that whatever effect influential vertices had, it was not correlated with the height of the cascade window. However, by observing the relative positions of the the tipping points it became clear that influentials significantly affected the height of this window by shifting these points to higher z values and causing global cascades to occur where they had not occurred for average seeds. The possibility that the effects induced by clustering may be of a similar nature, i.e., that they are directly connected to the change in the position of the cascade transition, with any change in the relative cascade size occurring only secondarily to

1 We shall have more to say on this point in Sections 4.3 and 5.1.3.

this, was acknowledged in the closing remarks of [99], though different error measures — incorporating, for example, the critical bond occupation probability — were not considered. Second, the philosophy of [99] leaves something to be desired. Generally, one does not look for a specific set of circumstances under which a low-order approximation may give accurate results for some observed dynamics, thereby narrowing the real world to meet one's theory; but rather one seeks to improve one's approximation by accounting for more of the relevant features of those dynamics, thereby broadening one's theory to meet the real world. This is particularly so when those features have an established mathematical interpretation. Thus, since the analysis of [99] cannot not rule out the relevance of clustering, which as we already know is a very important structural feature of real networks, in any other sense beyond its lack of correlation with E, we are justified in pursuing our extension of the basic approach of [73] to modelling cascade dynamics on highly clustered networks.

Note, however, that this critique is not designed to diminish the significance of [99]. Until this very recent work of ours, it had been more or less tacitly accepted, in part because of the difficulty in creating highly clustered structural models, that clustering must be by far the most significant factor determining the behaviour of cascading processes on real networks. Our work in [99] has shown that while clustering may be important structurally, accounting for it will by no means mark the pinnacle of all higher order analytical descriptions of dynamics.

Bearing this in mind then, let us begin our pursuit by reviewing the historical and current state of affairs as regards structural models.

4.1 A GAP IN THE LITERATURE

A lot of progress has been made in modelling the structure of complex networks over the past decade or so. In Chapter 2 we outlined the four major network growth models that underlie this progress, and cited a number of their important variants. However, we also saw that these modelling efforts are significantly limited in that, while each of them captures some of the characteristics of real-world networks, namely high clustering coefficients, short average geodesic path lengths, and power law scaling in the tails of their degree distributions ($p_k \sim k^{-\gamma}$, usually with $\gamma \in [2, 3]$), no one model has been found to capture all of these features simultaneously. In particular, an analytically tractable method for the creation of random graphs with realistically high (and preferably tunable) levels of clustering has been conspicuously absent from the networks literature.

To recap briefly, we saw that Erdös and Rényi gave two of the earliest prescriptions for the creation of ensembles of random graphs, $G_{n,M}$ [56] and $G_{n,p}$ [57], the latter of which led directly to the creation of the Poisson random graph model; the null model for all subsequent models of network growth. While the Poisson random graph does have a short average geodesic path length ($L \sim \log(n)/\log(z)$), its clustering coefficient vanishes as $n \to \infty$ and its degree distribution is necessarily limited to a Poisson distribution with mean z: $p_k \sim (z^k/k!)e^{-z}$.

On the other hand, in the configuration model of Molloy and Reed [104] the desired degree distribution $\widehat{p_k}$ is first given as a parameter; a degree sequence, prescribing the degree of each vertex, is then drawn uniformly at random from this distribution; and finally a random graph is created following a fixed and easily repeatable method (see Section 2.2.2). In theory any well-defined degree distribution can be fed in at the beginning and the match between the desired and actual degree distribution should improve as $n \to \infty$. However, just like Erdös and Rényi's model, the configuration model suffers from the significant drawback of producing graphs with vanishing clustering coefficients in the $n \to \infty$ limit.

Watts and Strogatz's small-world network model [147], and variants thereof, may be used to interpolate continuously from a completely ordered lattice where every vertex has the same degree to an unordered structure with a random sequence of degrees. The major drawback of this approach is the form of the degree distributions that it creates. Starting with a delta spike at z when the rewiring probability $p = 0$, as p increases shortcuts are added between vertices and the variance in their degrees increases until at $p = 1$ we reach a form similar to a Poisson degree distribution, only narrower. Significantly, it has been shown [110] that the degree distribution of a graph constructed in this way is never as broad as a Poisson distribution for any $p \in [0, 1]$. This is rather unfortunate since the levels of clustering produced by this model can be quite high.

Finally, the Barabási-Albert model gives us scale-free random graphs with power law degree distributions but again, similar to those created by both the Erdös-Rényi and configuration model, the clustering coefficient in these graphs vanishes as $n \to \infty$.

This lack of a plausible mechanism for the creation of random graphs with high of levels clustering is strikingly at odds with the enormous strides that have been made in our overall understanding of network-based phenomena. In light of this general success there must, we feel, be some simple, though as yet undeveloped, model that will capture all of the relevant features of real-world networks at once, including high clustering, and which will readily lend itself to an analytical treatment. For a long

time, surprisingly little progress had been made in this direction beyond that constituted by the aforementioned models; in the past couple of years, however, this has begun to change.

4.2 TWO NOVEL APPROACHES

4.2.1 *Edge-Triangle Graphs*

In a paper published in July of 2009 Newman [111] introduced an analytically solvable model for the creation of ensembles of random graphs with tunable clustering coefficients.[2] Viewed as a generalisation of the classical configuration model, the distinguishing features of this new method of graph generation can be summarised in the following steps. First, let the variable t_i denote the number of triangles (3-cliques) attached to an arbitrary vertex i, and rename k_i, the number of single edges attached to i, as s_i. Next, assign values to these two variables for all vertices $1 \leqslant i \leqslant n$, thereby creating the sequences s and t. Finally, expand s and t into their respective *stubslists* (see Section 2.2.2), and connect together *stubs* of triangle edges between vertex triples, and stubs of single edges between pairs of vertices. In theory, having done all this one should end up with a realisation of a clustered random graph whose localized topology, resembling Fig. 4.1, consists of non-overlapping triangles and single edge pairs.

The traditional graph descriptors can still be defined using the variables s and t; for example, the degree of each vertex, k, is simply $k = s + 2t$, and the degree distribution, p_k, is now given (after [111]) by

$$p_k = \sum_{s,t=0}^{\infty} p_{s,t}\delta_{k,s+2t}, \tag{4.1}$$

where $p_{s,t}$ is a joint distribution defining the probability that a vertex is attached to s single edges and t triangles, and $\delta_{i,j}$ is the Kronecker delta.

Similar to Newman's earlier papers on graph structure, the analysis given to these new edge-triangle (or simply $p_{s,t}$) graphs in [111] relied heavily on the use of probability generating functions (see Section 2.1). From [111], the generating function for $p_{s,t}$ is

$$g_p(x,y) = \sum_{s,t=0}^{\infty} p_{s,t}x^s y^t, \tag{4.2}$$

2 A model closely analogous to Newman's was independently proposed by Miller [101], and first appeared in print little over a week after the publication of [111].

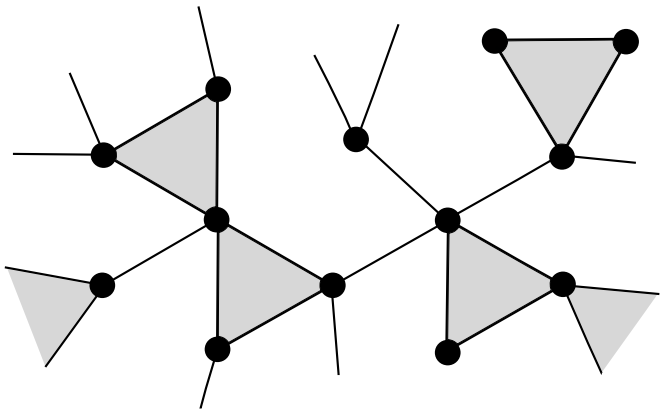


Figure 4.1: In Newman's model one specifies separately the number of single edges and triangles (shaded) attached to each vertex. Reproduction of Fig. 1 of [111].

while the generating function for p_k is

$$f(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{s,t=0}^{\infty} p_{s,t} z^{s+2t} = g_p(z, z^2). \tag{4.3}$$

Equations (4.2) and (4.3) were used in [111] to calculate the clustering coefficient C_1, defined (see Section 2.1) as

$$C_1 = \frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triples of vertices}} = \frac{3N_\triangle}{N_3}, \tag{4.4}$$

where the numerator

$$3N_\triangle = n \sum_{s,t} t p_{s,t} = n \left(\frac{\partial g_p}{\partial y} \right)_{x=y=1}, \tag{4.5}$$

and the denominator

$$N_3 = n \sum_k \binom{k}{2} p_k = \frac{n}{2} \left(\frac{\partial^2 f}{\partial z^2} \right)_{z=1}. \tag{4.6}$$

We can now appreciate the usefulness of the sequence t and the joint distribution $p_{s,t}$. By varying the number of triangles attached to each vertex in Eq. (4.5) one can directly increase or decrease the value of the clustering coefficient. Most importantly, the variable n, denoting the number of vertices in the graph, is present in both Eq. (4.5) and Eq. (4.6); it therefore cancels upon substitution into Eq. (4.4). Thus C_1 is independent of n and, in contrast to the models discussed earlier, diverging graph size ($n \to \infty$) cannot have a diminishing effect on the level of clustering.

In order to investigate dynamical processes on these graphs Newman defined what he calls excess degree distributions,

$$q_{s,t} = \frac{(s+1)p_{s+1,t}}{\langle s \rangle}, \tag{4.7}$$

and

$$r_{s,t} = \frac{(t+1)p_{s,t+1}}{\langle t \rangle}, \tag{4.8}$$

where $\langle s \rangle$ and $\langle t \rangle$ are the averages of s and t over all vertices. Here, $q_{s,t}$ is the distribution of the number of single edges and triangles attached to a vertex reached by traversing a single edge, excluding the traversed edge, and $r_{s,t}$ is the corresponding distribution associated with a vertex reached by traversing a triangle edge.[3] The generating functions for these distributions are

$$g_q(x,y) = \sum_{s,t} q_{s,t} x^s y^t = \frac{1}{\langle s \rangle} \sum_{s,t} s p_{s,t} x^{s-1} y^t = \frac{1}{\langle s \rangle} \frac{\partial g_p}{\partial x}, \tag{4.9}$$

and

$$g_r(x,y) = \sum_{s,t} r_{s,t} x^s y^t = \frac{1}{\langle t \rangle} \sum_{s,t} t p_{s,t} x^s y^{t-1} = \frac{1}{\langle t \rangle} \frac{\partial g_p}{\partial y}, \tag{4.10}$$

respectively. These two equations facilitate the derivation of analytical expressions for a number of important structural properties, including average path lengths, vertex connection probabilities, the size distribution of small components and the condition for the existence of the GCC. One can also use them to investigate various percolation processes: site percolation, bond percolation, joint site-bond percolation, etc.

In [111] Newman derived the following expression for the birth point of the GCC in terms of $\langle s \rangle$ and $\langle t \rangle$:

$$\left[\frac{\langle s^2 \rangle}{\langle s \rangle} - 2 \right] \left[2 \frac{\langle t^2 \rangle}{\langle t \rangle} - 3 \right] = \frac{2 \langle st \rangle^2}{\langle s \rangle \langle t \rangle}. \tag{4.11}$$

He also considered bond percolation on a graph with the doubly Poisson degree distribution

$$p_{s,t} = e^{-\mu} \frac{\mu^s}{s!} e^{-\nu} \frac{\nu^t}{t!}, \tag{4.12}$$

3 The concept of excess degree was first introduced in [114].

with μ chosen to be $\mu = \langle s \rangle$ and ν set to $\nu = \langle t \rangle$. In this case, the size of the GCC, called S, was shown to be

$$S = 1 - e^{-[\mu S + \nu S(2-S)]}. \tag{4.13}$$

In the left panel of Fig. 2 in [111] S was plotted against the bond occupation probability, which here we call ϕ_b, for several values of the clustering coefficient C_1. However, Newman did not calculate the critical values at which the GCC first appears in this figure. To do this it is necessary to derive an expression for the existence of the GCC in a bond percolation process on edge-triangle graphs in terms of $\langle s \rangle$, $\langle t \rangle$, and also ϕ_b. We will demonstrate in Section 5.1.2 how such a condition can be defined as part of our own generalized analytical approach to modelling the cascade dynamics of a broad range of processes run on Newman's graphs.

4.2.2 *Clique-based Graphs*

Newman's was not the only model of random graphs with tunable clustering coefficients to appear in 2009. In a paper published in September of that year Gleeson [71] also proposed a new way to generate such graphs. Like Newman's model, Gleeson's may be seen as a modifying the classical configuration model, with clustering introduced by embedding cliques of connected vertices within an otherwise tree-like structure. However, while Newman considered only 3-cliques, i.e. triangles, in his model, Gleeson's approach permits an entire distribution of clique sizes to be prescribed. Thus, [71] generalizes work based on similar ideas by Trapman [136], and Gleeson and Melnik [75].

To construct a realisation of a clustered graph in the manner described in [71] we first make a distinction between edges that connect vertices in the same clique, which we call *internal* edges, and those that connect vertices in different cliques, called *external* edges. We also stipulate that each vertex may be a member of at most one clique. Doing this allows the desired graph to be decomposed into a set of disjoint cliques connected together by external edges. In this way we can treat each clique as a vertex (or node) in its own right, called a *supernode* in [71]. With these conditions in place, a unique graph realisation can be drawn simply by employing the configuration model technique of pairing up at random stubs of edges. The only difference here is that the stubs we join together are now those of external edges between cliques (see Fig. 4.2).

The defining feature of a clique-based graph is its joint probability distribution $\gamma(k, c)$, which determines the probability that a randomly chosen

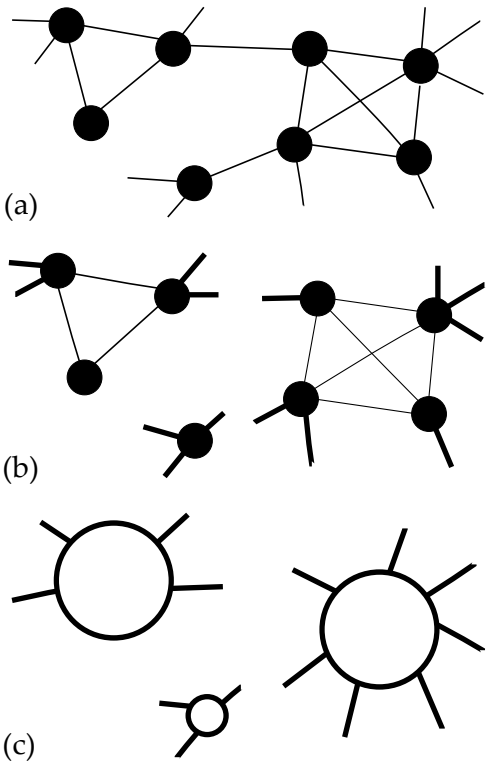


Figure 4.2: (a) Segment of a clustered random graph; (b) decomposed into disjoint cliques with external edges emphasized; (c) cliques seen as *supernodes* in a tree-like *supergraph*. Reproduced from Fig. 1 of [71].

vertex has degree k and belongs to a c-clique. Thus, the approach of [71] is sometimes referred to as γ-theory. Note, $\gamma(k,c) = 0$ for $k < c-1$ since each member in a clique of c vertices must possess enough incident edges to connect to every other member (excluding itself).

In the analysis presented in [71], Gleeson first demonstrated how the degree distribution of a γ-theory graph, p_k, is obtained by averaging $\gamma(k,c)$ over all clique sizes:

$$p_k = \sum_{c=1}^{k+1} \gamma(k,c) = \sum_{c} \gamma(k,c). \tag{4.14}$$

Next he showed how the degree dependent clustering coefficient c_k (see Section 2.1) can be defined in terms of $\gamma(k,c)$ as

$$c_k = \sum_{c} \frac{\gamma(k,c)}{p_k} \frac{(c-1)(c-2)}{k(k-1)}. \tag{4.15}$$

Having derived these expressions, he then proceeded to investigate some of the dynamical properties of his graphs. Formulas were found for the expected fractional size of the GCC in a bond percolation process, and the

critical bond occupation probability $\widehat{\phi_b}$. In contrast to [111], the analytical framework used to obtain these results did not involve generating functions. Instead, it was similar to, but evidently more complicated than, the tree-based approach derived in [73], and applied by us throughout Chapter 3. In Section 5.2 we will significantly extend this framework by presenting our own generalized analytical description of cascade dynamics on γ-theory graphs, which will include bond percolation as a special case. Therefore, to avoid repeating similar ideas here as later on, we omit for now the specific details of the analysis of [71], and state in short order its main outcome.

To summarize very briefly, then, by applying the concept of child to parent activation (see Section 3.1, and Fig. 2 of [71]) to graphs with embedded cliques, Gleeson derived the following expression for the expected GCC size in a bond percolation process with occupation probability ϕ_b:

$$S = \sum_{k,c} \gamma(k,c)\left[1-(1-\phi_b q_\infty)^{k-c+1}(1-Q_c)\right], \tag{4.16}$$

where Q_c is the probability that the parent vertex in a c-clique is active (with the remaining $c-1$ vertices in the clique treated as its children), and q_∞ is the steady state value of the level by level iteration, $q_{n+1} = G(q_n)$, defined by Eq. (5) of [71]. The same iterative equation also provided the first-order cascade condition $G'(0) = 1$ [70], which was expressed as a polynomial in ϕ_b (see Eq. (7) of [71]), and then solved to determine the critical value $\widehat{\phi_b}$.

The theory behind these two results will be more clearly elaborated in the generalized scheme put forth in Section 5.2, and we postpone all further discussion of dynamics on γ-theory graphs until then. There is, however, one other important aspect to the structural characterisation given in [71] worth highlighting here before we proceed.

We have mentioned that this model permits a distribution of clique sizes to be prescribed to a graph. In fact, it permits something much more significant than this. In [71] Gleeson showed that $\gamma(k,c)$ can be fitted to the degree distribution, p_k, and clustering spectrum, c_k, of any real-world network using the following parametrization

$$\gamma(k,c) = p_k \binom{k}{c-1} g_k^{c-1}(1-g_k)^{k-c+1}, \tag{4.17}$$

which is defined for $c = 1$ to $k+1$. To see how this idea works, first understand that Eq. (4.17) causes the distribution of clique sizes, c, occupied by vertices of degree k in a generated graph to adhere to a binomial distribution whose specific form is determined by the parameter g_k. Next, notice that substituting Eq. (4.17) into Eq. (4.15) reveals the remarkably

simple relationship $g_k = \sqrt{c_k}$. Thus, given a pair of measured p_k and c_k sequences, one can feed these directly into the right hand side of Eq. (4.17), produce a fitted joint distribution $\gamma(k, c)$, and then use that to define a more realistic ensemble of clustered random graphs. Furthermore, when substituted into Eq. (4.16) this fitted $\gamma(k, c)$ improves the accuracy of ρ and $\widehat{\phi_b}$, as evidenced by Fig. 3 of [71].

Finally note, however, that the binomial parametrization defined by Eq. (4.17) is fundamentally arbitrary, in that it was chosen without any theoretical justification other than the simple fact that it introduces no obvious structural bias into the distribution of c over k. Therefore, the equality of g_k to $\sqrt{c_k}$ was a rather serendipitous discovery, upon which a more in depth analysis may perhaps provide an improvement.

4.3 COMPARISON OF MODELS

The common goal of Newman's work in [111] and Gleeson's in [71] was to lay the theoretical foundations for the creation, ultimately, of a realistic network model that will accurately replicate the relevant structural features of observable complex systems.[4] In both papers the first steps towards this goal were taken by generalizing the configuration model such that it may used to create ensembles of highly clustered random graphs. Despite the appreciable unity of purpose between Newman and Gleeson in this respect, the distinguishing features of their respective analyses, most of which we have already outlined, may be viewed as setting the frameworks offered in [111] and [71] somewhat at odds with each other. However, at this early stage of development preference for one approach over the other remains largely a matter of opinion, or perhaps convenience.

For example, as regards the methods used by each author to model bond percolation on their graphs, Newman quite naturally employed an adaptation of the, by now classic, generating function techniques of [24], whereas Gleeson opted for a modified version of the tree-based theory of [73]. Undoubtedly, generating functions do carry with them an intrinsic mathematical elegance, and one may feel that they provide a more parsimonious framework in which, unlike [71], no reference need be made to levels of activation, or child and parent vertices, etc. On the other hand, some noteworthy commentators have argued quite convincingly that it is precisely this latter sort of physically intuitive grounding that makes the tree-based approach, and other complementary approaches [43, 44], easier

4 The adjective *relevant* in our usage means any feature that has an impact on the functional behaviours of the system; behaviours which we gain insight into by considering various idealized processes. After [99] this word will likely denote more than just high clustering, and may even include structural metrics that have not been thought of yet.

to understand and perhaps more readily generalizable to a wider range of processes. (Our work in Chapter 5 will shed further light on this point.)

There does, however, appear to be one aspect of the use of generating functions that stands clearly to Newman's advantage, and that is that they allow him to calculate the distribution of small component sizes up to the critical point at which the GCC first forms. The tree-based approach allows Gleeson to calculate the expected fractional size of the GCC but he cannot say anything about small components, at least not before first developing extensions to the current set of expressions given in [71]. Yet, from an applications perspective, this is not necessarily such a major shortcoming of [71] since, apart from the value of the critical point, the size of the GCC is usually the quantity of primary concern to us, particularly when we consider the question of resilience.

Besides these differences in the authors' approaches to modelling dynamical properties, there are also a number of fundamental differences in terms of the characteristics of real networks that their separate graph ensembles can capture. A number of pros and cons can be weighed in favour of or against either ensemble. For one, it may be argued that the way Newman introduces clustering is rather artificial. Nonoverlapping triangles are by no means the only structural motif by which clustering may be identified in a real network. An obvious advantage of Gleeson's graphs, in this respect, is that they permit an entire spectrum of clique sizes incorporating higher order motifs than simply triangles. Furthermore, as we have just seen, this clustering spectrum can be parametrized to fit measured values. But, then again, his graphs are constructed by first stipulating that each vertex may be a member of at most one clique. No such condition exists in the real-world, nor does it exist in Newman's model since more than one triangle may be attached to a single vertex. Thus the scales appear to be fairly evenly matched with regard to the question: Which model creates the most realistic network topology? The simple answer is that neither model creates graphs that look very much like real networks. Of course, this does not take away from the importance of either model as a theoretical point of departure for more ambitious investigations.

Lastly, however, in relation to clustering, we note that there is one other property of Gleeson's approach, besides its parametrizability, that would seem to tip the scales slightly in his favour. In Newman's model the local clustering coefficient is necessarily limited by $c_k \leqslant 1/(k-1)$ since a k-degree vertex can belong to at most $k/2$ disjoint triangles, and there are also strict bounds on the range of values achievable for the global clustering coefficient C_1 which depend on the graph topology under consideration. Restrictions of this sort do not apply to Gleeson's model: $C_1 \in [0, 1)$.

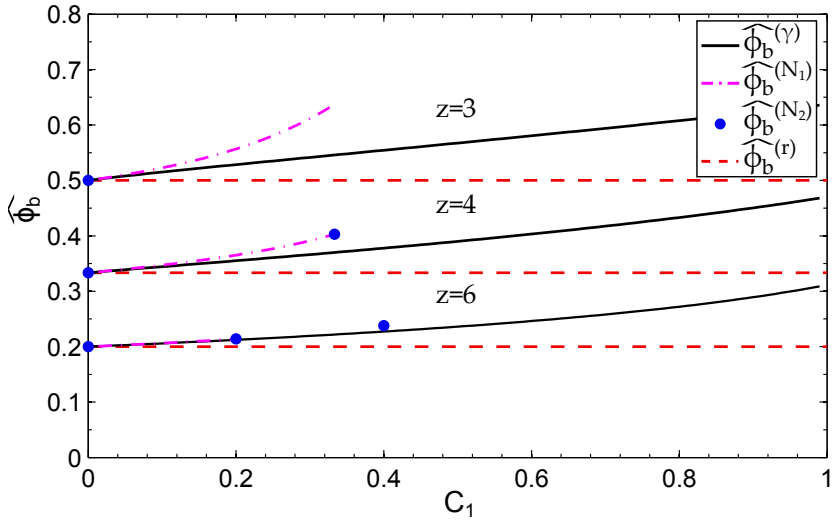


Figure 4.3: Bond percolation threshold $\widehat{\phi_b}$ vs. clustering coefficient C_1 for z-regular graphs with $z = 3$ (top), $z = 4$ (middle), and $z = 6$ (bottom). Red dashed lines correspond to the nonclustered value $\widehat{\phi_b}^{(r)} = 1/(z-1)$; Black solid: values of $\widehat{\phi_b}^{(\gamma)}$ from [71]; magenta dot-dashed: $\widehat{\phi_b}^{(N_1)}$ from [111]; and blue circles: $\widehat{\phi_b}^{(N_2)}$ from [109]. Source [76].

In [76], we (Gleeson, Melnik, and Hackett) looked very closely at the levels of clustering permissible in either approach and also at the effects that different levels may have on the bond percolation threshold. To get a flavour of the type of analysis carried out in [76], consider a bond percolation process on a z-regular (each vertex has z neighbours) γ-theory graph. By adapting the results of [71], we showed that the threshold for such a graph is given by the root in ϕ_b of the equation

$$\sum_c \frac{(z-c+1)}{(1-\sqrt{C_1})z}\binom{z}{c-1}C_1^{\left(\frac{c-1}{2}\right)}(1-\sqrt{C_1})^{z-c+1} \times\left[(z-c)\phi_b+(z-c+1)D_c(\phi_b)\right]=1, \tag{4.18}$$

where $D_c(\phi_b)$ are polynomial functions of ϕ_b defined in [71]. For any $C_1 \in [0,1)$, Eq. (4.18) can be solved to find the corresponding critical value, $\widehat{\phi_b}$. In [76] we used this equation to compare the clustering properties of the z-regular graphs produced by Gleeson's model [71] to those of Newman's [111], and also those of an earlier bipartite model of Newman's [109].[5] The most explicit illustration of the differences between these three models is given by Fig. 1(a) of [76] which we reproduce here as Fig. 4.3.

This figure clearly demonstrates our point about the range of C_1 values attainable in each model; although, note that the full analysis of [76] concerning this question extends beyond the z-regular subclass to the arbitrary

5 For [109] $\widehat{\phi_b}$ can be calculated only at certain $\{C_1, z\}$ pairings. See Appendix A of [76].

p_k ensemble. Moreover, it also illustrates a very important fact about the effects of clustering on dynamics; a fact which strongly supports our criticism of the error measure E [99] at the beginning of the chapter. Namely, in each model the introduction of clustered vertices increases the percolation threshold above its value on the nonclustered graph (red dashed). This effect may be observed consistently across each range of C_1 values and for varying z. The extent of the increase is greater in Newman's models than in Gleeson's; however, these differences diminish as z grows and the graphs become more heterogeneous.[6]

This latter aspect of bond percolation on clustered graphs, which from what we can tell so far appears to be a generic phenomenon among existing structural models, can be framed as part of the broader problem of determining the shift in the critical value for a more general class of dynamical processes. For instance, we might ask: Does increased clustering have a similar effect for site percolation, or Watts's model? We will address this question in some detail in Section 5.1.3.

6 At $z = 6$ the $\widehat{\phi_b}^{(N_1)}$ (magenta) line almost coincides with the $\widehat{\phi_b}^{(\gamma)}$ (black) line.

5

CASCADES AND CLUSTERING: A SYNTHESIS

In Section 3.3 we discussed the effectiveness and ultimate limitations of the locally tree-like approximation in application to real-world network topologies. Referring to our work in [99], we showed that the accuracy of an analytical theory of dynamics derived from this approximation (specifically $P(k, k')$-theory [70]) appears to depend more on the mean intervertex distance, L, of the observed network than on either definition of the global clustering coefficient, C_1 or C_2. However, we also established that the analysis of [99] did not present enough evidence to rule out entirely the possibility that clustering has some sort of effect on dynamics, particularly in the range of parameters close to the global cascade transition. Thus, given the broadly recognized significance of clustering as a structural characteristic of real networks, we proceeded in Section 4.2 to present two of the most successful recent attempts [71, 111] at creating ensembles of random graphs with tunable clustering coefficients. Comparing these models revealed (see Fig. 4.3) that indeed clustering does have an effect on cascade dynamics similar to what we had speculated, at least apropos the process of uniform bond percolation, and perhaps other processes as well. Thus we arrive at the beginning of this penultimate chapter of our dissertation confident that the task set before us of providing independent analytical descriptions of cascade dynamics on these two new ensembles of clustered random graphs is one which may provide insights of genuine novelty and importance.

We have said that we will undertake this task by seeking extensions of the original approach of [73]. To elaborate, what we have in mind here is to account for clustering from within the framework of level-by-level tree-based activations. In this way we aim to provide a unique synthesis of tree-like (branching) cascade propagation and clustered structural motifs; combining what we have learned about modelling processes in Chapter 3 with the insights into network structure gained in Chapter 4. We fully appreciate that, in view of all that we have learned so far, this stated goal of ours may appear to the reader a contradiction in terms; however, we shall demonstrate in the results that follow that the distinctive features of Newman's and Gleeson's graph ensembles permit its realization without recourse to excessively complicated mathematical techniques, and allow us to retain the familiar language of Chapter 3. We consider each model

separately, dividing our presentation into two broad sections. The analytical approach of Section 5.1 applies to cascades on edge-triangle graphs, and has appeared previously in [85]. The approach described in Section 5.2, applies to cascades on clique-based graphs, and is as yet unpublished [83].

We concede that this dependence on the individual traits of each graph ensemble means that both of these approaches, like the ensembles themselves, will fall considerably short of being ideal analogues of the real world. Notwithstanding, together they will undoubtedly provide the theoretical scope for further development in this direction. One other recent highlight from the networks literature [91] also provides this scope. Thus, we will discuss in the closing remarks of Section 5.3 potential avenues towards a more fully realized, and preferably unified framework; one in which the results of this chapter may perhaps be derived as special cases.

5.1 CASCADES ON EDGE-TRIANGLE GRAPHS

We begin in this section by showing how the theory put forth in [73] (see Section 3.1) and further developed in [70] for cascades on locally tree-like graphs can be modified such that it is applicable to the ensemble of clustered random graphs introduced in [111]. The work of [70], which we have previously merely alluded to, showed how through the response function mechanism the basic theory of [73] can be applied to any process that satisfies the following set of properties: (i) each vertex is assigned a binary value specifying its current state, *active* (*damaged* or *infected*) or *inactive* (*undamaged* or *susceptible*); (ii) the probability of a vertex becoming active (in a synchronous update of all vertices) depends only on its degree k and the number m of its neighbours who are already active, i.e., the response function $F(m, k)$; (iii) for any fixed degree k, $F(m, k)$ is a nondecreasing function of m; and (iv) once active, a vertex cannot become deactivated. Our work in Chapter 3 verified that the process described by Watts's model satisfies these constraints. Other confirmed members of the class of processes that can be modelled by choosing an appropriate $F(m, k)$ [70] are site and bond percolation [21, 134], and k-core decomposition [77, 47]. (It remains to be seen if more may be included.) Thus, when we refer to providing a *generalized* analytical description for Newman's ensemble, what is implied by this is an approach that can model the same broad class of processes but on edge-triangle graphs, with joint distribution $p_{s,t}$.

The next subsection describes in detail the methods behind our extension of the approach of [73], and includes our analytical calculation of the expected cascade size and our condition for the existence of global cascades. Following that, Section 5.1.2 will show how to model different processes

using these expressions by defining appropriate response functions. We conclude our discussion of $p_{s,t}$ graphs in Section 5.1.3 with a thorough investigation of the effects of clustering on cascades.

5.1.1 *Cascade Propagation*

The key observation that justifies our use of the locally tree-like approximation in the presence of nonzero clustering and permits our modification of the equations of Section 3.1 is simply the following: In $p_{s,t}$ graphs clustering is generated solely through the motif of nonoverlapping triangles. Therefore, we posit that fitting this specific type of clustering into the tree-based framework is quite straightforward: a triangle exists whenever an edge connects two vertices on the same level (see Fig. 5.1). This supposition appears to us to be fairly intuitive; however, it will be validated conclusively only by the accuracy of the equations derived. Taking it for the moment to be legitimate, we see that the same basic conception of propagation of activations up the levels of a tree towards a randomly chosen root vertex as was applied in Section 3.1 (see Fig. 3.1) may also be applied here. The only significant difference pertaining uniquely to $p_{s,t}$ graphs is that now we are faced with two distinct ways in which activations may spread from one level to the next. They may spread as in Fig. 5.1(a) from a child (c) to its parent (p) across a single edge or as in Fig. 5.1(b) from either child at the base of a triangle to the parent at its apex.

Following the methodology of Section 3.1 then, let us model a generalized cascade as a recursive sequence of activations from child to parent and set up self-consistent equations for the probabilities involved.

Considering first Fig. 5.1(a), let σ_1 be the probability that the child is active conditional on its parent being inactive, and let $\sigma_0 = 1 - \sigma_1$ be the corresponding conditional probability that the child is inactive. For convenience we represent this set of probabilities with the generating function $\sigma(x) = \sigma_0 + \sigma_1 x$. Similarly, in Fig. 5.1(b), let τ_2 be the probability that both children are active, conditional on their parent being inactive, let τ_1 be the conditional probability that only one child is active, and let $\tau_0 = 1 - \tau_1 - \tau_2$ be the conditional probability that neither child is active. The generating function for these probabilities is $\tau(x) = \tau_0 + \tau_1 x + \tau_2 x^2$.

Of course, the vertex arrangements represented by Figs. 5.1(a) and 5.1(b) usually exist in various combinations, and not exclusively of each other. By definition, in any given graph realization a randomly chosen vertex will be directly connected to s vertices via single edges and to $2t$ vertices via triangle edges, with probability $p_{s,t}$. Therefore, letting $\Pi_m^{s,t}$ be

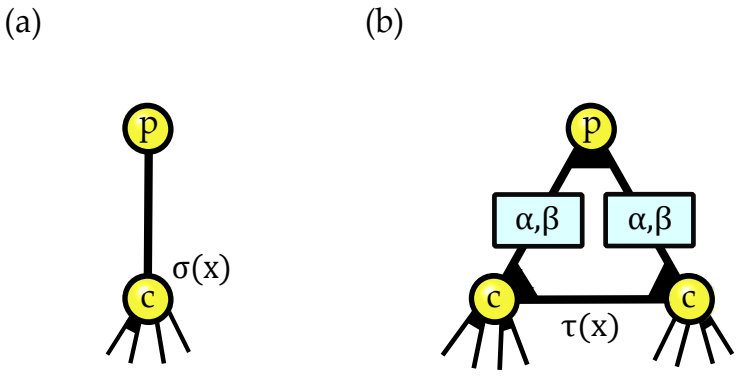


Figure 5.1: Level-by-level cascade propagation in a $p_{s,t}$ graph using the tree approximation. Triangle corners are marked in black. Source [85].

the probability that m of these neighbouring vertices are active, $\sigma(x)$ and $\tau(x)$ can be related to that probability by a third generating function

$$G(x) = \sum_{m=0}^{s+2t} \Pi_m^{s,t} x^m = \left[\sigma(x)\right]^s \left[\tau(x)\right]^t, \tag{5.1}$$

which is defined for each pairing of s and t.

We are now in a position to write an analytical expression for σ_1. In terms of an arbitrary response function $F(m, s+2t)$, written F_m for short, we have

$$\sigma_1 = \rho_0 + (1-\rho_0) \sum_{s,t} \frac{s p_{s,t}}{\langle s \rangle} \sum_{m=0}^{s+2t-1} \Pi_m^{s-1,t} F_m, \tag{5.2}$$

where ρ_0 is the seed fraction and $\langle s \rangle = \sum_{s,t} s p_{s,t}$ is the average number of single edges per vertex. Equation (5.2) is a self-consistent equation for σ_1 since according to Eq. (5.1), $\Pi_m^{s,t}$ is itself a function of the coefficients of $\sigma(x)$ and $\tau(x)$. We can read Eq. (5.2) as follows: The probability of the child vertex in a randomly chosen single edge pair being active, conditional on its parent being inactive, is equal to the probability that it was either initially active (ρ_0), or that $(1-\rho_0)$ it subsequently became active by copying the behaviour of the m out of $s+2t-1$ of its own children that were already active. Note, the term $s p_{s,t}/\langle s \rangle$ is the probability of reaching a child with s single edges by travelling along a random single edge from its parent (see Eq. (3.4) of Section 3.1, and [110]).

To obtain similar expressions for τ_1 and τ_2 we must reflect the fact that in a triangle the state of either child may influence the state of the other. Referring to Fig. 5.1(b), the probability that one child is active regardless of the state of the other is

$$\alpha = \rho_0 + (1-\rho_0) \sum_{s,t} \frac{t p_{s,t}}{\langle t \rangle} \sum_{m=0}^{s+2(t-1)} \Pi_m^{s,t-1} F_m, \tag{5.3}$$

the probability that one child is inactive if the other is inactive but will activate if the other is active is

$$\beta = (1-\rho_0)\sum_{s,t}\frac{tp_{s,t}}{\langle t\rangle}\sum_{m=0}^{s+2(t-1)}\Pi_m^{s,t-1}\left[F_{m+1}-F_m\right], \tag{5.4}$$

and finally the probability that one child is inactive even if the other is active is $\gamma = 1-\alpha-\beta$. In Eqs. (5.3) and (5.4), we use the fact that following a triangle edge from the parent leads to a child with t triangles with probability $tp_{s,t}/\langle t\rangle$. This child then has s single edges and $t-1$ triangles available to connect to its own children, giving its maximum number of active children (for the sum over m) as $s+2(t-1)$. Expressed in terms of the probabilities α and β, self-consistent expressions for τ_1 and τ_2 are given by

$$\tau_1 = 2\alpha\gamma, \tag{5.5}$$

and

$$\tau_2 = \alpha^2 + 2\alpha\beta. \tag{5.6}$$

The form of Eq. (5.5) arises from the fact that the probability of the parent in a triangle of vertices having one active child is equal to the probability that one child is active regardless of the state of the other (α), while the other is inactive regardless of the state of the other (γ), and there are two different ways in which this may be the case. Reading Eq. (5.6) in the same way, we see that the probability of the parent vertex in a triangle having two active children is equal to the probability that both children are active regardless of each others' states (α^2) plus the probability that one child is active and the other activates because of this with probability β; again there are two ways in which the latter may occur.

The propagation of a cascade through a $p_{s,t}$ graph is now almost fully defined. Given a seed fraction ρ_0, we solve Eqs. (5.1)-(5.6) to find the steady-state values of the coefficients of the polynomials $\sigma(x)$ and $\tau(x)$, and then, using these, we determine the expected cascade size in the familiar manner by calculating the probability of activation of the root vertex. This final probability is given by

$$\rho = \rho_0 + (1-\rho_0)\sum_{s,t}p_{s,t}\sum_{m=0}^{s+2t}\Pi_m^{s,t}F_m. \tag{5.7}$$

Comparing this equation to Eq. (5.2), we see that here the root vertex, which has s single edges and t triangles with probability $p_{s,t}$, has no parent and so has $s+2t$ children.

We leave the verification of the analytical approach derived here to Section 5.1.3. The figures presented there will show that Eq. (5.7) and Ineq. (5.11) (below) are in excellent agreement with the results of numerical simulations of site percolation and Watts's model on $p_{s,t}$ graphs.

5.1.1.1 *Cascade condition*

Having established an analytical expression for the expected cascade size in Eq. (5.7), we now turn to the derivation of a cascade condition. This will determine the circumstances under which the process of propagating activations described by Eqs. (5.1)-(5.6) can generate a nonvanishing mean cascade size from an infinitesimally small seed fraction $\rho_0 \to 0$.

We begin by observing that Eqs. (5.1)-(5.6) can be represented as the steady state of a nonlinear system of the general form $\mathbf{v}^{(n+1)} = \mathbf{H}(\mathbf{v}^{(n)})$, where $\mathbf{v}^{(n)} = \left[{\sigma_1}^{(n)}, {\tau_1}^{(n)}, {\tau_2}^{(n)}\right]$.[1] The trivial solution $\mathbf{v} = \mathbf{0}$ corresponds to an equilibrium state where cascades do not occur. We can look for other solutions by applying a small perturbation away from this equilibrium and then considering the trajectories in a linearized version of the system.

Applying this method we first linearize the generating function $G(x)$ of Eq. (5.1) about $\mathbf{v} = \mathbf{0}$ using a small parameter ϵ to measure the magnitude of the perturbation. Scaling the coefficients of $\sigma(x)$ and $\tau(x)$ as $\mathcal{O}(\epsilon)$, that is $\sigma_1 \simeq \epsilon\widetilde{\sigma_1}$, $\tau_1 \simeq \epsilon\widetilde{\tau_1}$ and $\tau_2 \simeq \epsilon\widetilde{\tau_2}$, we expand $G(x)$ as

$$G(x) \simeq 1 - \epsilon\left[s\widetilde{\sigma_1} + t(\widetilde{\tau_1} + \widetilde{\tau_2}) - (s\widetilde{\sigma_1} + t\widetilde{\tau_1})x - t\widetilde{\tau_2}x^2\right], \tag{5.8}$$

up to terms of $\mathcal{O}(\epsilon^2)$.

Our next step will be to substitute the coefficients of $G(x)$ from Eq. (5.8) into Eqs. (5.2)-(5.6). Before doing this, however, we further simplify our analysis by assuming $F_0 = 0$. This implies that a vertex will never activate if none of its neighbours are active, and this is true, or a good approximation, in many cases of interest. With $F_0 = 0$, then, said substitution gives us a linear system that may be represented in the matrix form $\widetilde{\mathbf{v}}^{(n+1)} = \mathbf{A} \cdot \widetilde{\mathbf{v}}^{(n)}$, where

$$\widetilde{\mathbf{v}}^{(n)} = \begin{bmatrix} \widetilde{\sigma_1}^{(n)} \\ \widetilde{\tau_2}^{(n)} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{5.9}$$

and the elements of $\mathbf{A}$ are

$$A_{11} = \frac{\langle (s^2 - s)F_1 \rangle}{\langle s \rangle}, \quad A_{12} = \frac{\langle stF_2 \rangle}{\langle s \rangle} + \frac{\langle stF_1 \rangle}{\langle s \rangle}\frac{\langle t \rangle - \langle tF_1 \rangle}{\langle tF_1 \rangle};$$

1 In fact, this is how the expressions of Section 3.1 were presented. We could have dropped the n and $n+1$ subscripts from Eqs. (3.7) and (3.8), and referred to them also as *self-consistent*; however, for the sake of simplicity we kept them iterative.

$$A_{21} = \frac{2\langle stF_1\rangle\langle tF_1\rangle}{\langle t\rangle^2},$$

$$A_{22} = \frac{2\langle (t^2-t)F_1\rangle}{\langle t\rangle} + \frac{2\langle (t^2-t)(F_2-F_1)\rangle\langle tF_1\rangle}{\langle t\rangle^2}. \tag{5.10}$$

Note, the application of Eq. (5.8) has allowed us to express $\widetilde{\tau_1}^{(n)}$ in terms of $\widetilde{\tau_2}^{(n)}$ as $\widetilde{\tau_1}^{(n)} = (\langle t\rangle - \langle tF_1\rangle)\widetilde{\tau_2}^{(n)}/\langle tF_1\rangle$, hence the reduction to the 2×2 system of linear equations represented by Eqs. (5.9) and (5.10).

In order for this system to produce trajectories that will diverge from $\mathbf{v} = \mathbf{0}$, in other words in order to produce cascades, we require that the larger eigenvalue of $\mathbf{A}$ (both eigenvalues are real) be greater than one, $\lambda_+ > 1$.[2] This condition is satisfied (see Appendix B.1) if

$$\begin{aligned}\langle t\rangle\big\{&2\langle stF_1\rangle^2 - \big[\langle (s^2-s)F_1\rangle - \langle s\rangle\big]\big[2\langle (t^2-t)F_1\rangle - \langle t\rangle\big]\big\}\\ &- 2\langle tF_1\rangle\big\{\big[\langle (s^2-s)F_1\rangle - \langle s\rangle\big]\langle (t^2-t)(F_2-F_1)\rangle\\ &- \langle stF_1\rangle\langle st(F_2-F_1)\rangle\big\} > 0.\end{aligned} \tag{5.11}$$

Conversely, if the left-hand side of Ineq. (5.11) is negative, then $\lambda_+ < 1$, and the trivial equilibrium is stable, so cascades do not occur. The boundary between these two regimes, one where cascades are observed and the other where they are not, is located precisely at the point where $\lambda_+ = 1$; that is, where the left-hand side of Ineq. (5.11) is equal to zero.

5.1.2 *Response Functions*

In this subsection we will show how our new generalized analytical approach may be used to model a range of processes on $p_{s,t}$ graphs. As stated previously each specific process will be defined by choosing an appropriate response function, and Eq. (5.7), in combination with Eqs. (5.1)-(5.6), will then give the expected cascade size (or, for percolation, the fractional GCC size). We consider in detail site and bond percolation, and Watts's model.

5.1.2.1 *Site and bond percolation*

Beginning with percolation, we frame our description of this process in the language of successive activations already introduced. Thus, we define a vertex as active if it is part of the GCC (percolating cluster) of the graph, and our choice of response function, Eq. (5.12) or Eq. (5.16) below, will

2 In general, the trivial equilibrium $\mathbf{v} = \mathbf{0}$ is unstable when the matrix $\mathbf{A}$ has at least one eigenvalue λ such that $|\lambda| > 1$. Our simplified requirement follows from the fact that $\mathbf{A}$ consists of real positive elements and thus according to the Perron-Frobenius theorem at least one of the eigenvalues of $\mathbf{A}$ is real and positive, and is greater than the other in absolute value.

determine the type of percolation under consideration, either uniform site percolation or uniform bond percolation, respectively. This interpretation works because when the activation process defined by Eqs. (5.1)-(5.7), with the appropriately substituted response function, reaches the steady state, any vertex that is labelled as active will have at least one active neighbour to which it is connected. Therefore, the fraction ρ of these active vertices will correspond to the size of the connected component, expressed as a fraction of the graph size n. In the $n \to \infty$ limit, only the GCC size scales with n, and so ρ will match the fractional size of the GCC, S. This can be seen also from the fact that in the limit of zero clustering, our equations reduce to the standard percolation equations for the GCC size in configuration model graphs, as given in [24]. Note, however, that, like every other variation of the tree-based theory [70, 71, 73, 75], this method does not permit the calculation of finite-size connected components.

In uniform site percolation, each vertex is occupied with independent probability ϕ_s, and an occupied vertex can become active in the cascade, i.e., form part of the GCC, if it has one or more active neighbours (who are already in the GCC). Unoccupied vertices can never become active. The response function for site percolation is therefore [70]

$$F(m, s+2t) = \begin{cases} \phi_s & \text{if} \quad m > 0, \\ 0 & \text{if} \quad m = 0. \end{cases} \tag{5.12}$$

Using Eq. (5.12) in the $\rho_0 \to 0$ limit of Eqs. (5.2)-(5.7), and noting that with this choice of response function

$$\sum_{m=0}^{s+2t} \Pi_m^{s,t} F(m, s+2t) = \phi_s \left[1 - {\sigma_0}^s {\tau_0}^t\right], \tag{5.13}$$

the fractional size of the GCC (as $n \to \infty$) is given by Eq. (5.7) (with $\rho \equiv S$) and reduces to the simple form

$$S = \phi_s - \phi_s \sum_{s,t} p_{s,t} {\sigma_0}^s {\tau_0}^t. \tag{5.14}$$

Likewise, substituting Eq. (5.12) into the left hand side of Ineq. (5.11) and setting it equal to zero we derive the equation [84]

$$\left(\phi_s \langle s^2 - s\rangle - \langle s\rangle\right)\left(2\phi_s \langle t^2 - t\rangle - \langle t\rangle\right) - 2{\phi_s}^2 \langle st\rangle^2 = 0, \tag{5.15}$$

whose solution in ϕ_s determines the critical site occupation probability $\widehat{\phi_s}$.

In uniform bond percolation, on the other hand, each edge is occupied with independent probability ϕ_b, and a vertex can become active only if it

is linked to another active vertex by an occupied edge. Thus, a vertex with m active children has probability $1-(1-\phi_b)^m$ of becoming active itself. The appropriate choice of response function in this case is therefore [70]

$$F(m, s+2t) = \begin{cases} 1-(1-\phi_b)^m & \text{if} \quad m > 0, \\ 0 & \text{if} \quad m = 0, \end{cases} \tag{5.16}$$

which upon substitution into Eqs. (5.2)-(5.7) yields

$$\begin{aligned} S = 1 - \sum_{s,t} p_{s,t} \left[\sigma_0 + \sigma_1(1-\phi_b)\right]^s \\ \times \left[\tau_0 + \tau_1(1-\phi_b) + \tau_2(1-\phi_b)^2\right]^t, \end{aligned} \tag{5.17}$$

as the fractional GCC size for this type of percolation. Similarly, applying Eq. (5.16) to Ineq. (5.11) gives us the following equation for the critical bond occupation probability $\widehat{\phi_b}$ [76]:

$$\left(\phi_b\langle s^2 - s\rangle - \langle s\rangle\right)\left(2f\phi_b\langle t^2 - t\rangle - \langle t\rangle\right) - 2f{\phi_b}^2\langle st\rangle^2 = 0, \tag{5.18}$$

where $f : \phi_b \to f(\phi_b) = 1 + \phi_b - {\phi_b}^2$.[3]

The approach outlined here is also applicable to two other closely related problems: SIR contagion dynamics [80, 108] and k-core decomposition [47, 77]. From [108] (see Section 2.3.2) we know that the steady state infected fraction in an SIR process can be mapped directly to the GCC size in bond percolation. The topic of k-cores was discussed in detail in [70] and the relevant response function for standard configuration model graphs was provided in Eq. (10) of that paper. With the introduction of triangles we simply update that response function by setting $k = s + 2t$, and then continue as above by performing appropriate substitutions of $F(m, s+2t)$.

Finally, in relation to network-oriented epidemiology, we note that the question of how clustering in networks of human interactions may influence the size and persistence of outbreaks of infectious diseases has motived a number of recent studies [10, 20, 52, 87, 102, 136]. In fact, much of the impetus for considering more complex topological motifs in networked structures in general has come from this source (see [75, 101, 136] and references therein). We will see in Section 5.1.3 how the results obtained by us for site and bond percolation echo (albeit indirectly) some of the major results from this literature concerning the effects of clustering.

3 Note that by setting $\phi_s = 1$ in Eq. (5.15) and $\phi_b = 1$ in Eq. (5.18) both equations reduce to Newman's original Eq. (22) of [111] (reproduced in Section 4.2.1 as Eq. (4.11)).

5.1.2.2 *Watts's model*

Turning next to Watts's model, we recall from our review in Section 2.3.2.1 that in its most abstract formulation this model describes a type of binary-state dynamics on the vertices of a random graph of arbitrary degree distribution, p_k. The active fraction requires no great elaboration in this case, since it has already been discussed at length throughout Chapter 3. To recap briefly, it simply corresponds to the relative number of participants in the cascade (of some fashion, rumour, etc.), where the participation of a vertex i at time t is indicated by the state $\nu_i(t) = 1$, and non-participation by $\nu_i(t) = 0$. From the definition of the decision rule in, Eq. (2.13), a vertex will change state from zero to one if the fraction of its neighbouring vertices that are already active (m/k) exceeds the value of its random (frozen) threshold r, which is drawn from a specified distribution $q(r)$. Otherwise, $\nu_i(t)$ will remain unchanged. Thus, if during a synchronous update of all states, a vertex decides to activate, from that point on it may not change state ever again. It is this crucial feature that permits us to map the average steady-state active fraction in this process to the expected cascade size, ρ. The results of Chapter 3 have demonstrated this last point quite extensively.

In [70] Gleeson defined the response function for Watts's model on a graph with both arbitrary $q(r)$ and arbitrary p_k (see Eq. (2) of [70]). We can extend this definition to $p_{s,t}$ graphs, in precisely the same way as before, simply by setting $k = s + 2t$. From Eq. (2) of [70] this gives us

$$F(m, s + 2t) = H_r\left(\frac{m}{s + 2t}\right), \tag{5.19}$$

where m is the number of active neighbours and H_r denotes the CDF of the thresholds. The form that Eq. (5.19) takes when a uniform threshold distribution is applied can be seen by substituting $k = s + 2t$ into Eq. (3.3) of Section 3.1. If we require a Gaussian threshold distribution with mean R and standard deviation σ, then Eq. (5.19) becomes

$$F(m, s + 2t) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{m/(s + 2t) - R}{\sigma\sqrt{2}}\right)\right], \tag{5.20}$$

where $\operatorname{erf}(x)$ is the error function. Note, $F(0, s + 2t) > 0$ here, meaning some vertices have negative thresholds, and so will activate even if none of their neighbours are active. It is possible, therefore, for such vertices to instigate a cascade even when $\rho_0 = 0$ (see the discussion pertaining to Fig. 3.3).

As usual, we obtain the expected cascade size and the cascade condition for Watts's model by substituting the appropriate form of Eq. (5.19) into Eqs. (5.1)-(5.7) and Ineq. (5.11) from above.

5.1.3 *The Effects of Clustering*

The results of Section 5.1.1 in combination with the response functions listed in the previous section will supply us with a diverse set of tools with which to model cascades on $p_{s,t}$ graphs. Provided, that is, we can verify that the reasoning that underlies their derivation is, indeed, sound. In this section we shall seek this verification by the familiar method of plotting predicted values from our analytical calculations against the results of numerical simulations of various processes. However, we shall do more than merely verify. These plots will also be used to address an important question concerning the effects of clustering on cascades.

We showed in Fig. 4.3 of Section 4.3 that increasing the level of clustering in the z-regular graphs of Newman's $p_{s,t}$, or Gleeson's $\gamma(k,c)$ ensemble will unambiguously increase the value of the critical occupation probability (percolation threshold) in a bond percolation process run on either of these types of graphs [76]. Therefore, we asked: Does a similar effect occur for a more general class of processes? Evidently, the class we were alluding to is the one cited at the beginning of this chapter. Also, after our review of [99] in Section 3.3 we proposed at the beginning of Chapter 4 that the relationship between the expected cascade size[4], ρ, and the level of clustering may be determined by looking closely at the parameter range near the cascade threshold. Both of these aspects of the effects of clustering are directly related to each other since a change in the position of the threshold is, in point of fact, a change in the cascade size over the distance the threshold has moved relative to its original position. An increased threshold corresponds to a reduction in the value of ρ, and a decreased threshold corresponds to an increase in ρ (see Fig. 5.2 below).

In order, therefore, to address both aspects simultaneously, we reframe the question above as: Does the presence of clustering in $p_{s,t}$ graphs increase or decrease the expected cascade size relative to its value in a nonclustered graph with the same degree distribution, p_k? The answers provided to this question are to be taken with the proviso that since they will be inferred from the position of the cascade threshold they will apply conclusively only to the region over which it has moved. Furthermore, we must ensure that the way we add clustering does not change p_k, since were it to do so, we would not be permitted to unambiguously attribute the change in the threshold to the effects of clustering.

This last point indicates why we have chosen to conduct our investigation from within the conceptual framework of the edge-triangle model. The simplicity of both the model itself and our description of cascades on the

4 The height of the cascade window in Fig. 3.8.

graphs that it produces affords us the flexibility to vary clustering in an uncomplicated manner, without altering p_k.

Consider, for example, the following choice joint distribution:

$$p_{s,t} = p_k \delta_{k,s+2t} \left[(1-g)\delta_{t,0} + g\delta_{t,\lfloor (s+2t)/2 \rfloor} \right], \tag{5.21}$$

where $g \in [0,1]$, $\delta_{i,j}$ is the Kronecker delta, and $\lfloor \cdot \rfloor$ is the floor function. By applying this definition, we construct $p_{s,t}$ from a given degree distribution p_k such that a fraction g of the vertices in our graph are attached to the maximum possible number of triangles $t = \lfloor (s+2t)/2 \rfloor$ while the remaining $(1-g)$ are attached to single edges only ($t = 0$). Upon substitution of Eq. (5.21) into Eq. (4.4) of Section 4.2.1 we find that the clustering coefficient C_1 can be expressed as

$$C_1 = g \frac{\sum_k k(p_{2k} + p_{2k+1})}{\sum_k \binom{k}{2} p_k}. \tag{5.22}$$

This linear relationship between C_1 and g allows us to increase C_1 continuously from its minimum value at $g = 0$ to its maximum possible value obtained at $g = 1$, while preserving p_k throughout. We cannot guarantee, however, that degree-degree correlations will be preserved [76].

In Fig. 5.2 (below) we have used Eq. (5.21) to verify our approach of Sections 5.1.1 and 5.1.2 in the case of site percolation on $p_{s,t}$ graphs with Poisson degree distribution $p_k = z^k e^{-z}/k!$. We plot our result for the GCC size from Eq. (5.14) against numerical simulations for two different values of the mean degree $z = \sum_k k p_k$. In both cases we consider minimum clustering ($g = 0$) and maximum clustering ($g = 1$). Threshold values defined by Eq. (5.15) are also plotted as yellow pentagrams (see caption).

Observing the relative positions of the percolation thresholds in Fig. 5.2 we note that they lend support in favour of (or, at least, do not contradict) the hypothesis that adding clustering will decrease the cascade size by virtue of increasing the threshold. From Fig. 4.3 (and [76]), we know that this is true for bond percolation on $p_{s,t}$ graphs with $p_k = \delta_{k,z}$; however, since the presence of clustered edges in a z-regular graph cannot affect its correlation structure, this means that any effects that may have been introduced by allowing correlations to vary were automatically negated in that figure. Furthermore, it was explicitly demonstrated in [76], and also [101], that such effects may significantly complicate matters. In Fig. 5.2, on the other hand, degree-degree correlations are not preserved. Therefore, while this figure does provide a validation of our generalized approach, it does not permit us to draw definitive conclusions as regards the question of the change in the expected cascade size due to clustering alone.

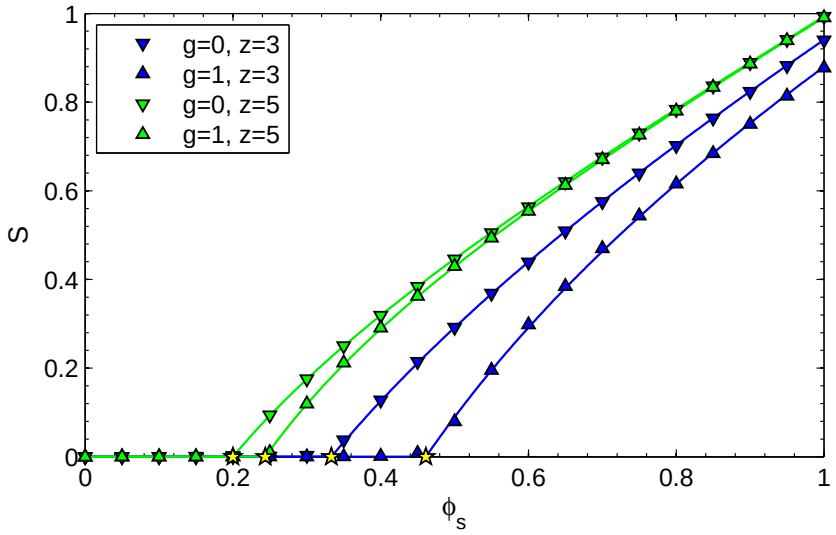


Figure 5.2: Site percolation on $p_{s,t}$ graphs with $n = 10^5$ and Poisson p_k. Numerical simulations (triangles) averaged over 100 realisations and theory of Section 5.1.1 (lines). GCC size S vs. site occupation probability ϕ_s. Colour indicates mean degree: $z = 3$ blue; and $z = 5$ green. In both cases we consider minimum clustering $g = 0$ and maximum clustering $g = 1$. In each of the four parameter settings we calculate $\widehat{\phi_s}$ from Eq. (5.15) and mark its position on the ϕ_s axis with a yellow pentagram.

In order to do that we will follow the approach of [76] (see also [25]) and focus our investigation on $p_{s,t}$ graphs with z-regular p_k. In particular, we consider the following joint distribution

$$p_{s,t} = \delta_{z,s+2t}\left[(1-g)\delta_{t,0} + g\delta_{t,1}\right], \tag{5.23}$$

where $z > 2$. This choice shares some similarities with Eq. (5.21); however, here we are adding only one triangle to each of a fraction g of the vertices in a z-regular graph. Substituting Eq. (5.23) into Ineq. (5.11) we have, as the condition for cascades to occur (corresponding to $\lambda_+ > 1$),

$$F_1(z^2 - z) - z + g\zeta_c > 0, \tag{5.24}$$

where

$$\begin{aligned}\zeta_c = {} & 2 + F_1(6-4z) + 2{F_1}^2(z-2)^2 \\ & + 2{F_1}^2F_2(z-2)^2 - 2{F_1}^3(z-2)^2,\end{aligned} \tag{5.25}$$

denotes the sum of the terms which introduce clustering into the graph. This expression gives us an insight into how adding clustering (specifically nonverlapping triangles in this case) alters the cascade size. Given a specific z we can determine the qualitative effect of clustering in the following way. First, set the expression on the left-hand side of Ineq. (5.24) equal

to zero and solve for F_1 at $g = 0$. This determines the value of F_1 at the transition to the cascade regime in the nonclustered graph; the well-known result of Watts [144], $F_1 = 1/(z-1)$. Next, substitute that F_1 into Eq. (5.25) and observe the sign of ζ_c. If ζ_c is negative we conclude that introducing clustering will decrease the expected cascade size. If, on the other hand, ζ_c is positive, clustering will increase the cascade size.

The justification for these last two statements follows simply from the fact that if ζ_c constitutes a negative contribution to the expression on the left-hand side of Ineq. (5.24), then increasing g, given that $F_1 = 1/(z-1)$, will break the inequality in Ineq. (5.24) and take us into the regime where cascades do not occur. Alternatively, if ζ_c is shown to be positive, then increasing the parameter g will ensure the inequality holds and cascades do occur at these parameter values. Furthermore, because we add only one triangle to each vertex, the level of clustering will be varied by only a small amount by g. Therefore, this analysis provides a zoomed in view of the region near the threshold. In this way we can be certain that the conditions set in relation to the question posed at the outset are met. This is another advantage of using the language of $p_{s,t}$ graphs to address this question.[5]

In Fig. 5.3 (below) we have plotted ζ_c against z for the three processes described in Section 5.1.2: site percolation, bond percolation and Watts's model. In this last case we have chosen the following parameters: seed fraction $\rho_0 = 0$, and a Gaussian threshold distribution with mean R fitted to $F_1 = 1/(z-1)$, and standard deviation $\sigma = 0.1$.

This figure indicates that clustering will decrease the expected cascade size in both site percolation and bond percolation. In other words, the value of the occupation probability needed for a giant connected component to exist is increased in the presence of clustering. We have already demonstrated in [76] that this is the case for the latter of these two processes; to our knowledge this is the first statement of the corresponding result for site percolation. While these results are not directly applicable to models of the spread of disease, in light of the established connection between SIR epidemics and bond percolation we suggest that they may, nonetheless, be of some considerable interest to researchers in that field. This statement is vindicated by the fact that analogous results have recently been established in a number of epidemiological studies that have shown that clustering can adversely affect the propagation of a disease [10, 52, 87, 102].

Also of interest is the behaviour of ζ_c for Watts's model. As z increases in Fig. 5.3, we see ζ_c vary from negative values for $z \leqslant 3$, through a regime of positivity, and back again to negative values for $z \geqslant 29$. This tells us

5 A similar, though more complicated, analysis is most likely possible with z-regular $\gamma(k, c)$ graphs; however, we have yet to verify this.

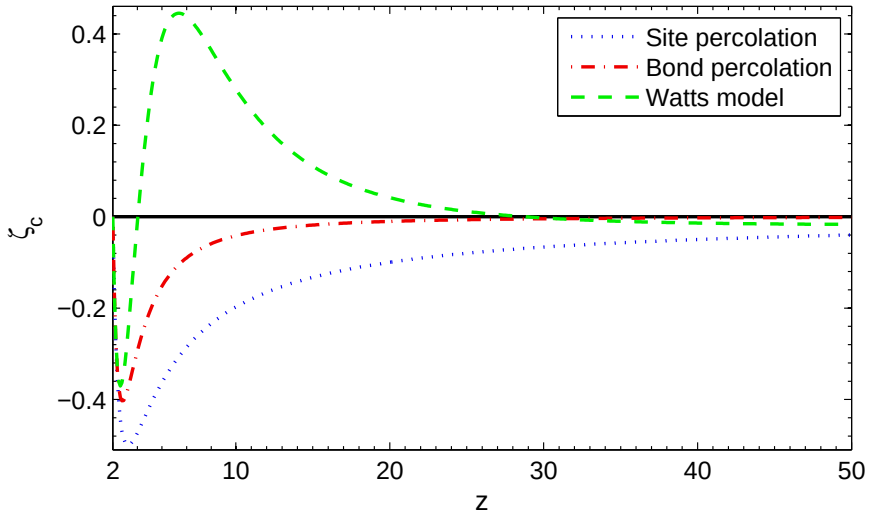


Figure 5.3: Sum of the clustering terms from Ineq. (5.24), ζ_c, vs. mean degree z on $p_{s,t}$ graphs with z-regular degree distribution. Results from site percolation, bond percolation, and Watts's model are shown. As in Section 5.1.2, each process is defined by choosing an appropriate response function. For Watts's model the threshold distribution is Gaussian with standard deviation $\sigma = 0.1$ and mean R, such that $F_1 = 1/(z-1)$. Note, only integer z values are realizable as z-regular graphs.

that for $z \leqslant 3$ the presence of clustering will decrease the left-hand side of Ineq. (5.24) below zero, thereby *decreasing* the expected cascade size; for $3 < z < 29$ clustering will *increase* the expected cascade size; and finally for $z \geqslant 29$ clustering will once more tend to *decrease* the expected cascade size. We note that qualitatively similar results (not reproduced here) are seen for different values of σ, the standard deviation of the thresholds.

By way of validation, in Fig. 5.4 (below) we plot the cascade size ρ against the mean of the threshold distribution R for Watts's model with joint distribution defined by Eq. (5.21), and otherwise the same parameter settings as in Fig. 5.3 (see caption for details). We inferred from Fig. 5.3 that at $z = 3$ cascades become smaller as clustering is increased. This is what we observe in Fig. 5.4(a). Contrastingly, at $z = 5$ cascades should become larger as clustering increases. This is verified by Fig. 5.4(b).

This dependence of the cascade size on the sign of the sum of the clustering terms, ζ_c, in Ineq. (5.24) may be expressed succinctly as a condition on the response function F_2, the probability of activation in the presence of two active neighbours. Specifically, if the value of F_2 at the transition point for cascades in nonclustered z-regular graphs (i.e., F_2 evaluated at the parameters for which $F_1 = 1/(z-1)$) satisfies the condition

$$F_2\Big|_{F_1=\frac{1}{z-1}} > \frac{2z-3}{(z-2)(z-1)}, \tag{5.26}$$

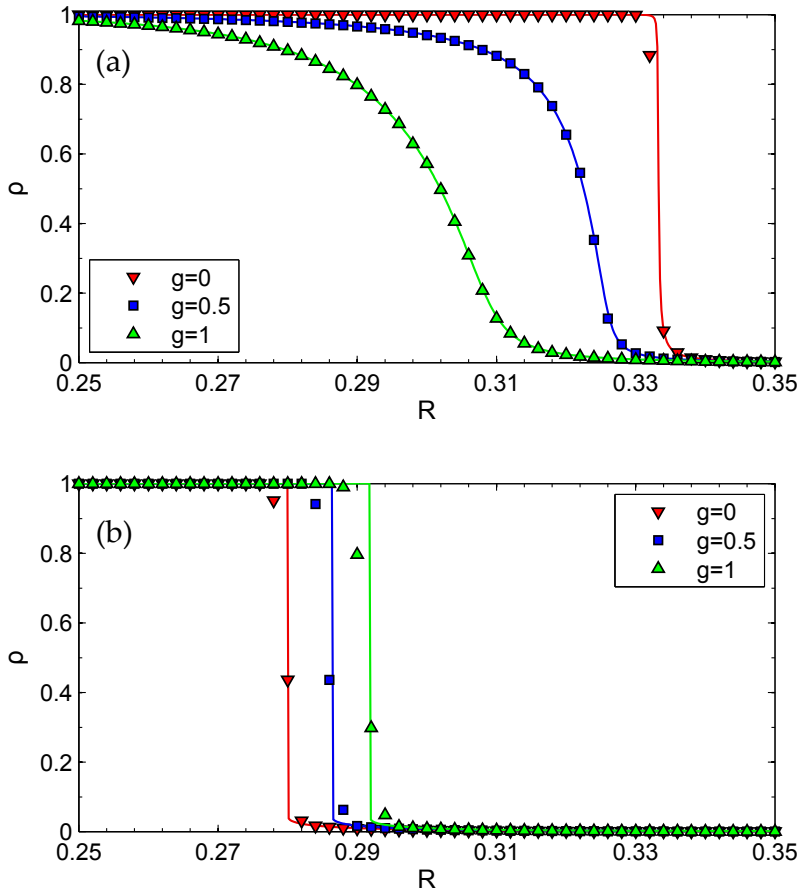


Figure 5.4: Cascade dynamics of Watts's model on graphs of $n = 10^5$ vertices with $p_k = \delta_{k,z}$ and $p_{s,t}$ defined by Eq. (5.21). Gaussian thresholds: mean R and standard deviation $\sigma = 0.1$. Numerical simulations averaged over 100 realizations (symbols) and theory of Section 5.1.1 (lines). Cascade size ρ vs. mean threshold R. (a) $z = 3$: here increasing the level of clustering decreases the expected cascade size at any given R value; (b) $z = 5$: increasing clustering increases the expected cascade size.

then adding triangles will increase the expected size of cascades. Alternatively, if F_2 does not satisfy this inequality, clustering will decrease the expected size of cascades. One may derive this condition by substituting the zero-clustering cascade condition $F_1 = 1/(z-1)$ into Ineq. (5.24) and then solving for F_2. Note that by substituting the respective response functions for site and bond percolation, Eq. (5.12) and Eq. (5.16), into Ineq. (5.26) one may confirm that for $z > 2$ this inequality is not satisfied, and thus that clustering decreases cascade sizes for both of these processes (increases the percolation threshold). Finally, note that Ineq. (5.26) can also be arrived at by a simple counting argument that compares the spread of activations in a clustered random graph to that in a nonclustered random graph. We leave this discussion to Appendix B.1.2.

* * *

This concludes for now our analysis of cascades on edge-triangle graphs. As indicated at the beginning of the chapter, we will pick up this topic again in Section 5.3 as part of a closing discussion of the potential for further development offered by ours and other complementary approaches [91] to synthesising cascade dynamics and high clustering. Our approach consists of two distinct parts, each of which can be thought of as an extended version of the tree-based theory of [73]. Having dealt at length with the first (and simpler) of these extensions in the preceding subsections, we now switch our focus for the duration the next major section of this chapter to the task of modelling cascades on clique-based graphs [71]. This will provide us with our second extension.

5.2 CASCADES ON CLIQUE-BASED GRAPHS

In this section we aim to derive a generalized analytical description of cascade dynamics on the ensemble of graphs defined by Gleeson's γ-theory [71]. Again, the word generalized here means that the approach being sought for will apply to the same broad class of models as discussed at the beginning of Section 5.1, and will include in its scope, at the very least, Watts's model; *k*- core decomposition; and both site and bond percolation.

As was also the case for $p_{s,t}$ graphs, the first problem that we are immediately confronted with in this endeavour is that of reconciling the presence of clustering — which in this case is introduced through a whole spectrum of different-sized cliques — with the tree-based framework of successive activations. Based on our work in Section 5.1.1, we can be hopeful that a congenial solution to this problem will open up the floodgates, so to speak, to the derivation of a series of self-consistent analytical expressions. However, by the same token, we can be certain that the solution provided there for $p_{s,t}$ graphs will be of little use to us here since it depends on the structural motif of nonoverlapping triangles. That is to say, despite the fact that in certain parameter settings (e.g., $\gamma(3,3) = 1$ and $p_{1,1} = 1$) both models will produce the same ensemble, in general clique-based $\gamma(k,c)$ graphs are distinct from edge-triangle graphs, and their cascade dynamics require a unique conceptualization.

In considering how best to proceed, let us return briefly to our review in Section 4.2.2 and remind ourselves of what we know (and do not yet know) about the γ-theory. As regards structural properties, we know that

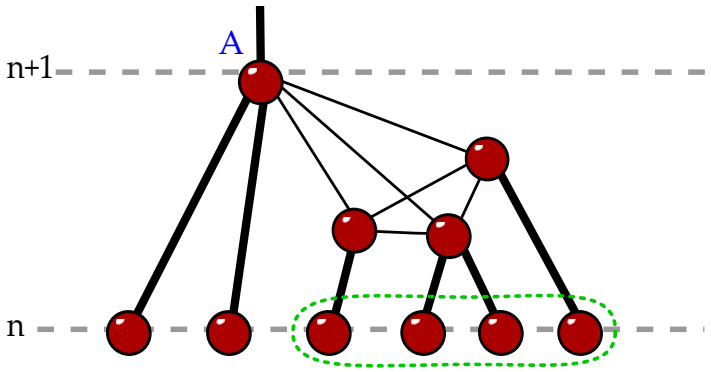


Figure 5.5: Level-by-level cascade propagation in a $\gamma(k, c)$ graph using the tree approximation. External edges emphasized. Reproduced from [71].

the joint distribution $\gamma(k, c)$ prescribes the probability that each vertex has degree k and is a member of a c-clique. We also know that each vertex may be a member of at most one clique, and that there is a distinction between internal edges that connect vertices in the same clique and external edges that connect vertices in different cliques. In terms of dynamics, we have seen that Gleeson [71] has, in fact, already provided an analytical description of cascade propagation on $\gamma(k, c)$ graphs; but that this description is limited strictly to the process of bond percolation (i.e., it is not defined by a response function). One may recall that we sketched the main outline of this approach, while carefully avoiding any discussion of its finer details. By far the most significant detail, which has been purposefully omitted from our initial review, is the precise method by which he was able to apply the concept of child to parent activation to graphs with embedded cliques. We can now reveal how this was achieved.

In Fig. 5.5 we have reproduced Fig. 2 of [71]. This figure shows a portion of an arbitrary $\gamma(k, c)$ graph that has been reconfigured into a tree-like formation. The essential characteristics of this reconfiguration can be explained most succinctly by looking at the local edge topology of the randomly chosen vertex A. This vertex, positioned on level $n + 1$ of the tree, has degree $k = 6$ and is a member of a 4-clique. Its six incident edges are made up of $c - 1 = 3$ internal edges, which connect A to its neighbouring clique members, and $k - c + 1 = 3$ external edges (emphasised). Of these external edges, one connects A to its parent vertex on the next level up, while the remaining $k - c = 2$ connect A to its external children on level n. As regards the location of the clique neighbours, they are positioned on an unlabelled, intermediate level between A and its *grandchildren* (circled in green) on level n. This categorization and positioning of edges is representative of how the tree-based framework operates throughout the graph. Notice that any vertex may be treated similarly to A regardless of the size

of the clique to which it belongs. For any (k, c) pairing such that $k \geqslant c - 1$ (see Section 4.2.2), $c - 1$ clique neighbours can always be made to reside in the interspace between a vertex and the level below; and one may also stipulate in general that at most one external edge leads to the parent above. In extreme cases, a vertex with no internal edges is simply a member of a 1-clique; and, therefore, all of its connections will pass directly from one level to the next ($c - 1 = 0$), as in [73]. A vertex with no external edges must reside either at the root of the tree, and have no parent, if it is part of a clique, or it must be entirely isolated and have zero connections in total.

This, then, was the key that allowed Gleeson to calculate the GCC size, S, in bond percolation on $\gamma(k, c)$ graphs. Equation 5 of [71] was used to determine the conditional probability that a vertex like A is active (part of the GCC) on each level of the tree, and Eq. 6 of [71] (see Eq. (4.16) of Section 4.2.2) then gave S as the probability of activation of the root vertex by using the steady-state value from Eq. 5. The restriction of this theory to bond percolation arises primarily from its reliance on a set of polynomial functions which were defined and tabulated in [109]. Crucially, however, those polynomials play no role in the conceptualization described above. Thus, it turns out that the solution to our problem of understanding how cliques can fit into a tree approximation has already been provided for us in [71]. Our task of generalization, therefore, amounts to taking this set-up and introducing the response function mechanism. However, given the limitations of the equations of [71], this will necessarily involve much more than a series of straightforward substitutions of $F(m, k)$. In fact, as we will now show, it requires a fundamentally new set of equations.

5.2.1 *Cascade Propagation*

With the theoretical foundations now in place, we can begin in earnest to derive analytical expressions for generalized cascade dynamics on $\gamma(k, c)$ graphs. We proceed in the familiar manner by considering (from scratch) the probability, q_{n+1}, that the randomly selected vertex A in Fig. 5.5 is active, conditional on its parent vertex being inactive. As is usual for tree-based propagation, we stipulate that a vertex can become active only by copying the states of the neighbouring vertices directly below it in the tree. In this case, however, the vertex A has two different types of neighbour: it has $k - c$ external children on level n and $c - 1$ clique neighbours on the intermediate level. Significantly, the ways in which these two types of neighbour can become active in their own right are quite distinct from each other. Thus, their contributions to the probability of activation of A must be calculated separately. This is the first problem to be addressed.

Starting with the simpler of the two contributions, let us write down the probability that an arbitrary number, call it j, of A's external neighbours are active. Since there is no clustering between these vertices, each one is independently activated by its own children on level $n-1$ with probability q_n. Therefore, the probability that a total of j out of $k-c$ external neighbours are activated in this way is given simply by the binomial PMF

$$B_j^{k-c}(q_n) = \binom{k-c}{j} {q_n}^j (1-q_n)^{k-c-j}. \tag{5.27}$$

For the second contribution to A, matters are made considerably more complicated by the fact that its $c-1$ clique neighbours are fully connected. This means that the probability that each of these clique neighbours is active depends not only on the states of their children — the four grandchildren of A on level n — but also on the states of one another. Recall from the derivation of our theory for cascades on $p_{s,t}$ graphs in Section 5.1.1 that we had to account for the fact that each vertex at the base of a triangle can directly influence the state of the other. We are faced with a similar problem here; however, since we are now dealing with $\gamma(k,c)$ graphs we have a whole spectrum of clique sizes to contend with. One can appreciate how much more intricate this will make our calculations, by imagining that A were part of a very large clique (as it could be if we were to choose a power-law degree distribution, $p_k \sim k^{-\alpha}$). For example, if A were in a 20-clique, then $c-1 = 19$ intermediate vertices would each have a role to play in determining each others' states. The solution in this case, would require an extensive list of combinatorial expressions, similar to, but extending far beyond, Eqs. (5.5) and (5.6) of Section 5.1.1. However, the sheer number of vertices that may be involved in our calculations is not the only major difficulty that confronts us. Ideally, we would like to avoid tabulating combinatorial terms altogether and instead have a single, compact analytical expression that is flexible enough to deal with any clique size. This expression (behaving as a function) would allow us to feed in the total number of clique neighbours as a variable and would then return the probability that a certain fraction of them are active.

It is not at all obvious that a function can even be defined in the foregoing terms. And, in fact, our demonstration that, indeed, this can be done is of such a distinct character from the rest of our theory as to warrant its own separate subsection. Thus, we continue with our analytical description of cascade propagation by simply providing the name of this function, and taking it for grated that later in Section 5.2.2 we will define precisely how it operates. Let us call the relevant function $R_m^{c-1}(q_n)$, and in doing so refer to it as the probability that in a clique of $c-1$ intermediate vertices a

total of m are active. The dependence on q_n arises from the fact that each intermediate vertex has its own set of children on level n, and each of those children (A's grandchildren in Fig. 5.5) is active with probability q_n.

If we accept the meaning of the label $R_m^{c-1}(q_n)$ and combine it with Eq. (5.27) above, we now have the necessary terms in which to express the contribution of A's external children and clique neighbours towards its probability of activation, q_{n+1}. This takes us very close to defining an iterative equation for q_{n+1} in terms of q_n. The last missing ingredient is the probability, $\Psi_{k,c}$, that the random vertex A, while having degree k and being a member of a c-clique, is also the child of a random vertex on level $n+2$. This probability plays a role similar to that of the term $(k/z)p_k$ in Eq. (3.4) of Section 3.1; which, one may recall, gave the probability of reaching a child of degree p_k by travelling along a randomly chosen edge from its parent in a non-clustered graph. Similarly, here $\Psi_{k,c}$ closes our iteration by allowing us to average over all vertices on level $n+1$ in the correct manner. We express this probability as

$$\Psi_{k,c} = (k-c+1)\gamma(k,c)/z_e, \tag{5.28}$$

where $z_e = \sum_{k',c'}(k'-c'+1)\gamma(k',c')$ is the average number of external edges per vertex.[6]

Combining all three of our ingredients, we can now write our generalized iterative equation, in terms of an arbitrary response function $F_{m+j} \equiv F(m+j,k)$, as

$$q_{n+1} = \rho_0 + (1-\rho_0)\sum_{k,c}\Psi_{k,c}\sum_{j=0}^{k-c}\sum_{m=0}^{c-1} B_j^{k-c} R_m^{c-1} F_{m+j}. \tag{5.29}$$

Thus, we have derived an analytical expression for the probability that a random vertex on the next level up, generically called $n+1$, is active, conditional on its own parent being inactive. Reading this equation from left to right, we see that it is quite easy to interpret its meaning. Referring once again to Fig. 5.5, Eq. (5.29) tells us that the randomly chosen vertex A will be found active if it was initially activated as part of the seed fraction ρ_0, or $(1-\rho_0)$ if it subsequently became activated by copying the states of the neighbours directly below it in the tree. For the latter, there are two distinct contributions from two different sets of neighbours: one from the external children of A, and the other from the intermediate clique members. A total of j of the first type of neighbour are active with probability $B_j^{k-c}(q_n)$, and m of the second type with probability $R_m^{c-1}(q_n)$. (For compactness, these

6 The dashes have been introduced in the definition of z_e to avoid confusion when nesting different sums over k and c inside of each other.

probabilities have been written as B_j^{k-c} and R_m^{c-1}, respectively.) Whether the sum of m and j is sufficient to activate A is determined (as always) by how we define the response function $F(m+j,k)$.

In the usual manner, iterating Eq. (5.29) to the steady-state will give us q_∞. This value can then be substituted into B_j^{k-c+1} and R_m^{c-1} (again both shortened) in the following expression to determine the probability of activation of the root vertex:

$$\rho = \rho_0 + (1-\rho_0)\sum_{k,c}\gamma(k,c)\sum_{j=0}^{k-c+1}\sum_{m=0}^{c-1}B_j^{k-c+1}R_m^{c-1}F_{m+j}. \tag{5.30}$$

The probability ρ is, of course, equivalent to the expected fractional cascade size (see the discussion in Section 3.1). The differences between this equation and Eq. (5.29) above are all attributable to the fact that the root vertex has no parent. The sum over j extends to $k-c+1$ because all of the root's external edges connect downwards to its children, and the correct term for averaging over k and c is simply $\gamma(k,c)$ since the root's activation probability is not conditional on any parent.

Taken together, then, Eqs. (5.29) and (5.30) constitute the core of our new analytical approach. Using these it should be possible (*i*) to derive a generalized cascade condition, and (*ii*) to investigate various cascading processes by applying an appropriately defined response function. We will only consider the second of these tasks, and not in any great detail since the main ideas behind the various response function definitions have already been discussed at length in Section 5.1.2.[7] The appropriate versions of $F(m+j,k)$ will be summarized later in Section 5.2.3; where we will also verify our approach against the results of numerical simulations of bond percolation and Watts's model (see Figs. 5.7 and 5.8 below). Before that, there remains for us, in the following subsection, the not so trivial task of deriving an analytical expression for $R_m^{c-1}(q_n)$.

5.2.2 *Active Clique Neighbours*

Backtracking now slightly in the flow of our presentation, we will offer over the course of the next few pages a series of arguments that will lead ultimately to our derivation of a concise closed-form expression for the probability labelled above as $R_m^{c-1}(q_n)$. Let us begin by recapitulating the precise meaning of this label. According to our earlier definition, it is the probability that m out of $c-1$ intermediate level c-clique vertices are active, given that their own externally linked children are each independently ac-

7 In addition, one may note that if we were to derive a cascade condition, the technique used would be similar to the perturbative method of Section 5.1.1.1.

tive with probability q_n. In Fig. 5.5, for example, $R^3_m(q_n)$ is the probability that m of the vertex A's three clique neighbours are active, given each of the four grandchildren of A (circled) has an activation probability of q_n.

In considering how to calculate $R^{c-1}_m(q_n)$ in general, we see immediately that it is not the states of the external grandchildren that will cause us difficulty; but, rather, the fact that the state of each intermediate clique member can influence the states of all other members. In the framework of [71], every c-clique has one of its (internally linked) members designated as the parent and placed on level $n+1$. This leaves each of the remaining $c-1$ clique members on the intermediate level with $k-c+1$ external edges to connect to its own children on level n. The probability that some number, j, of these children are active is given by the binomial PMF $B^{k-c+1}_j(q_n)$. Thus, the probability that an intermediate clique member is activated by its children is quite easy to calculate. On the other hand, in order to deal with the influence of these $c-1$ clique members on one another, we will have to consider carefully the various combinations of states that may exist within the intermediate portion of the clique. However, given that these states are conditional on the parent vertex in the c-clique being inactive, this intermediate portion may be treated as clique of size $c-1$ in its own right. Our main problem, therefore, is to figure out how to count every possible active-vertex configuration in a $(c-1)$-clique, for arbitrary c.

Our first step in tackling this problem is to provide a mechanism for the intermediate clique members to be activated, which combines both internal and external influences. We define

$$G(d, c-2) = \sum_{k'} \frac{\gamma(k', c)}{\sum_{k''} \gamma(k'', c)} \sum_{j=0}^{k'-c+1} B^{k'-c+1}_j(q_n) F(d+j, k'), \qquad (5.31)$$

for $c \geqslant 2$ and $d \leqslant c-2$, as the conditional probability that an intermediate c-clique vertex will be activated if d of its $c-2$ clique neighbours on the same level are active, given its external children are each active with probability q_n. The term $\gamma(k', c)/\sum_{k''} \gamma(k'', c)$ in this expression is the degree distribution of vertices that belong to a c-clique.[8] And, the response function $F(d+j, k')$ will determine whether d neighbours plus j children are enough to cause activation. Defined as such, $G(d, c-2)$, provides us with a fundamental term in which to express the various possible active configurations, thus permitting us to begin the procedure of counting.

We consider first the simplest non-trivial case, namely $c = 3$. Suppose we pick, from some arbitrary $\gamma(k, c)$ graph, a vertex with degree k, that is also a member of a 3-clique. If, following [71], we let this vertex reside on level $n+1$ of a tree, and also position its $c-1 = 2$ clique neighbours

8 See footnote [6] above on p. 87.

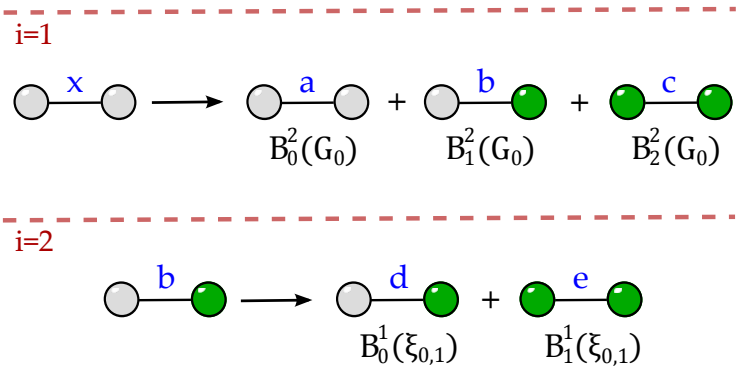


Figure 5.6: Transition probabilities for a pair ($c-1=2$) of intermediate clique neighbours in a $\gamma(k,c)$ graph, expressed in terms of the binomial PMF from Eq. (5.27). Colour indicates state: grey: inactive; green: active.

between level $n+1$ and level n below, our task then is to calculate $R_m^2(q_n)$. To do this, let us refer to Fig. 5.6, and look at the possible states of these two vertices in isolation from their inactive parent.

Starting with both vertices inactive — the configuration labelled x in Fig. 5.6 — we first count the possible configurations of states after one round ($i=1$) of synchronous updating. Since we have started from x, with both vertices inactive, the probability of either vertex becoming active in this first round is simply $G(0,1)$. Therefore, each possible outcome (a, b, or c in Fig. 5.6) is determined by a binomial PMF with probability of success $G(0,1)$. Configuration a, in which both vertices have remained inactive, will occur with probability $B_0^2(G_0)$. Similarly, configuration b, in which one vertex has been activated and the other has remained inactive, will occur with probability $B_1^2(G_0)$. Finally, configuration c, in which both vertices have been activated, will occur with probability $B_2^2(G_0)$. (Note that in each term $G_d \equiv G(d,c-1)$; we will use this abbreviation throughout.)

Having determined the three distinct outcomes of the first round of updates, we will now categorize each configuration into either of two types: *terminal* or *volatile*. In a terminal configuration no further changes of state are possible because all vertices have reached their own steady-state of either permanent activation or inactivation. In a volatile configuration, on the other hand, there exists at least one inactive vertex that is liable to become active. Thus, as long as volatile configurations are produced we must continue with another round of synchronous updates. The process of updating will reach its end when all configurations are terminal. Categorizing the outcomes of round one tells us whether or not a second round is necessary, and, also, indicates which configurations need to be updated. Configuration a is obviously terminal, since the transition from x to a has established that neither vertex can activate while the other remains inactive. Similarly c is also terminal for the simple reason that we do not allow active

vertices to revert to being inactive. Configuration b, however, is volatile, since the transition from x to b has shown us that one of these vertices can activate without the other first being active, but that the same is not true of this other vertex. That is to say, we know that the inactive vertex in b cannot activate without an active neighbour. What is not clear from b is whether the vertex that did activate in round one is now sufficient to activate the vertex that remained inactive in that round. The only way to determine this is to run a second round ($i = 2$) of updates on b.

As was the case in the first round, to begin the second round we must provide an appropriate probability of activation. We want to know if the active vertex in b is enough to activate the inactive vertex in b, given that the inactive vertex cannot activate without an active neighbour. This can be decided upon by using the activation probability $\xi_{0,1} = (G_1 - G_0)/(1 - G_0)$. The configurations produced by updating with this probability are, once again, given by a binomial PMF. With probability $B_0^1(\xi_{0,1})$ the inactive vertex will remain inactive, thereby producing configuration d. Conversely, with probability $B_1^1(\xi_{0,1})$ the inactive vertex will activate, thereby producing configuration e. Categorizing d and e, we find both configurations are terminal, and therefore the process of updating may now cease.

With all terminal configurations now achieved, the next step in our derivation of $R_m^2(q_n)$ will be to combine the various transition probabilities listed in Fig. 5.6, and use them to calculate each of $R_0^2(q_n)$, $R_1^2(q_n)$, and $R_2^2(q_n)$. Tracing our way through Fig. 5.6, we reach a terminal state in which no vertices are active by following the route $x \to a$. Similarly, we end with one active vertex by following $x \to b \to d$. Finally, a terminal state with two active vertices is given by either of the routes $x \to c$ or $x \to b \to e$. All of this information can be expressed succinctly using the various transition probabilities associated with each route, if we bear in mind that a transition from one configuration to another, symbolized by $\to$, corresponds to the multiplication of probabilities, and also that the word *or* corresponds to addition. To summarize, the set of routes described here yields the following set of equations:

$$\begin{aligned} R_0^2(q_n) &= B_0^2(G_0), \\ &= (1 - G_0)^2. \end{aligned} \tag{5.32}$$

$$\begin{aligned} R_1^2(q_n) &= B_1^2(G_0)B_0^1(\xi_{0,1}), \\ &= 2G_0(1 - G_1). \end{aligned} \tag{5.33}$$

$$\begin{aligned} R_2^2(q_n) &= B_1^2(G_0)B_1^1(\xi_{0,1}) + B_2^2(G_0), \\ &= 2G_0(G_1 - G_0) + {G_0}^2. \end{aligned} \tag{5.34}$$

The final step towards our goal of writing a closed-form expression for $R_m^2(q_n)$ is, obviously, to find a way of expressing Eqs. (5.32)-(5.34) as the outputs of a single function which has been given the inputs $m = 0$, $m = 1$, and $m = 2$, respectively. There may be a number of different ways of defining such a function; some of which may appear more elegant than others. For our own part, we can offer a particularly concise definition by introducing a new variable, and considering how the various combinations of G_d in Eqs. (5.32)-(5.34) can be produced in a parsimonious manner.

Our new variable is called l_i. We define it as the number of new activations in round i of synchronous updates. In the scheme presented above we had two rounds; therefore, we also define the pair $l = (l_1, l_2)$ as the sequence of new activations over both rounds. This allows us to represent all possible routes through the configurations of Fig. 5.6 as a collection of ordered pairs. For example, $l = (1, 0)$ means that there is one activation in round $i = 1$ and no activations in round $i = 2$, and therefore corresponds to the route $x \to b \to d$. (Similarly, $l = (1, 1)$ corresponds to $x \to b \to e$.) By applying this notation we have determined (through observation) that the following equation will reproduce each of the Eqs. (5.32)-(5.34) above:

$$R_m^2(q_n) = \sum_{l_1+l_2=m} \binom{2}{l_1, l_2} \frac{\left[1 - G_{l_1}\right]^{2-(l_1+l_2)}}{\left[2 - (l_1 + l_2)\right]!} {G_0}^{l_1} (G_{l_1} - G_0)^{l_2}. \tag{5.35}$$

There are two points worthy of note concerning this equation. The first is that $\sum_{l_1+l_2=m}$ means one must sum over all pairs $l = (l_1 + l_2)$ such that $l_1 + l_2 = m$, where m is the total number of activations. The second point relates to the appearance, directly after this summation, of the multinomial coefficient $\binom{2}{l_1, l_2}$. For those who are unfamiliar with multinomial coefficients, the general form of this term for an arbitrary number of vertices, v, is defined as

$$\binom{v}{l_1, l_2, \ldots, l_v} = \binom{l_1}{l_1} \binom{l_1 + l_2}{l_2} \cdots \binom{l_1 + l_2 + \ldots + l_v}{l_v}. \tag{5.36}$$

In Eq. (5.35), therefore, $\binom{2}{l_1, l_2}$ acts as a neat way of writing the multiplication of binomial coefficients $\binom{l_1}{l_1}\binom{l_1+l_2}{l_2}$ (see [61, 149] and references therein).

To see how Eq. (5.35) operates let us calculate $R_1^2(q_n)$ by setting $m = 1$. The set of all l-pairs that add up to this value of m is $l \in \{(0, 1), (1, 0)\}$. Substituting each of these pairs in turn into the right hand side of Eq. (5.35) and then summing gives $R_1^2(q_n) = [0 + 2G_0(1 - G_1)]$, thereby reproducing Eq. (5.33) above. The values of $R_0^2(q_n)$ and $R_2^2(q_n)$ are found, sim-

ilarly, by using the parameters $m = 0$ and $l = (0,0)$, and $m = 2$ and $l \in \{(0,2),(1,1),(2,0)\}$, respectively.

Thus, in Eq. (5.35) we have found an expression for $R^2_m(q_n)$ — which, we remind ourselves once more, is the probability that m of the two intermediate vertices in a 3-clique are active, given that each of their own children are active with probability q_n. Recall, however, that our ultimate goal is to provide a general expression for $R^{c-1}_m(q_n)$. In the remaining few paragraphs of this section we will summarize very briefly how this expression can be defined. Our technique of finding $R^{c-1}_m(q_n)$, to put it bluntly, is to simply determine a series of expressions for increasing values of $c-1$, and then to try to uncover a unifying pattern between them. In other words, we first find a series of expressions for $R^3_m(q_n)$, $R^4_m(q_n)$, etc. And, following that, we determine, by means of various notational devices, how to write these expressions as a single function.

We spare the reader the details of the derivations involved in this procedure. Each of our individual expressions for $R^{c-1}_m(q_n)$, where $c > 3$, can be found by a method similar to the one described above for $R^2_m(q_n)$. The core of this method is the same regardless of the value of c, and can be summarized in general as follows:

i) Synchronously update the states of all inactive vertices.

ii) Categorize the resulting configurations of states as either terminal or volatile, removing those that are terminal from further consideration.

iii) Repeat steps (*i*) and (*ii*) until no volatile configurations remain.

For example, in determining $R^3_m(q_n)$, the application of these three steps reveals every possible active configuration in a triangle of connected vertices, and each associated transition probability. The full scheme of activations for this case is illustrated in Fig. B.2 of Appendix B.2.1. As above, following the different routes in this figure towards each terminal configuration indicates the correct sequence of multiplications and additions to employ to calculate the values of $R^3_m(q_n)$ for $0 \leqslant m \leqslant 3$. Doing these calculations yields the following set of equations:

$$\begin{aligned} R^3_0(q_n) &= B^3_0(G_0), \\ &= (1-G_0)^3. \end{aligned} \tag{5.37}$$

$$\begin{aligned} R^3_1(q_n) &= B^3_1(G_0)B^2_0(\xi_{0,1}), \\ &= 3G_0(1-G_1)^2. \end{aligned} \tag{5.38}$$

$$\begin{aligned} R_2^3(q_n) &= B_2^3(G_0)B_0^1(\xi_{0,2}) + B_2^3(G_0)B_1^2(\xi_{0,1})B_0^1(\xi_{1,2}), \\ &= 3{G_0}^2(1-G_2) + 6G_0(G_1-G_0)(1-G_2). \end{aligned} \tag{5.39}$$

$$\begin{aligned} R_3^3(q_n) &= B_3^3(G_0) + B_1^3(G_0)B_2^2(\xi_{0,1}) + B_2^3(G_0)B_1^1(\xi_{0,2}) \\ &\quad + B_1^3(G_0)B_1^2(\xi_{0,1})B_1^1(\xi_{1,2}), \\ &= {G_0}^3 + 3G_0(G_1-G_0)^2 + 3{G_0}^2(G_2-G_0) \\ &\quad + 6G_0(G_1-G_0)(G_2-G_1). \end{aligned} \tag{5.40}$$

Note, for these equations we have extended our earlier definition of $\xi_{0,1}$ to the create the function $\xi_{a,b} = (G_b - G_a)/(1-G_a)$.

A generalized expression for $R_m^3(q_n)$, which contains Eqs. (5.37)-(5.40) as special cases can be defined by a similar method to that used for Eq. (5.35). By applying the variable l_i, and considering the sequences of activations, $l = (l_1, l_2, l_3)$, associated with each route through Fig. B.2 we have found that the equation

$$\begin{aligned} R_m^3(q_n) = \sum_{l_1+l_2+l_3=m} \binom{3}{l_1, l_2, l_3} \frac{\left[1-G_{l_1+l_2}\right]^{3-(l_1+l_2+l_3)}}{\left[3-(l_1+l_2+l_3)\right]!} {G_0}^{l_1} \\ \times (G_{l_1}-G_0)^{l_2}(G_{l_1+l_2}-G_{l_1})^{l_3}, \end{aligned} \tag{5.41}$$

will produce expressions in G_d in agreement with Eqs. (5.37)-(5.40).

Observe the striking similarities between equation Eq. (5.41) and Eq. (5.35). They indicate that to create an expression for $R_m^3(q_n)$ from that for $R_m^2(q_n)$, above, all one must do (besides set $c-1=3$) is place extra indices, l_2 and l_3, in appropriate positions, and include one more multiplicative term, namely $(G_{l_1+l_2}-G_{l_1})^{l_3}$. By running the entire scheme of categorization and route counting over again with $c-1=4$ and $l=(l_1,l_2,l_3,l_4)$, we have observed (in calculations not reproduced here) that a similar relationship also holds between $R_m^3(q_n)$ and $R_m^4(q_n)$. And so on for higher c values, in a matryoshka-like sequence.

The pattern of similarities detected in our calculations strongly suggests the following form for a single unifying expression for $R_m^\nu(q_n)$, where ν is arbitrary:

$$R_m^\nu(q_n) = \sum_{l_1+\ldots+l_\nu=m} \binom{\nu}{l_1,\ldots,l_\nu} \frac{\left[1-\sum_{i=1}^{\nu}\theta_i\right]^{\nu-\sum_{i=1}^{\nu} l_i}}{\left[\nu-\sum_{i=1}^{\nu} l_i\right]!} \prod_{i=1}^{\nu} {\theta_i}^{l_i}. \tag{5.42}$$

The variable θ_i in this equation is defined recursively as $\theta_i = G_{S_l} - S_\theta$ for $i \geqslant 2$, with initial value $\theta_1 = G_0$. The terms S_l and S_θ denote the partial sums $\sum_{j=1}^{i-1} l_j$ and $\sum_{j=1}^{i-1}\theta_j$, respectively.

Finally, we observe that Eq. (5.42), may be rewritten, using multi-index[9] notation, in the remarkably concise form

$$R_m^v(q_n) = \sum_{|l|=m} \binom{v}{l} \frac{(1-|\theta|)^{v-|l|}}{(v-|l|)!} \theta^l, \tag{5.43}$$

where, following the conventional usage of this notation, the multi-indices $l = (l_1, l_2, \ldots, l_v)$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_v)$ are ordered v-tuples of l_i and θ_i, for $1 \leqslant i \leqslant v$. The summations over l and θ are defined, respectively, as $|l| = l_1 + \ldots + l_v$ and $|\theta| = \theta_1 + \ldots + \theta_v$, and exponentiation of θ by l means $\theta^l = {\theta_1}^{l_1} {\theta_2}^{l_2} \ldots {\theta_v}^{l_v}$.

By setting $v = c - 1$ in Eq. (5.43) we have the function $R_m^{c-1}(q_n)$, expressed in closed form. Applying this definition of $R_m^{c-1}(q_n)$ in Eqs. (5.29) and (5.30) above completes our analytical description of cascades on clique-based graphs, and permits us to proceed with the task of verifying of our approach. This will be done in the next section by comparing predicted values of the expected cascade size from Eq. (5.30) against the results of numerical simulations of bond percolation and Watts's model.

5.2.3 *Response Functions*

To test the theory of the previous two sections we require an appropriate set of definitions for the response function $F(m + j, k)$ corresponding to the processes in our familiar *broad class* (see Section 5.1). The function F, however, is the same one that we have used throughout our presentation. We started in Chapter 3 by writing it in its simplest generalized form: $F(m, k)$ (see Eq. (3.3)). There, it defined the probability that a k-degree vertex in a locally tree-like graph may be activated by m active neighbours. In Section 5.1, $F(m, s + 2t)$ gave the probability that a k-degree vertex in an edge-triangle graph may be activated by m active neighbours, where $k = s + 2t$. In this section, $F(m + j, k)$ prescribes the probability that a k-degree vertex in a clique-based graph may be activated by $m + j$ active neighbours, where j and m are the numbers of external and internal neighbours, respectively. Since F has not changed (only its arguments have), the same justifications of our use of the response function mechanism as were given in Section 5.1.2 apply equally here. Furthermore, the same method of substitution of arguments into F also applies. Therefore, similarly

9 Multi-index notation is little more an aesthetic tool. It is typically used to simplify formulae in multivariable calculus and distribution theory, by representing an ordered n-tuple of indices as an integer index on which a standard set of operations are defined (see for example [65, 127]). The l_i indices in Eq. (5.42) are ideally suited to this representation.

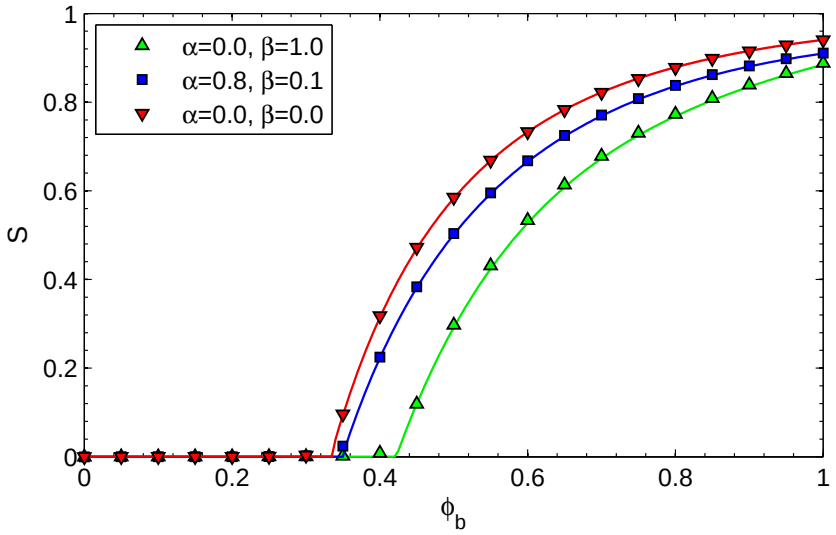


Figure 5.7: Bond percolation on $\gamma(k,c)$ graphs with $n = 10^5$ vertices and Poisson degree distribution p_k, mean degree $z = 3$. Numerical simulations (symbols) averaged over 100 realisations and theory of Section 5.2.1 (lines). GCC size S vs. bond occupation probability ϕ_b. Colour indicates values of α and β used to create the joint distribution $\gamma(k,c)$, and also the level of clustering: $C_2 = 0$ red; $C_2 = 0.31$ blue; $C_2 = 0.35$ green.

to Section 5.1.2, the definitions of $F(m+j,k)$ for different processes are found by setting $m = m+j$ in the definitions of $F(m,k)$ given in [70].

With this aspect clarified, we can begin testing our theory against numerical simulations of various processes. We consider first the process of uniform bond percolation. Setting $m = m+j$ in the right-hand side of Eq. (5.16) defines $F(m+j,k)$ for this process. Applying that definition in the respective $\rho_0 \to 0$ limits of Eqs. (5.29) and (5.30) allows us to use these two equations to calculate the expected GCC size, S. We choose not to reproduce here the simplified forms of Eqs. (5.29) and (5.30) after these substitutions have been applied. Instead, let us go directly to our results.

In Fig. 5.7 we have plotted our calculations of S from Eq. (5.30) against the results of numerical simulations on γ-theory graphs (see caption). The parameters chosen for this figure are the same as those used in Fig. 3(a) of [71]. Each graph has a Poisson degree distribution $p_k = z^k e^{-z}/k!$, with mean degree $z = 3$. Following [71], we set $\gamma(k,c) = \big[(1-\alpha-\beta)\delta_{c,1} + \alpha\delta_{c,3} + \beta\delta_{c,4}\big]p_k$ for $k \geqslant 3$, where $\alpha, \beta \in [0,1]$. In this way we create nonzero clustering by assigning a fraction α of k-degree vertices to 3-cliques and a fraction β to 4-cliques. Additionally, since a 2-degree vertex cannot belong to a clique of size $c > 3$, we assign a fraction α of these vertices to 3-cliques using $\gamma(2,c) = \big[(1-\alpha)\delta_{c,1} + \alpha\delta_{c,3}\big]p_2$. We let vertices of degree zero or one belong to 1-cliques: $\gamma(k,c) = p_k\delta_{c,1}$. By varying α and β different levels of clustering can be prescribed. Again following [71], we use three (α,β) pairs: $(0,0)$, $(0.8,0.1)$, and $(0,1)$. Obviously, $(0,0)$ produces a nonclustered

graph (red). We can use Eq. (4.15) of Section 4.2.2 to define the global clustering coefficient $C_2 = \sum_k p_k c_k$ (see Section 2.1). From this one may show that $(0.8, 0.1)$ produces a clustered graph with $C_2 = 0.31$ (blue), and also that $(0, 1)$ gives a graph with $C_2 = 0.35$ (green). The match obtained between theory and numerics in Fig. 5.7 is excellent, and thereby provides a validation of our generalized approach in the case of bond percolation.[10]

Furthermore, because we have chosen the same parameters as Fig. 3(a) of [71], the results shown in that figure should correspond exactly with the results shown here in Fig. 5.7. Comparing these two figures will reveal to the reader that they do indeed match. This illustrates that our approach contains within its scope the ability to produce the same predicted values of S as the theory of [71]. Note, however, that we have not attempted to explicitly reproduce the equations of [71] from our Eqs. (5.29) and (5.30). As noted earlier at the beginning of Section 5.2, Gleeson's equations depend on a set of polynomial functions defined and tabulated in [109]. These polynomials limit the application of his equations to bond percolation. They also make his equations different to ours. Specifically, the polynomials of [109] give the probability that a randomly chosen vertex in a damaged (i.e., by bond percolation) c-clique belongs to a connected cluster of a certain number of vertices. It is unclear whether this feature can be mapped directly to our generalized framework; although, it appears the correspondence must reside somewhere in the combination of $R_m^{c-1}(q_n)$ and $F(m+j, k)$.

The advantage of our approach over that of [71] is, of course, its supposed applicability to other processes besides bond percolation. To confirm that it really does posses this power we consider for our second, and final, test Watts's model [144]. In Fig. 5.8 (below) we present values of the expected cascade size ρ from Eq. (5.30) plotted against the results of numerical simulations on γ-theory graphs. The thresholds in each of these graphs are drawn from a Guassian distribution: $q(r) = N(R, 0.1)$ (see caption). Therefore, the response function for our theory is defined by setting $m = m + j$ in Eq. (5.20). The structural variables used are the same as those given for Fig. 5.7. All graphs have Poisson p_k with $z = 3$, and $\gamma(k, c)$ is defined by the same three equations as above. The match between theory and numerics in Fig. 5.8 is excellent, once again validating our approach.

This draws to an end all we have wished to say concerning the topic of verification. We take it that the results of Figs. 5.7 and 5.8 have illustrated to ample effect that our approach is extremely accurate, and that the response function mechanism provides us with the flexibility to model a range of

10 One may note that, like Fig. 5.2 earlier, the percolation thresholds in Fig. 5.7 are shifted to the right by increasing the level of clustering. However, the restrictions on what we can infer from this observation are the same as those applied to Fig. 5.2. In particular, we cannot rule out the effects of degree-degree correlations.

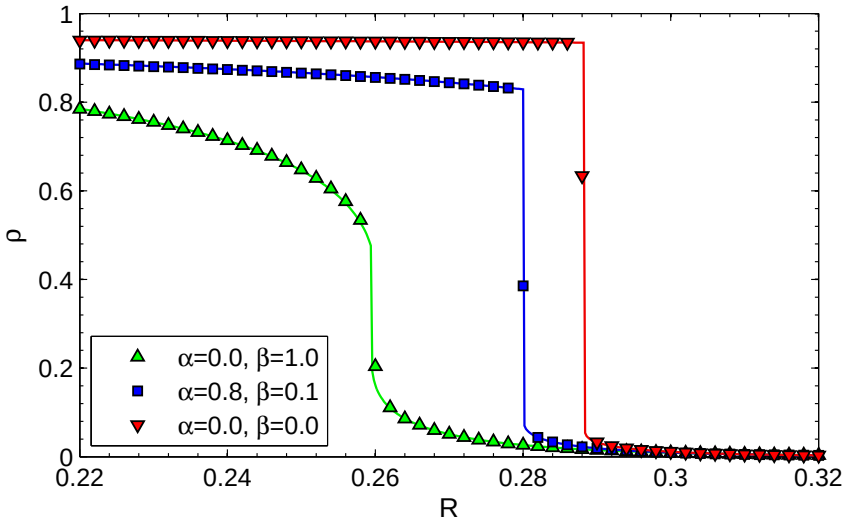


Figure 5.8: Cascade dynamics of Watts's model on $\gamma(k, c)$ graphs with $n = 10^6$ vertices and Poisson degree distribution p_k, mean degree $z = 3$. Gaussian thresholds: mean R and standard deviation $\sigma = 0.1$. Numerical simulations averaged over 100 realizations (symbols) and theory of Section 5.2.1 (lines). Cascade size ρ vs. mean threshold R. Colour indicates values of α and β used to create the joint distribution $\gamma(k, c)$, and also the level of clustering: $C_2 = 0$ red; $C_2 = 0.31$ blue; $C_2 = 0.35$ green.

processes. Based on this evidence, we posit that site percolation and *k*-core decomposition may also be modelled by defining the appropriate response functions (see Section 5.1.2).

In the next, and final, section of this chapter we will provide a brief overview of the approaches to modelling cascades presented here and in Section 5.1. We will also offer some concluding remarks concerning the potential for further development contained within the work of these two sections and some recent work by Karrer and Newman [91].

5.3 TOWARDS A UNIFIED FRAMEWORK

In this chapter we have provided analytical methods for modelling the cascade dynamics of a broad, but precisely defined, class of processes on each of a pair of highly clustered random graph ensembles [71, 111]. Both of these methods have been derived as extensions of the tree-based theory of cascade propagation first put forth in [73]. In each case, the kernel of our approach has been the way in which we have reconciled conceptually the locally tree-like approximation with the explicit presence of clustering in the graphs under consideration. Ostensibly, these two concepts are diametrically opposed; however, the distinctive features of each type of graph have enabled us to subsume clustering within a tree structure and, following [73], to treat the cascade dynamics as a process of successive

activations from level to level, towards the root. In recalling our review of Section 4.3, we are reminded that, for Newman's edge-triangle model, the enabling feature in question is the fact that in his graphs clustering exists solely through the motif of nonoverlapping triangles. Similarly, the vital feature of Gleeson's clique-based model in this respect is the fact that in his graphs a vertex can belong to at most one clique.

If, however, our approaches depend on these two fundamentally unrealistic aspects of Newman's and Gleeson's models, it raises a question concerning the significance of the insights that we can provide. Setting aside the inherent limitations of the idealized processes that our theories can model. We would like to be able to say that the structures to which our theories apply contain at least some degree of realism. The ultimate goal of work like ours is to create an analytical theory that offers highly accurate predictions for cascade dynamics on real-world complex networks. There are many aspects underlying the complexity of these networks; perhaps more aspects than we can ever hope to fully describe from within a mathematical framework. Based on this understanding, we try to build our theories from the bottom up by stripping away as much information as possible and considering instead of real networks, ensembles of undirected and unweighted random graphs. However, even at this degree of abstraction it is not clear which graphical structures should be accounted for, and which others are redundant. For example, the degree distribution is the cornerstone of most sensible first-order theories. To achieve the next level of complexity one can include degree-degree correlations and/or clustering. But, from there it is uncertain whether there are other structural features that should be included to achieve a further level, and if so, how. In the work presented here we have focused on the problem of accounting for clustering for the practical reason that unlike degree-correlated graphs analytical models of dynamics on highly clustered graphs are severely lacking from the networks literature. We also believe that this problem offers a broad prospect for further development, and that, judging by the consensus of empirical observations, clustering is a genuine hallmark of complexity across many domains. Thus, even though models like Newman's and Gleeson's may contain unrealistic features, the mere fact that they include clustering places these models, and also theories of dynamics derived from them, at the forefront of the current state of knowledge.

Concerning the way forward, it appears that Gleeson's model is better suited as a point of departure for the goal we have in mind. This is in virtue of the fact that in a clique-based graph the spectrum of degree-dependent clustering (see Section 2.1) can be fitted to values observed from real networks. There are also no upper bounds on the conventional

clustering measures C_1 and C_2. In contrast, in an edge-triangle graph the level of clustering achievable does have strict upper bounds (see Fig. 4.3 and [76]), and it is generally not possible to match empirical clustering spectra using only nonoverlapping triangles.

However, Newman's model appears to provide the most natural language with which to address questions related to the effects of clustering on cascades. By the addition of nonoverlapping triangles we can increase the level of clustering in an edge-triangle graph in a very precise way, and in small increments each time. This allows us to pinpoint certain effects, such as changes in the position of the cascade threshold (see Section 5.1.3). We suggest that further work in this direction should seek to clarify fully the interplay between clustering and degree-degree correlations (see [76] and [101]) in determining such effects. A similar analysis should be possible from within the clique-based model; however, is it not clear that this would provide any greater insight into the true effects of these properties in the real-world than the corresponding analysis on an edge-triangle graph.

This latter point highlights another important fact about the current state of knowledge concerning network structure. There are many different ways to include clustering in a graph. Two of the most elegant procedures, in terms of analytical tractability, are those provided in [71] and [111]. At certain parameter settings the ensembles of graphs produced by these models may correspond; however, for the most part these ensembles are distinct from each other. This means that each model describes, in a sense, its own universe. Certainly, neither of those universes corresponds to reality. What is missing it seems from each model is the fundamental aspect of arbitrariness. Expressed in the terms in which it actually exists in the real-world, clustering refers to the presence of densely connected subgraphs. Evidently, there are no a priori restrictions on the characteristics of these subgraphs. For example, for most real networks only a small percentage of the subgraphs in their corresponding graphs will be cliques. And so, there appears to be limited scope in these two models for achieving the ideal of a universal model of clustering. Crucially, however, this does not mean that their analysis, and in particular their individual use as a basis for descriptions of dynamics (such as we have provided), will be rendered meaningless should such a structural model be defined. On the contrary, we suggest that our work concerning cascade dynamics [83, 85] may bear fundamental relevance to such a model. In justification of this claim we highlight, finally, the recent paper of Karrer and Newman [91].

In [91] Karrer and Newman have introduced a general formalism for creating ensembles of random graphs that contain arbitrary distributions of subgraphs. The most remarkable feature of their approach is that in

theory the subgraphs in question can be of any desired type. Hence, cliques of various sizes and motifs such as nonoverlapping triangles, are a small subset in the range of possible choices. In particular, the edge-triangle model of [111] is contained within this new class of models as a special case. In fact, it appears likely that many other models are also special cases. In this way, [91] presents a great leap forward towards a universal framework for modelling clustering. What is not clear, however, is how one should select subgraphs such that the random graphs created may mimic in a parsimonious manner the structure of a given real-world network.

We find it particularly interesting that every ensemble generated by the formalism of [91] contains graphs that, like those of [71, 111], can be treated as being tree-like. As the authors of [91] have shown, if we consider arbitrary subgraphs and single edges as the two fundamental units of a graph ensemble, then in each graph realization "local neighbourhoods are tree-like at level of these units" [91]. This suggests the possibility of further extending the tree-based approach of [73] to model cascades on some, or perhaps all, of the ensembles described by this formalism. We have already provided an extension to one of these ensembles in Section 5.1. Finally, it may be possible to generalize our result of Section 5.2.2 for the probability that a clique of arbitrary size contains a certain number of active vertices, into a corresponding result (or set of results) for arbitrary subgraphs.

6

SUMMARY AND CONCLUSIONS

We began by introducing the idea of complexity, and by providing a few examples of its manifestation in cascading phenomena. We also offered a minor critique of certain contemporary connotations of this concept by showing that it is, in fact, deeply rooted in the history of Western intellectual thought. In its most sensible usage the epithet complex applies to any system whose global behaviours are not reducible to a simple average or sum of the individual properties of its components. When expressed in these terms, it becomes apparent that complexity is an increasingly pervasive feature of the modern world. This renders imperative the task of determining precisely why it comes about and how it can be modelled. By its very nature, the study of complexity requires much more than the narrow set of techniques specific to any one field of scientific inquiry. Success, in these terms, is predicated on the knitting together of a diverse array of tools and concepts. This is borne out by the recent prominence of network theory among the different branches of complexity science.

Network theory refers to a broadly interdisciplinary approach to understanding the structural and dynamical characteristics of the networked architectures on which many complex systems are built. It combines aspects of fields ranging in diversity from condensed matter physics to sociometry into a framework based on the theory or random graphs. The core principle of this approach is that insight into complex global phenomena can be gained by determining the properties of these underlying networks. Thus, it adheres to a fundamentally reductionist philosophy. For our own part, we have sought to contribute towards a mathematical interpretation of cascading phenomena by describing analytically a certain class of idealized processes on random graphs of various structural configurations.

After our review of the historical development of network theory in Chapter 1, we proceeded in Chapter 2 to provide details of some of the most significant advances offered by the contemporary study of networks in terms of modelling both structural properties and dynamics. As regards structure, we reviewed four important models of random graph generation [13, 57, 104, 147]. We saw that while each of these models captures some of the salient features of real-world networks — namely, heterogeneous degree distributions, short average geodesic path lengths, and high levels of clustering — none capture all. We attached particular significance to the

lack of a robust method of creating highly clustered random graphs since this feature is widely deemed indispensable to an accurate characterisation of the patterns of interaction between the components of a complex system. In graph theoretic terms clustering translates as the presence of a large number of (typically small) fully connected subgraphs, and is commonly measured by counting the fraction of vertex triples that form triangles. Given the simplicity of this idea, it appeared to us at odds with the general success of network theory that an ensemble of highly clustered graphs, with broad degree distributions, had not been defined many years ago.

Models of the processes that take place on networks are many and varied. If we accept the consensus that networks are ubiquitous, then certainly there are too many processes to consider at any one time. Therefore, for the second part of Chapter 2 we offered detailed reviews of only two dynamical models. The first, percolation, was presented as a way of modelling the resilience of networks under targeted or random removal of sites (vertices) or bonds (edges). And we discussed some important results [24, 31, 104] relating to the size of the GCC of the graph and the percolation threshold, at which the GCC disappears. The second, Watts's model [144], provides insight into cascading phenomena in society, such as fashions, rumours, and opinions. It offers a numerical procedure for computing the the steady-state fraction of active vertices in a particular type of binary-state dynamics on a random graph. This number is used as an estimate of the expected cascade size (i.e., the steady-state density of participants in the cascade).

And so, Chapter 2, as well as providing the fundamental definitions and concepts used in subsequent chapters, also introduced the two focal points of our thesis: random graphs with clustering, and cascade dynamics.

In Chapter 3 we began our discussion of the modelling of cascades with a review of the analytical approach of Gleeson and Cahalane [73]. Motivated by Watts's model, in [73] Gleeson and Cahalane provided an analytically tractable framework for modelling cascade dynamics on networks. Their approach was derived from work on the zero-temperature RFIM on a Bethe lattice [41]. The authors of [73] showed how this model could be generalized to provide a method of calculating analytically the expected cascade size and the position of the cascade threshold for Watts's model on locally tree-like random graphs. Later, it was shown by Gleeson [70] how the response function mechanism of the approach of [73] facilitates its application to a broad class of problems, which includes, as well as Watts's model, site and bond percolation, k-core decomposition, and SIR contagion dynamics.

The tree-based approach of [73] is the cornerstone of the analyses offered in this thesis. Our objective has been to extend this approach in suitable

ways in order to account for certain aspects of network dynamics and/or structure that lay outside the remit of the original theory.

In our first extension we considered the idea of targeted activation of seed vertices (those who instigate the cascade) in Watts's model, and showed how the equations of [73] can be modified to calculate the expected cascade size and the position of the cascade threshold (Appendix A) when the seed is chosen at random from amongst either all vertices in the graph, or a specific subset of vertices with the highest degrees. From this we investigated the so-called *influentials* hypothesis, whereby it is posited that there are certain individuals in society who drive the propagation of information cascades in this domain (like fashions, rumours, and opinions). Following the approach of Watts and Dodds [146], we compared the expected size of cascades instigated by a seed of influentials to the expected size of those instigated by average members of society. Influenatials were defined as those vertices selected from the high-degree subset. Specifically, we chose vertices that had degrees corresponding to the top 10% of the degree distribution. An average member was any vertex in the graph. Using both numerical simulations of Watts's model and our modified expressions we calculated the expected cascade size ρ on Poisson random graphs, and scale-free networks of various mean degrees z, and for various seed fractions ρ_0. For infinite seeds (i.e., those that scale with the size of the graph n as $n \to \infty$), our analysis showed that at values of z for which both types of seed produced global cascades ($\rho \approx 1$) the expected size of those produced by influential seeds was never considerably greater than the expected size of those produced by average seeds. However, there were other z values at which influentials caused global cascades, while average vertices did not ($\rho \approx 0$). We inferred from the figures produced that it may be possible to replicate the effects induced by targeting influentials simply by picking a larger seed of average vertices. Hence, we devised an heuristic approximation, whereby the seed fraction of average vertices is rescaled according to the ratio between the mean degree of the influential subset and the mean degree of all vertices in the graph. Figures 3.6 and 3.7 confirmed that this approximation is valid, at least in a qualitative sense.

The conclusions drawn from this investigation were addressed towards the subject of mass-marketing. We proposed that instead of engaging in the costly endeavour of tracking down influential individuals, companies may be better served by seeking to activate as many "average" members of the population as possible. However, our results do not refute the intrinsic value of influentials (as defined above) as spreaders of information; rather, they reaffirm it. This is borne out also by our work in Appendix A, where we have looked at the case of single seed activation. Here, targeting an

influential produced consistently larger cascades than those produced by picking a random vertex. Thus, in our view, the truth of the influentials hypothesis is largely a matter of interpretation. We agree with Watts and Dodds that real-world cascades are most likely driven by a critical mass of easily influenced (average) individuals, but depending on the costs involved targeting a small number of high degree vertices (influentials) may be an effective strategy for marketers to consider.

We rounded off our discussion of cascade dynamics on locally tree-like graphs in Section 3.3 by mentioning some of the ways in which the approach of [73] has been further modified. For example, versions applying to degree-correlated graphs and directed graphs have been derived in [70] and [51], respectively. And, we mentioned some recent work of ours (Gleeson, Hurd, Melnik, and Hackett) [74] in which we have modelled default contagions in banking networks as binary-state cascades on a certain augmented class of directed graphs. We also considered in Section 3.3 the broader effectiveness of the tree-based framework for modelling cascades on real-world networks; i.e., networks for which structural data has been measured and compiled into adjacency matrices. We reviewed in detail a recent paper of ours (Melnik, Hackett, Porter, Mucha, and Gleeson) [99] that has provided conditions under which the application of the tree-based approach (specifically, the degree-correlated version [70]) may be expected to give accurate results. In this paper we found that, for a range of processes, the discrepancy, in terms of vertical distance, between predicted values of the expected cascade size from the tree-based theory, and values determined by numerical simulations is strongly correlated with $(L - L_1)/z$. That is to say, the accuracy of the theory, depends on the difference between L, the mean intervertex distance in the original network, and L_1, the mean intervertex distance in the rewired version of the network, divided by the mean degree z. Rewiring refers, here, to running an adjacency matrix through an algorithm (Appendix C) that removes clustering, but preserves degree-degree correlations. Furthermore, we found that the vertical distance between theory and numerics is poorly correlated with the level of clustering in the network. Thus, the results of [99], presented us with the counterintuitive proposition that the tree-based theory may work well in the presence of clustering, provided $(L - L_1)/z$ is small. This implies that clustering does not affect the expected cascade size.

Having reviewed [99], we reached the end of Chapter 3 uncertain whether our planned second extension of the theory of [73], to model cascade dynamics on highly clustered networks, was a worthwhile endeavour. However, in the introduction to Chapter 4 we argued that clustering does affect the expected cascade size (and thereby the accuracy of the tree-based

theory), and that this effect may be detected by observing closely the region near the cascade threshold. Therefore, we proceeded with our extension.

We began by reviewing two methods of generating ensembles of clustered random graphs that have appeared recently in the networks literature: Newman's edge-triangle model [111], and Gleeson's model of clique-based clustering [71]. A comparison of these two models revealed that both create graphs that are more structurally realistic than those created by any of the four models reviewed in Chapter 2. In addition, Gleeson's model has the advantage over Newman's that it allows one to create graphs with clustering spectra that match empirical measurements. In Fig. 4.3 [76], we showed also that the range of clustering achievable in Gleeson's model is significantly broader than in Newman's. This figure corroborated our argument that clustering affects the expected cascade size by showing us that increasing the level of clustering in z-regular graphs, generated by either model, increases the value of the critical bond occupation probability.

Each graph ensemble reviewed in Chapter 4 provided a unique structural foundation on which to build an analytical model of cascades on highly clustered networks. Therefore, in Chapter 5 the task of extending in this way the tree-based theory of [73] branched into two distinct tasks. The first, described in Section 5.1, was to extend the theory of [73] to provide an analytical description of cascades on Newman's graphs. The work presented in this section was published in May of this year [85], and has been cited four times since then [26, 43, 122, 150]. The second, described in Section 5.2, was to extend the theory of [73] to provide an analytical description of cascades on Gleeson's graphs. The work of this section is currently being prepared for publication.

In both sections we demonstrated how motifs of clustered vertices can be included in the conceptual framework of child-to-parent activation that defines the tree-based approach. For edge-triangle graphs this demonstration was quite straightforward (see Fig. 5.1). Since nonoverlapping triangles is the only motif that Newman defines for his graphs, we posited that such a triangle exists whenever an edge connects two vertices on the same level of the tree. From this basis we derived self-consistent equations for the activation probabilities of random vertices on each level that included the interplay of influences between the child vertices in a triangle. In this analysis we derived an analytical expression for the expected cascade size. The corresponding demonstration for Gleeson's graphs of how cliques of various sizes can be included in the framework of child-to-parent activation was first given by Gleeson himself in [71], and we simply reviewed his approach (see Fig. 5.5). This, however, did not make the derivation of an analytical expression for the expected cascade size any easier in this case.

The broad spectrum of clustering motifs defined by Gleeson's model made this task significantly more complicated than the corresponding task had been for the edge-triangle model by virtue of the fact that the interplay of influences between neighbouring vertices in cliques of various sizes had to be accounted for. In Newman's graphs we had only the two children in a 3-clique (triangle) to contend with, for Gleeson's graphs we had the $c-1$ children in a c-clique, where c is arbitrary. Nevertheless, in Section 5.2.2 we presented an extensive series of arguments to show how one may count the number of active vertices in a clique of arbitrary size. This lead, ultimately, to our derivation in Eq. (5.43) of a concise closed-form expression for the probability, R_m^{c-1}, that m of the $c-1$ children in a c-clique are active. Earlier, in Section 5.2.1 we had written an iterative equation for the level-by-level activation probabilities in a clique-based graph, and also an equation for the expected cascade size, before knowing how to express one of the key components of each of these equations, namely R_m^{c-1}, analytically. Equation (5.43) completed the definitions of both equations.

Thus, in Sections 5.1 and 5.2 we have extended the tree-based theory of [73] to model cascade dynamics on two classes of highly clustered graphs. In both cases, our expressions for level-by-level activation probabilities and the expected cascade size provide us with the necessary tools to model a broad range of processes (see above). As Gleeson [70] has demonstrated, this is facilitated through the response function mechanism.

We validated our results for edge-triangle graphs by comparing our calculations of the expected GCC size in site percolation and the steady-state active fraction in Watts's model to the corresponding values determined by numerical simulations. We validated our results for clique-based graphs in a similar manner but with bond percolation and Watts's model. For both types of graphs our analytical calculations provided extremely accurate matches to the numerical output. Additionally, in the case of edge-triangle graphs we used our expressions to determine a cascade condition in terms of arbitrary response functions $F(m, s+2t)$. Therefore, this condition (Ineq. (5.11)) contains as special cases expressions for the threshold values in every process within our broad range.

Finally, in Section 5.1.3 we conducted a close analysis of the effects of clustering on cascades. We demonstrated that on edge-triangle graphs with z-regular (every vertex has degree z) degree distributions clustering will increase the percolation threshold in both site and bond percolation for all values $z > 2$ (see Fig. 5.3 and Ineq. (5.26)). We suggested that these results may be of considerable interest to epidemiologists since, as Newman [108] has shown, the steady-state infected fraction in SIR dynamics on a random graph can be mapped onto the GCC size in bond percolation. Furthermore,

results analogous to ours have been established in a number of recent network-based epidemiological studies that have shown that clustering can negatively impact the spread of disease [10, 52, 87, 102]. In relation to Watts's model, we showed that the effects of clustering may vary. For $z \leqslant 3$ adding clustering to the graph will decrease the expected cascade size ρ, for $3 < z < 29$ clustering will increase ρ, and for $z \geqslant 29$ clustering will once again decrease ρ. In Fig. 5.4 we confirmed these results at $z = 3$ and $z = 5$. This insight into the idiosyncrasies of the cascade dynamics of Watts's model may bear direct relevance to studies of the spread of behaviour in human populations, such as [25].

* * *

Some suggestions for future research relating to the analytical approaches of Chapter 5 and the clustering formalism of [91] have been given in Section 5.3. Adding to these, we suggest that it is desirable to broaden the range of processes we can model to include those — such as SIS — that exhibit non-monotone binary-state dynamics. In [72] Gleeson has provided master equations for some of these processes on locally tree-like graphs. The combination of our work in this thesis (including that presented in Section 3.2) with these master equations and the ensembles of clustered graphs defined by [91] represents an intriguing possibility.

A

OTHER ASPECTS OF INFLUENTIALS THEORY

In this appendix we demonstrate the application of the theory of Section 3.2 to other more technical questions related to the influentials hypothesis of information dynamics on social networks. We encourage the reader to review Section 3.2 before proceeding as extensive reference is made to the material presented there throughout the following discussion.

A.1 CRITICAL SEED FRACTION

Returning to Fig. 3.4 of Section 3.2.1, if instead of reading the cascade size ρ as a function of the mean degree z, we fix on a specific z and consider the change in ρ as the seed fraction ρ_0 increases, we see that at certain z values there is a discontinuous transition from $\rho \approx 0$ to $\rho \approx 1$. To illustrate this point, let us extract a slice from Fig. 3.4 at $z = 8$ and compare ρ against ρ_0. The result is shown below in Fig. A.1. With regards to this plot, observe that for each seed type (average and influential) the transition to the global cascade regime occurs at a critical value of ρ_0, which we denote as $\widehat{\rho_0}$. For both parameter settings ($\tau = 1$ and $\tau = 0.1$), we have calculated this number and marked its position on the ρ_0 axis.

What is the significance of $\widehat{\rho_0}$? For a fixed graph topology it represents the minimum seed fraction necessary to instigate a global cascade. According to our theory, if the seed is smaller than $\widehat{\rho_0}$ global cascades are not likely to occur; and if it is larger global cascades are likely to occur. When generalized to include varying τ, this concept is of significant interest to those concerned with the effectiveness of marketing strategies, as it provides a theoretical insight into not only the type but also the relative number of people that should be targeted.

In this section we show how, for either type of seed, this critical value of ρ_0 can be approximated using analytical techniques. Our first step is to rewrite the extended iterative equation (3.18) as

$$\lambda(q) = \frac{\hat{z}}{z}\rho_0 + \zeta(q) - \rho_0\hat{\zeta}(q), \tag{A.1}$$

where

$$\zeta(q) = \sum_{k=1}^{\infty} \frac{k}{z} p_k \sum_{l=0}^{k-1} \widetilde{C}_l q^l, \tag{A.2}$$

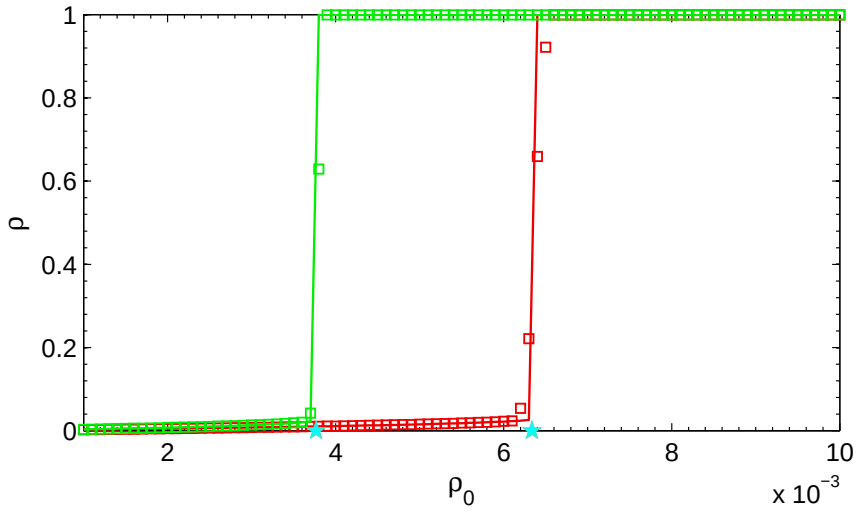


Figure A.1: Cascade dynamics of Watts's model on a PRG with $n = 10^6$, $z = 8$, and $R = 0.18$. Numerical simulations (squares) averaged over 100 realisations and extended tree-based theory (lines). Final active density ρ vs. seed fraction ρ_0. Colour indicates seed type: red: $\tau = 1$; green: $\tau = 0.1$. Critical seed fractions, $\widehat{\rho_0} = 3.768 \times 10^{-3}$ for $\tau = 0.1$ and $\widehat{\rho_0} = 6.336 \times 10^{-3}$ for $\tau = 1$, marked with cyan pentagrams.

$$\hat{\zeta}(q) = \frac{\alpha k^\star p_{k^\star}}{\tau z} \sum_{l=0}^{k-1} \widetilde{C_l} q^l + \frac{1}{\tau} \sum_{k \geqslant k^\star+1}^{\infty} \frac{k}{z} p_k \sum_{l=0}^{k-1} \widetilde{C_l} q^l, \tag{A.3}$$

and

$$\widetilde{C_l} = (-1)^l \binom{k-1}{l} \sum_{m=0}^{l} (-1)^m \binom{l}{m} F(m,k). \tag{A.4}$$

For convenience we have dropped the n and $n+1$ subscripts from q; so that the function $\lambda(q)$ now represents q_{n+1}. And, we have introduced $\widetilde{C_l}$ in order to make simpler the differentiation of $\zeta(q)$ and $\hat{\zeta}(q)$. We will see why this is necessary shortly.

The key insight that suggested the possibility of an analytical approximation was the realisation that the critical value of ρ_0 occurs when $\lambda(q) = q$ has a double root; that is, when its discriminant $\Delta = 0$. This emerged from the analysis of cobweb plots of the function $\lambda(q)$. One such plot is shown below in Fig. A.2. This figure illustrates the behaviour of $\lambda(q)$ at $\rho_0 = 10^{-3}$ for Watts's model on a Poisson random graph with the same parameters as Fig. A.1 (see caption). The size of a cascade depends on where the function $\lambda(q)$ first hits the diagonal (grey). If it hits close to $q = 0$ a global cascade does not occur; however, if it does not hit the diagonal here it will continue to grow until it reaches a value close to $q = 1$, in which case a global cascade will occur. Significantly, it is the value of ρ_0 that determines the shape of $\lambda(q)$ around $q = 0$, and thereby determines where it will first cross the diagonal. As can be seen in the zoomed-in version, Fig. A.2(b), for

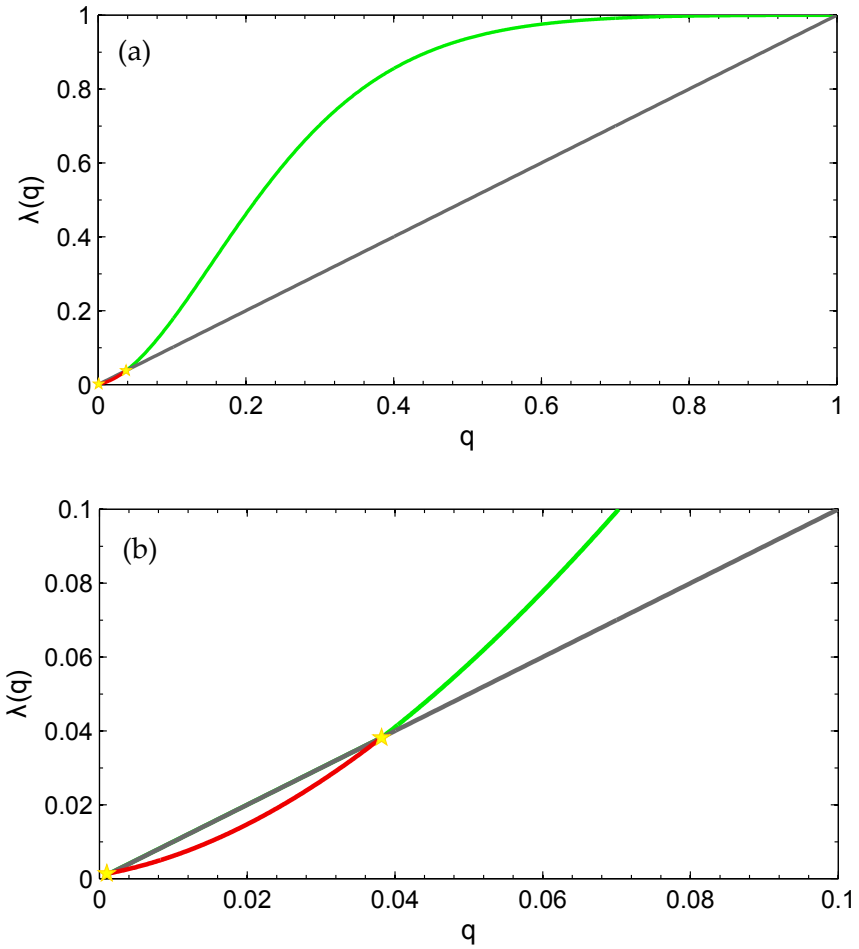


Figure A.2: Cobweb plot of $\lambda(q)$ at $\rho_0 = 10^{-3}$, for Watts's model on a PRG with $n = 10^6$, $z = 8$, and $R = 0.18$. Colour indicates the behaviour of $\lambda(q)$: red: $\lambda(q) < q$; green: $\lambda(q) > q$. Points where $\lambda(q) = q$ are marked with yellow pentagrams. In (a) $q, \lambda(q) \in [0, 1]$, in (b) $q, \lambda(q) \in [0, 0.1]$.

$\rho_0 = 10^{-3}$ the function $\lambda(q)$ dips below the diagonal very close to $q = 0$ (values marked in red) and therefore a global cascade does not occur in this case, which tallies with Fig. 3.4. In our analysis we have observed that the trajectory of $\lambda(q)$ over the interval $0 \leqslant q \leqslant 1$ is always qualitatively similar to the one shown in Fig. A.2(a). The differences are usually discernible only by looking closely at the $q \leqslant 0.1$ region. For example, we know from Fig. 3.4 that with $\rho_0 = 10^{-2}$ a global cascade should occur at $z = 8$, and we would recognize this in a cobweb plot (if we were to draw one) by the fact $\lambda(q)$ would not fall below q, as it does in Fig. A.2(b).

Note, we do not attempt to find an exact expression for the double root of $\lambda(q) = q$; instead, we approximate this value using Taylor series. In doing this we take the third order Taylor polynomials of $\zeta(q)$ and $\hat{\zeta}(q)$ about the point $q = 0$:

$$\zeta(q) = q\zeta'(0) + \frac{q^2}{2}\zeta''(0) + \frac{q^3}{6}\zeta'''(0), \tag{A.5}$$

and

$$\hat{\zeta}(q) = q\hat{\zeta}'(0) + \frac{q^2}{2}\hat{\zeta}''(0) + \frac{q^3}{6}\hat{\zeta}'''(0). \tag{A.6}$$

Substituting Eqs. (A.5) and (A.6) into $\lambda(q) = q$ gives us a third order polynomial in q, with ρ_0 and the derivatives of $\zeta(q)$ and $\hat{\zeta}(q)$ appearing in the coefficients:

$$\Big(\frac{\zeta'''(0) - \rho_0\hat{\zeta}'''(0)}{6}\Big)q^3 + \Big(\frac{\zeta''(0) - \rho_0\hat{\zeta}''(0)}{2}\Big)q^2 + (\zeta'(0) - \rho_0\hat{\zeta}'(0) - 1)q + \frac{\hat{z}}{z}\rho_0 = 0. \tag{A.7}$$

Letting Δ denote the discriminant of Eq. (A.7), our critical value $\widehat{\rho_0}$ is found by solving $\Delta = 0$ for ρ_0. This task is made easier by the introduction of the $\widetilde{C_l}$ function, but is still sufficiently complicated for us to skip the details here. Suffice to say that given the relevant parameters, p_k, n, R, τ and z, we can solve $\Delta = 0$ using a combination of pen and paper calculation and computer-based iterative solvers. For example, in Fig. A.1 we made use of the built-in '*fzero*' function from MATLAB R2010b® to find that $\widehat{\rho_0} = 3.768 \times 10^{-3}$ for $\tau = 0.1$, and $\widehat{\rho_0} = 6.336 \times 10^{-3}$ for $\tau = 1$. As one can see, these are very accurate approximations of the true transition points.[1]

Finally, note that Fig. A.1 lends further credence to our assertion of Section 3.2.2 that the effect of influentials on cascade dynamics in Watts's model can be accurately replicated by renormalizing the seed; i.e., by choosing a greater number of average degree vertices to initiate the spreading dynamics. Notice, the green line ($\tau = 0.1$) is essentially similar to the red line ($\tau = 1$) only with a lower $\widehat{\rho_0}$. Thus, this figure tells us, and marketing executives, that on a PRG with the given parameters — which is obviously a trivialized example for real-world applications — one is likely to initiate a cascade by turning on a fraction of the population greater than 3.768×10^{-3} consisting entirely of influentials, or, failing that, simply by increasing to a fraction 6.336×10^{-3} consisting of average degree people. We have observed qualitatively similar pictures to Fig. A.1 (not reproduced here) for other choices of z, and also for power law p_k; therefore, this interpretation is quite robust.

1 A similar analysis can be applied also to the power-law distributed graphs of Section 3.2.1.

A.2 SINGLE SEED ADJUSTMENT

In [144] Watts considered a seed consisting a single vertex of average degree, and showed how in this setting the existence of global cascades depends on the relative size of the vulnerable cluster, S_v, and that of the extended vulnerable cluster, S_e. From our review in Section 2.3.2.1 we know that the vulnerable cluster consists of those vertices that require only a single neighbour to be active in order for them to join in the cascade themselves. The extended vulnerable cluster is a superset of the vulnerable cluster consisting of all vulnerable vertices, plus any of their immediately adjacent neighbours. From this definition, it is clear that a single seed cannot be the spark that ignites a cascade unless it is a member of the extended vulnerable cluster. In any given realization the probability that we pick such a vertex as our seed is S_e. Thus here our calculation of the expected size of an ensuing cascade, which we usually call ρ, requires an adjustment to reflect this restriction on where the seed is placed. An intuitive, and as Gleeson [70] has shown effective, way to factor in this adjustment is to simply multiply these probabilities. This gives us the following approximation for the expected size of a cascade instigated by a single seed of average degree:

$$S \approx \rho S_e. \tag{A.8}$$

The generalization of Eq. (A.8) to seed fractions larger than $\rho_0 = 1/n$ was expressed in [70] as

$$S \approx \rho\left[1 - (1 - S_e)^{\lfloor n\rho_0 \rfloor}\right], \tag{A.9}$$

where $\lfloor \cdot \rfloor$ is the floor function. This version of the approximation applies to any ρ_0 that does not scale with the number of vertices in the graph, n, as $n \to \infty$. For the figures shown in Section 3.2 $\rho_0 \to \infty$ with n, so that Eq. (A.9) reduces to ρ.

In [70] Gleeson also derived an analytical expression for the size, S_e, of the extended vulnerable cluster. Watts had calculated this value by numerical simulations in [144], whereas in [70] in was shown that

$$S_e = \sum_{k=1}^{\infty} p_k\left[1 - (1 - q_\infty)^k\right], \tag{A.10}$$

where q_∞ is the steady state of the iteration

$$q_{n+1} = \sum_{k=1}^{\infty} \frac{k}{z} p_k\left[1 - (1 - q_n)^{k-1}\right] F(1, k). \tag{A.11}$$

By substituting Eq. (A.10) into Eq. (A.8) it is possible to accurately approximate S (see Fig. 5 of [70]). However, this adjustment factor applies only when the seed vertex is chosen at random ($\tau = 1$). For our extended theory, in which we account for the targeting of vertices of specific degree ($\tau \in [0, 1]$), we must modify Eq. (A.10) slightly. In calculating S_e we are now no longer interested merely in the probability p_k that a vertex has degree k but rather the probability that an initially active vertex has degree k. We call this probability S_k, it is given by:[2]

$$S_k = \begin{cases} 0, & \text{if} \quad k < k^\star, \\ \alpha p_k/\tau, & \text{if} \quad k = k^\star, \\ p_k/\tau, & \text{if} \quad k > k^\star. \end{cases} \tag{A.12}$$

And so, Eq. (A.10) now becomes

$$S_e = \sum_{k=1}^{\infty} S_k \left[1 - (1 - q_\infty)^k\right]. \tag{A.13}$$

Figures A.3 and A.4 illustrate the application of the approximation of S given by Eq. (A.8) with Eq. (A.13) used to calculate S_e. Figure A.3 shows the result of average and influential single seed activation in Watts's model on Poisson random graphs; Fig. A.4 shows the same but on scale-free networks created in a similar manner those analysed in Fig. 3.5 of Section 3.2.1 (see captions for details).

As regards the influentials hypothesis, these figures provide a unique and interesting insight. They demonstrate that in this particular setting, where one is not afforded the freedom to simply increase the relative size of the seed, an influential may be significantly more successful than an average vertex in terms of the extent of the cascade that it may trigger. While in either figure the bounds on the window of global cascades is the same for both types of seed, inside each of these windows influentials generally produce larger cascades. Thus, Figs. A.3 and A.4 show that there is a genuinely intrinsic value which influentials posses over average degree vertices. This does not, however, contradict our thesis that a larger group of regular Janes and Joes could, if mobilized, outdo this one trendsetter.

2 Refer to Section 3.2.1 and [70] for the relevant definitions.

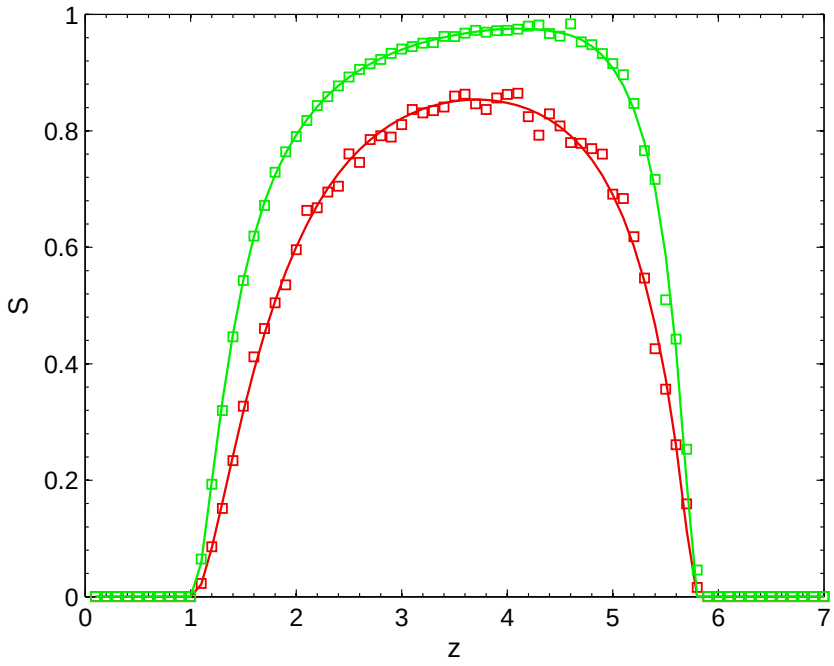


Figure A.3: Cascade dynamics of Watts's model on PRGs with $n = 10^6$ and uniform thresholds, $R = 0.18$. Numerical simulations (squares) averaged over 10^3 realisations and extended tree-based theory (lines). Final active density S vs. mean degree z. Single-seed initially active, $\rho_0 = 1/n$. Colour indicates seed type: red: $\tau = 1$; green: $\tau = 0.1$.

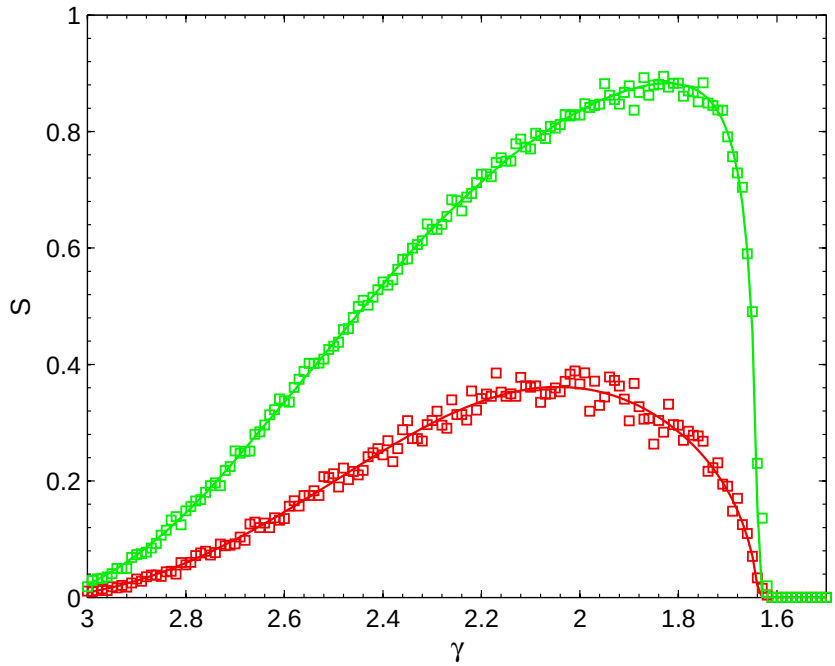


Figure A.4: Cascade dynamics of Watts's model on SFNs with $n = 10^6$ and uniform threshold $R = 0.06$. Numerical simulations (squares) averaged over 10^3 realisations and extended tree-based theory (lines). Final active density S vs. slope γ. Single-seed initially active, $\rho_0 = 1/n$. Colour indicates seed type: red: $\tau = 1$; green: $\tau = 0.1$.

B

FURTHER DETAILS OF A SYNTHESIS

This appendix offers further details of some of the arguments put forth in Chapter 5 concerning cascade dynamics on highly clustered graphs. Section B.1 relates to our approach to modelling cascades on Newman's edge-triangle graphs [111] (see Section 5.1). Section B.2 deals with a certain aspect of our approach to modelling cascades on Gleeson's clique-based graphs [71] (see Section 5.2). (See Chapter 5 for definitions of terms.)

B.1 CONCERNING EDGE-TRIANGLE GRAPHS

B.1.1 *On the Edge-Triangle Cascade Condition*

The cascade condition for $p_{s,t}$ graphs represented by Ineq. (5.11) of Section 5.1.1.1 has its origin in the following simple argument. Returning to Eq. (5.9) we see that the largest eigenvalue of the matrix $\mathbf{A}$ can be expressed in the generalized form

$$\lambda_+ = \frac{p + \sqrt{q}}{2}, \tag{B.1}$$

where $p = A_{11} + A_{22}$ and $q = (A_{11} - A_{22})^2 + 4A_{12}A_{21}$. We have said if $\lambda_+ > 1$, then the vector of activation probabilities $\mathbf{v}$ will diverge from its trivial equilibrium $\mathbf{v} = \mathbf{0}$, taking us into the global cascade regime (see Section 5.1.1.1). From Eq. (B.1) this condition on λ_+ is equivalent to

$$q - (2 - p)^2 > 0. \tag{B.2}$$

In terms of the elements of $\mathbf{A}$, Ineq. (B.2) may be written as

$$-4\det(\mathbf{I} - \mathbf{A}) > 0, \tag{B.3}$$

where $\mathbf{I}$ is the identity matrix and $\det(\mathbf{I} - \mathbf{A}) = (1 - A_{11})(1 - A_{22}) - A_{12}A_{21}$. By substituting the values of the elements of $\mathbf{A}$ from Eq. (5.10) into Ineq. (B.3) and rearranging terms we arrived at Ineq. (5.11).

B.1.2 *Counting Argument for the Effects of Clustering*

Here we give an intuitive argument for the effects of clustering on cascades in z-regular $p_{s,t}$ graphs. This stands as an alternative derivation of the condition on the response function F_2 in Section 5.1.3, see Ineq. (5.26).

We compare the spread of activations from a single active vertex (coloured green in Fig. B.1(a) and B.1(b), below) to two of its neighbours, and then further into the graph. In configuration (a) the three vertices considered do not form a triangle, and up to $2(z-1)$ second neighbours may potentially be activated in this way. In configuration (b), the three vertices do form a triangle, and therefore only $2(z-2)$ second neighbours are available for activation. We proceed to calculate the expected number of edges that may activate second neighbours in each configuration, and derive a condition under which clustering (configuration (b)) gives a greater number of expected activations than the corresponding nonclustered case (configuration (a)). First, we consider configuration (a). Each of the two grey vertices will be activated by the green vertex with probability F_1. If activated, a grey vertex may in turn activate up to $z-1$ of its other neighbours. So we count the expected number of *active edges* (edges that are connected to an active vertex) on the right-hand side of Fig. B.1(a) as $2F_1(z-1)$.

In configuration (b), the two neighbours of the active vertex are also connected to each other, leaving each with $z-2$ edges to other neighbours. These edges may become active edges in one of three ways:

i) Both grey vertices are activated directly by their single active neighbour; this happens with probability ${F_1}^2$, and gives $2(z-2)$ active edges on the right-hand side of Fig. B.1(b).

ii) One grey vertex is activated directly by the active neighbour; the other grey vertex then becomes active because it now has two active neighbours. This happens with probability $2F_1(F_2-F_1)$, and gives $2(z-2)$ active edges.

iii) One grey vertex is activated directly by the active neighbour; the other grey vertex does not activate even though it has two active neighbours. This happens with probability $2F_1(1-F_2)$, and gives $z-2$ active edges.

The expected number of active edges on the right-hand side of Fig. B.1(b) is therefore

$$\begin{aligned} &2{F_1}^2(z-2)+4F_1(F_2-F_1)(z-2)+2F_1(1-F_2)(z-2) \\ &\qquad = 2F_1(z-2)(F_2-F_1+1). \end{aligned} \tag{B.4}$$

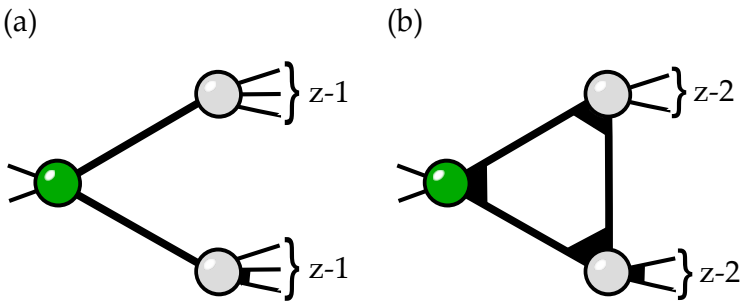


Figure B.1: Spread of activation from a single active vertex (green) to two of its inactive neighbours (grey) in (a) a nonclustered graph, and (b) a $p_{s,t}$ graph with nonzero clustering.

This is greater than the value $2F_1(z-1)$ found for configuration (a) if

$$F_2 - F_1 > \frac{1}{z-2}. \tag{B.5}$$

To examine the effect upon the cascade threshold, we substitute the cascade condition $F_1 = 1/(z-1)$ for the threshold in a nonclustered z- regular graph [144] into Ineq. (B.5)) to obtain the condition given in Ineq. (5.26). If this condition is satisfied, cascade propagation is more likely on the clustered z-regular graph than on the nonclustered version.

B.2 CONCERNING CLIQUE-BASED GRAPHS

B.2.1 *On Active Clique Neighbours*

Figure B.2, below, illustrates the various configurations of states associated with the process of updating and categorization described in Section 5.2.2, when applied to a clique of three intermediate vertices. Refer to the discussion in Section 5.2.2 for the meanings of the different labels and variables used in this figure. By following routes through the configurations shown here we can count every possible way of producing a certain number of permanently active vertices. For example, the process will end with no vertices active if we follow the route $x \to a$. We end with a single active vertex by the route $x \to b \to e$. Two permanently active vertices are given by either of the routes $x \to c \to h$ or $x \to b \to f \to j$. Finally, there are four different routes that each lead to three permanently active vertices: $x \to d$, or $x \to b \to g$, or $x \to c \to i$, or $x \to b \to f \to k$.

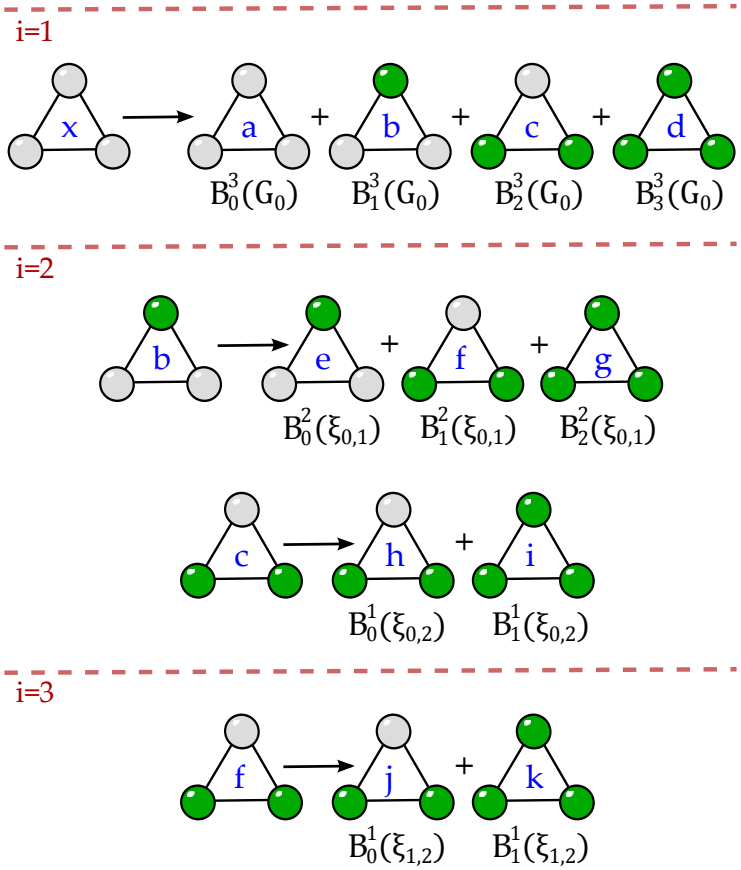


Figure B.2: Transition probabilities for a triple ($c - 1 = 3$) of intermediate clique neighbours in a $\gamma(k, c)$ graph, expressed in terms of the binomial PMF from Eq. (5.27). Colour indicates state: grey: inactive; green: active.

C

SOME NUMERICAL ALGORITHMS

Listing C.1: A MATLAB script for Watts's model

```
% watts_model.m: Watts's model [144] on adjacency matrix A. Here all
    vertices are assigned the same threshold R. Author: J. P. Gleeson.

% Input variables:
% A := adjacency matrix;
% Nrealiz := no. of runs to average over;
% R := value of uniform threshold;
% rho0 := seed fraction;
% Output variable:
% rho := expceted cascade size.

function[rho]=watts_model(A,Nrealiz,R,rho0)

Nnodes=size(A,1);
degree=A*ones(Nnodes,1);
realiz=zeros(Nrealiz,1);
threshold=R*ones(Nnodes,1); % set uniform treshold

    for r=1:Nrealiz
        n=zeros(Nnodes,1);
        randnodes=randperm(Nnodes); % randomly activate seed set
        n(randnodes(1:floor(rho0*Nnodes)))=1;
        Non=sum(n);
        vuln_n=A*n>degree.*threshold; % vulnerable set
        new_n=max(vuln_n,n); % newly active nodes
        % repeat until no new activations
        while max(new_n~=n)>0
              n=new_n;
              Non=sum(n);
              vuln_n=A*n>degree.*threshold;
              new_n=max(vuln_n,n); % assumes always on
        end
        realiz(r)=Non/Nnodes;
    end

rho=sum(realiz)/Nrealiz;
end
```

Listing C.2: A MATLAB script for Newman and Ziff's bond percolation algorithm

```
% newman_ziff_alg.m: Newman and Ziff bond percolation algorithm [116]
    on adjacency matrix A. Adds a specified fraction of edges to A in
    random order. Author: J. P. Gleeson.

% Input variables:
% A := adjacency matrix;
% edgefraction := fraction of edges to add.
% Output variables:
% Nnodes := no. of nodes;
% Q := size of largest cluster after bonds added;
% useNedges := no. of edges added.

function [Q,Nnodes,useNedges]=newman_ziff_algorithm(A,edgefraction)

Nnodes=size(A,1);
[fromvtemp,tovtemp,vals]=find(A);
kv=A*ones(Nnodes,1);
kmax=max(kv);

LUindices=fromvtemp<tovtemp;
fromv=fromvtemp(LUindices);
tov=tovtemp(LUindices);

randorder=randperm(length(fromv));
fromv=fromv(randorder);
tov=tov(randorder);

Nedges=length(fromv);
useNedges=floor(edgefraction*Nedges);
size_cluster=ones(Nnodes,1);
max_cluster_size=1;

Q=zeros(useNedges,1); % holds max cluster size with n edges in play

    for n=1:useNedges
        label_a=fromv(n);
        label_b=tov(n);
        % if label_a=label_b do nothing on this bond
        if label_a~=label_b
           if size_cluster(label_a)<size_cluster(label_b)
                label_min=label_a;
                label_max=label_b;
           else
                label_min=label_b;
                label_max=label_a;
```

```
            end % if
            % move cluster sizes from min to max
            size_cluster(label_max)=size_cluster(label_max)+size_
                cluster(label_min);
            if size_cluster(label_max)>max_cluster_size
                 max_cluster_size=size_cluster(label_max);
            end % if
            size_cluster(label_min)=0;
            % replace min labels in fromv
            indices=find(fromv(n+1:useNedges)==label_min);
            fromv(indices+n)=repmat(label_max,length(indices),1);
            % replace min labels in tov:
            indices=find(tov(n+1:useNedges)==label_min);
            tov(indices+n)=repmat(label_max,length(indices),1);
            % search only from n+1 to end, as no need to relabel
                earlier bonds
        end % if
        Q(n)=max_cluster_size;
    end % for n
end
```

Listing C.3: A MATLAB script for rewiring a clustered random network

```
% rewire_alg.m: Assumes matrix A loaded. Rewires A, removing clustered
     configurations of edges, preserves degree-degree correlations.
    Author: S. Melnik [99].

kc=kmax
kclow=1
kchigh=kmax

% shuffling routine:
N=size(A,1);
[fromv,tov,vals]=find(A);
kv=A*ones(N,1);
kmax=max(kv)
LUindices=fromv<tov;
LUfromv=fromv(LUindices);
LUtov=tov(LUindices);

% check recontructs correctly:
sm=sparse(LUfromv,LUtov,ones(size(LUfromv)),N,N);
A2=sm+sm';
k2=A2*ones(N,1);
z2=mean(k2)
max(k2)
% end reconstruction check
```

```
% now shuffle k-degree stubs:
LUfromv_degrees=kv(LUfromv);
LUtov_degrees=kv(LUtov);
new_LUfromv=[];
new_LUtov=[];

for knum=kclow:kchigh % 1:kc %1:kmax
    indices_fromv_deg_knum=find(LUfromv_degrees==knum);
    indices_tov_deg_knum=find(LUtov_degrees==knum);
    if length(indices_fromv_deg_knum)>0
      LUfromv_nodes=LUfromv(indices_fromv_deg_knum);
      rp=randperm(length(LUfromv_nodes));
      LUfromv_nodes_shuffled=LUfromv_nodes(rp);
      LUfromv(indices_fromv_deg_knum)=LUfromv_nodes_shuffled;
    end % if

    if length(indices_tov_deg_knum)>0
      LUtov_nodes=LUtov(indices_tov_deg_knum);
      rp=randperm(length(LUtov_nodes));
      LUtov_nodes_shuffled=LUtov_nodes(rp);
      LUtov(indices_tov_deg_knum)=LUtov_nodes_shuffled;
    end % if
end % for knum

% check post-shuffle:
sm=sparse(LUfromv,LUtov,ones(size(LUfromv)),N,N);
A3=sm+sm';
k3=A3*ones(N,1);
z3=mean(k3)
max(k3)

% end of shuffle routine
A=A3;
disp('end of shuffle');
```

BIBLIOGRAPHY

[1] Abbate, J. *Inventing the Internet*. MIT Press, Cambridge, MA, 1999.

[2] Abello, J., Buchsbaum, A., and Westbrook, J. A functional approach to external graph algorithms, in *Proceedings of the 6th European Symposium on Algorithms*. Springer, Berlin, 1998.

[3] Achlioptas, D., D'Souza, R. M., and Spencer, J. Explosive percolation in random networks. *Science* ***323***, 1453-1455 (2009).

[4] Aiello, W., Chung, F., and Lu, L. A random graph model for massive graphs, in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*. Association of Computing Machinery, New York, 2000.

[5] Albert, R., and Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* ***74***, 47-97 (2002).

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

[7] Albert, R., Jeong, H., and Barabási, A.-L. Attack and error tolerance of complex networks. *Nature* ***406***, 378-382 (2000).

[8] Amaral, L. A. N., Scala, A., Barthélémy, M., and Stanley, H. E. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* ***97***, 11149-11152 (2000).

[9] Austin, T. L., Fagen, R. E., Penny, W. F., and Riordan, J. The number of components in random linear graphs. *Ann. Math. Stat.* ***30***, 747-754 (1959).

[10] Ball, F., Sirl, D., and Trapman, P. Analysis of a stochastic SIR epidemic on a random network incorporating household structure. *Math. Biosci.* ***224***, 53-73 (2010).

[11] Ball, P. *Critical Mass: How One Thing Leads to Another*. Arrow Books, London, 2004.

[12] Barabási, A.-L. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume, New York, 2003.

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

[14] Biggs, N. L., Lloyd, E. K., and Wilson, R. J. *Graph Theory, 1736-1936*. Oxford University Press, Oxford, 1976.

[15] Bollobás, B. *Random Graphs*, 2nd ed. Cambridge University Press, Cambridge, 2001. Preface.

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

[17] Bollobás, B., and Riordan, O. *Percolation*. Cambridge University Press, Cambridge, 2006, ch. 1, p. 1.

[18] Bornholdt, S., and Schuster, H. G., Eds. *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH, Weinheim, Germany, 2003.

[19] Bringhurst, R. *The Elements of Typographic Style*. Version 2.5. Hartley & Marks Publishers, Point Roberts, WA, 2002.

[20] Britton, T., Deijfen, M., Lindholm, M., and Lagerås, A. N. Epidemics on random graphs with tunable clustering. *J. Appl. Prob.* **45**, 743-756 (2008).

[21] Broadbent, S. R., and Hammersley, J. M. Percolation processes. I. Crystals and mazes. *Proc. Cambridge Philos. Soc.* **53**, 629-641 (1957).

[22] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the Web. *Comput. Netw.* **33**, 309-320 (2000).

[23] Caldarelli, G., and Vespignani, A., Eds. *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science*. World Scientific Publishing, Singapore, 2007.

[24] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. Networks robsutness and fragility: Percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468-5471 (2000).

[25] Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194 (2010).

[26] Chatterjee, A. P. Connectedness percolation in monodisperse rod systems: clustering effects. *J. Phys.: Condens. Matter* **23**, 375101 (2011).

[27] Christakis, N. A., and Fowler, J. H. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little, Brown and Company, New York, 2009.

[28] Chung, F., and Graham, R. *Erdös on Graphs: His Legacy of Unsolved Problems*. A K Peters, New York, 1998.

[29] Chung, F., and Lu, L. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA* **99**, 15879-15882 (2002).

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

[31] Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. Resilience of the Internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626-4628 (2000).

[32] Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. Breakdown of the Internet under intentional attack. *Phys. Rev. Lett.* **86**, 3682-3685 (2001).

[33] Colizza, V., Flammini, A., Maritan, A., and Vespignani, A. Characterization and modeling of protein-protein interaction networks. *Physica A* **352**, 1-27 (2005).

[34] Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110-115 (2006).

[35] Colizza, V., Pastor-Satorras, R., and Vespignani, A. Reaction diffusion processes and metapopulation models in heterogeneous networks. *Nat. Phys.* **3**, 276-282 (2007).

[36] Darwin, C. *On the Origin of Species*. Penguin Classics. Penguin Books, London, 2009.

[37] de Sola Pool, I., and Kochen, M. Contacts and influence. *Soc. Networks* **1**, 1-48 (1978).

[38] de Solla Price, D. J. Networks of scientific papers. *Science* **149**, 510-515 (1965).

[39] de Solla Price, D. J. A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inform. Sci.* **27**, 292-306 (1976).

[40] Dewey, J. *Democracy and Education: An Introduction to the Philosophy of Education*. The Free Press, New York, 1966, ch. 6, pp. 76–77.

[41] Dhar, D., Shukla, P., and Sethna, J. P. Zero-temperature hysteresis in the random-field Ising model on a Bethe lattice. *J. Phys. A: Math. Gen.* ***30***, 5259–5267 (1997).

[42] Dobson, I., Carreras, B. A., Lynch, V. E., and Newman, D. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-orginization. *Chaos* ***17***, 026103 (2007).

[43] Dodds, P. S., Harris, K. D., and Payne, J. L. Direct, physically motivated derivation of triggering probabilities for spreading processes on generalized random networks. Submitted to *Phys. Rev. E*, arXiv:1108.5398v1.

[44] Dodds, P. S., Harris, K. D., and Payne, J. L. Direct, physically motivated derivation of the contagion condition for spreading processes on generalized random networks. *Phys. Rev. E* ***83***, 056122 (2011).

[45] Dodds, P. S., Muhamad, R., and Watts, D. J. An experimental study of search in global social networks. *Science* ***301***, 827-829 (2003).

[46] Dorogovtsev, S. N. *Lectures on Complex Networks*. Oxford University Press, Oxford, 2010.

[47] Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. *k*-core organization of complex networks. *Phys. Rev. Lett.* ***96***, 040601 (2006).

[48] Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* ***80***, 1275-1335 (2008).

[49] Dorogovtsev, S. N., and Mendes, J. F. F. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.

[50] Duch, J., and Arenas, A. Community detection in complex networks using extremal optimization. *Phys. Rev. E* ***72***, 027104 (2005).

[51] Dunne, A. Cascade dynamics on directed networks. Master's thesis, University of Limerick, 2010.

[52] Eames, K. T. D. Modelling disease spread through random and regular contacts in clustered populations. *Theor. Pop. Biol.* ***73***, 104-111 (2008).

[53] Ebel, H., Mielsch, L.-I., and Bornholdt, S. Phys. rev. e. *Scale-free topology of e-mail networks* **66**, 035103 (2002).

[54] Erdös, P. Some remarks on the theory of graphs. *Bull. Amer. Math. Soc.* **53**, 292-294 (1947).

[55] Erdös, P. Graph theory and probability II. *Canad. J. Math.* **13**, 346-352 (1961).

[56] Erdös, P., and Rényi, A. On random graphs I. *Publ. Math. Debreccen* **6**, 290-297 (1959).

[57] Erdös, P., and Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* **5**, 17-61 (1960).

[58] Erdös, P., and Rényi, A. On the evolution of random graphs. *Bull. Inst. Int. Statist. Tokyo* **38**, 343-347 (1961).

[59] Erdös, P., and Rényi, A. On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hungar* **12**, 261-267 (1961).

[60] Euler, L. Solutio problematis ad geometriam situs pertinentis. *Comment. Acad. Sci. Petrop.* **8**, 128-140 (1741).

[61] Evans, M., Hastings, N., and Peacock, B. *Statistical Distributions*, 3rd ed. Wiley, New York, 2000.

[62] Faloutsos, M., Faloutsos, P., and Faloutsos, C. On power-law relationships of the Internet topology. *Comput. Commun. Rev.* **29**, 251-262 (1999).

[63] Ford, G. W., and Uhlenbeck, G. E. Combinatorial problems in the theory of graphs. IV. *Proc. Natl. Acad. Sci. USA* **43**, 163-167 (1957).

[64] Fowler, J. H., and Christakis, N. A. Cooperative behaviour cascades in human social networks. *Proc. Natl. Acad. Sci. USA* **107**, 5334-5338 (2010).

[65] Friedlander, F. G., and Joshi, M. S. *Introduction to the Theory of Distributions*, 2nd ed. Cambridge University Press, Cambridge, 1998, ch. 1, p. 5.

[66] Gell-Mann, M. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Abacus, London, 1995.

[67] Gilbert, E. N. Enumeration of labelled graphs. *Canad. J. Math.* **8**, 405-411 (1956).

[68] Gillies, J., and Cailliau, R. *How the Web Was Born: The Story of the World Wide Web*. Oxford University Press, Oxford, 2000.

[69] Gladwell, M. *The Tipping Point: How Little Things Can Make a Big Difference*. Abacus, London, 2002.

[70] Gleeson, J. P. Cascades on correlated and modular random networks. *Phys. Rev. E* **77**, 046117 (2008).

[71] Gleeson, J. P. Bond percolation on a class of clustered random networks. *Phys. Rev. E* ***80***, 036107 (2009).

[72] Gleeson, J. P. High-accuracy approximation of binary-state dynamics on networks. *Phys. Rev. Lett.* ***107***, 068701 (2011).

[73] Gleeson, J. P., and Cahalane, D. J. Seed size strongly affects cascades on random networks. *Phys. Rev. E* **75**, 056103 (2007).

[74] Gleeson, J. P., Hurd, T. R., Melnik, S., and Hackett, A. Systemic risk in banking networks without Monte Carlo simulation. Submitted to appear as a chapter in *Financial Networks and Risk Assessment*, Springer, New York.

[75] Gleeson, J. P., and Melnik, S. Analytical results for bond percolation and k-core sizes on clustered networks. *Phys. Rev. E* ***80***, 046121 (2009).

[76] Gleeson, J. P., Melnik, S., and Hackett, A. How clustering affects the bond percolation threshold in complex networks. *Phys. Rev. E* ***81***, 066114 (2010).

[77] Goltsev, A. V., Dorogovtsev, S. N., and Mendes, J. F. F. *k*-core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects. *Phys. Rev. E* **73**, 056101 (2006).

[78] Goold, G. P., Ed. *Plutarch's Moralia*, vol. 1. Loeb Classical Library 197. Harvard University Press, Cambridge, MA., 1989, bk. III, 48c.

[79] Goold, G. P., Ed. *Aristotle's Metaphysics*, vol. 1. Loeb Classical Library 271. Harvard University Press, Cambridge, MA., 1989, bk. VIII, 1045a.

[80] Grassberger, P. On the critical behaviour of the general epidemic process and dynamical percolation. *Math. Biosci.* ***63***, 157-172 (1982).

[81] Grossman, J. W. The evolution of the mathematical research collaboration graph. *Congr. Numer.* ***158***, 202-212 (2002).

[82] Guare, J. *Six Degrees of Seperation: A Play*. Vintage Books, New York, 1990.

[83] Hackett, A., and Gleeson, J. P. Cascades on graphs with embedded cliques. In preparation.

[84] Hackett, A., Gleeson, J. P., and Melnik, S. Site percolation in clustered random networks. *Int. J. Comp. Syst. Sci.* ***1***, 25-30 (2011).

[85] Hackett, A., Melnik, S., and Gleeson, J. P. Cascades on a class of clustered random networks. *Phys. Rev. E* ***83***, 056107 (2011).

[86] Harary, F. *Graph Theory*. Addison-Wesley, Reading, MA, 1969.

[87] Hébert-Dufresne, L., Noël, P.-A., Marceau, V., Allard, A., and Dubé, L. J. Propagation dynamics on networks featuring complex topologies. *Phys. Rev. E* ***82***, 036115 (2010).

[88] Hierholzer, C. Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Math. Ann.* ***7***, 30-32 (1873). Published posthumously.

[89] Huberman, B. A. *The Laws of the Web*. MIT Press, Cambridge, MA, 2001.

[90] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* ***407***, 651-654 (2000).

[91] Karrer, B., and Newman, M. E. J. Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E* ***82***, 066118 (2010).

[92] Katz, E., and Lazarsfeld, P. F. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. The Free Press, Glencoe, IL., 1955.

[93] Kleinberg, J. Navigation in a small world. *Nature* ***406***, 845 (2000).

[94] Kleinfeld, J. The small world problem. (Could it be a big world after all? The 'six degrees of separation' myth.). *Society* ***39***, 62 (2002).

[95] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tompkins, A. S., and Upfal, E. The web as a graph, in *Proceedings of the 19th ACM Symposium on Principles of Database Systems*. Association of Computing Machinery, New York, 2000.

[96] Kuperman, M., and Abramson, G. Small world effect in an epidemiological model. *Phys. Rev. Lett.* ***86***, 2909-2912 (2001).

[97] Lazarsfeld, P. F., Berelson, B., and Gaudet, H. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press, New York, 1968.

[98] Maslov, S., Sneppen, K., and Zaliznyak, A. Detection of topological patterns in complex networks: Correlation profile of the Internet. *Physica A* ***333***, 529-540 (2004).

[99] Melnik, S., Hackett, A., Porter, M. A., Mucha, P. J., and Gleeson, J. P. The unreasonable effectiveness of tree-based theory for networks with clustering. *Phys. Rev. E* ***83***, 036112 (2011).

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

[101] Miller, J. C. Percolation and epidemics in random clustered networks. *Phys. Rev. E* ***80***, 020901(R) (2009).

[102] Miller, J. C. Spread of infectious disease through clustered populations. *J. R. Soc. Interface* ***6***, 1121-1134 (2009).

[103] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Networks motifs: Simple building blocks of complex networks. *Science* ***298***, 824-827 (2002).

[104] Molloy, M., and Reed, B. A critical point for random graphs with a given degree sequence. *Random Struct. Algor.* ***6***, 161-179 (1995).

[105] Molloy, M., and Reed, B. The size of the giant component of a random graph with a given degree sequence. *Combin. Prob. Comp.* ***7***, 295-305 (1998).

[106] Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* ***98***, 404-409 (2001).

[107] Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* ***89***, 208701 (2002).

[108] Newman, M. E. J. Spread of epidemic disease on networks. *Phys. Rev. E* ***66***, 016128 (2002).

[109] Newman, M. E. J. Properties of highly clustered networks. *Phys. Rev. E* ***68***, 026121 (2003).

[110] Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* ***45***, 167-256 (2003).

[111] Newman, M. E. J. Random graphs with clustering. *Phys. Rev. Lett.* ***103***, 058701 (2009).

[112] Newman, M. E. J. *Networks: An Introduction*. Oxford University Press, Oxford, 2010.

[113] Newman, M. E. J., Barabási, A.-L., and Watts, D. J. *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, 2006.

[114] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* ***64***, 026118 (2001).

[115] Newman, M. E. J., and Watts, D. J. Renormalization group analysis of the small-world network model. *Phys. Lett. A* ***263***, 341-346 (1999).

[116] Newman, M. E. J., and Ziff, R. M. Fast Monte Carlo algorithm for site or bond percolation. *Phys. Rev. E* ***64***, 016706 (2001).

[117] Ore, O. *Graphs and Their Uses*. Random House, New York, 1963.

[118] Palla, G., Derényi, I., and Vicsek, T. The critical point of k-clique percolation in the Erdös-Rényi graph. *J. Stat. Phys.* ***128***, 219-227 (2007).

[119] Pastor-Satorras, R., and Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* ***86***, 3200-3203 (2001).

[120] Pastor-Satorras, R., and Vespignani, A. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, Cambridge, 2004.

[121] Payne, J. L., Dodds, P. S., and Eppstein, M. J. Information cascades on degree-correlated random networks. *Phys. Rev. E* ***80***, 026125 (2009).

[122] Payne, J. L., Harris, K. D., and Dodds, P. S. Exact solutions for social and biological contagion models on mixed directed and undirected, degree-correlated random networks. *Phys. Rev. E* ***84***, 016110 (2011).

[123] Redner, S. How popular is your paper? An empirical study of the citation distribution. *Euro. Phys. J. B* ***4***, 131-134 (1998).

[124] Riordan, O., and Warnke, L. Explosive percolation is continuous. *Science* ***333***, 322-324 (2011).

[125] Rogers, E. M. *Diffusion of Innovations*. The Free Press, Glencoe, IL., 1962.

[126] Ryle, G. *The Concept of Mind*. Penguin Modern Classics. Penguin Books, London, 2000.

[127] Saint-Raymond, X. *Elementary Introduction to the Theory of Pseudodifferential Operators*. CRC Press, Boca Raton, FL, 1991, ch. 1, p. 2.

[128] Schelling, T. C. Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *J. Conflict Resolut.* ***17***, 381-428 (1973).

[129] Serrano, M. Á., and Boguñá, M. Clustering in complex networks. I. General formalism. *Phys. Rev. E* ***74***, 056114 (2006).

[130] Simon, H. A. On a class of skew distribution functions. *Biometrika* ***42***, 425-440 (1955).

[131] Smith, A. *The Theory of Moral Sentiments*. Penguin Classics. Penguin Books, London, 2010.

[132] Solé, R. V., Pastor-Satorras, R., Smith, E., and Kepler, T. B. A model of large-scale proteome evolution. *Adv. Complex Syst.* **5**, 43-54 (2002).

[133] Solomonoff, R., and Rapoport, A. Connectivity of random nets. *Bull. Math. Biophys.* ***13***, 107-117 (1951).

[134] Stauffer, D., and Aharony, A. *Introduction to Percolation Theory*, 2nd ed. Taylor and Francis, London, 1992.

[135] Thompson, C. Is the tipping point toast? *Fast Company* ***122***, 74-105 (2008).

[136] Trapman, P. On analytical approaches to epidemics on networks. *Theor. Popul. Biol.* ***71***, 160-173 (2007).

[137] Traud, A. L., Kelsic, E. D., Mucha, P. J., and Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* ***53***, 526-543 (2011).

[138] Traud, A. L., Mucha, P. J., and Porter, M. A. Social structure in Facebook networks. Submitted to *Soc. Net.*, arXiv:1102.2166.

[139] Travers, J., and Milgram, S. An experimental study of the small world problem. *Sociometry* ***32***, 425-443 (1969).

[140] Tyler, J. R., Wilkinson, D. M., and Huberman, B. A. Email as spectroscopy: Automated discovery of community structure within organizations, in *Proceeding of the First International Conference on Communities and Technologies*. M. Huysman, E. Wenger, and V. Wulf, Eds. Kluwer, Dordrecht, 2003.

[141] Vázquez, A., and Moreno, Y. Resilience to damage of graphs with degree correlations. *Phys. Rev. E* ***67***, 015101(R) (2003).

[142] Vázquez, A., Pastor-Satorras, R., and Vespignani, A. Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E* **65**, 066130 (2002).

[143] Wagner, A., and Fell, D. The small world inside large metabolic networks. *Proc. R. Soc. London B* **268**, 1803-1810 (2001).

[144] Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci. USA* **99**, 5766-5771 (2002).

[145] Watts, D. J. *Six Degrees: The Science of a Connected Age*. Vintage, London, 2004.

[146] Watts, D. J., and Dodds, P. S. Influentials, networks, and public opinion formation. *J. Consum. Res.* **34**, 441-458 (2007).

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

[148] White, J. G., Southgate, E., Thompson, J. N., and Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis Elegans*. *Phil. Trans. R. Soc. London* **314**, 1-340 (1986).

[149] Wilf, H. S. *Generatingfunctionology*, 3rd ed. A K Peters, Wellesley, MA, 2006.

[150] Yoon, S., Goltsev, A. V., Dorogovtsev, S. N., and Mendes, J. F. F. Belief-propagation algorithm and the Ising model on networks with arbitrary distributions of motifs. Submitted to *Phys. Rev. E*, arXiv:1106.4925v1.

COLOPHON

This thesis was typeset with LaTeX 2ε using Hermann Zapf's *Palatino* and *Euler* type faces (Type 1 PostScript fonts *URW Palladio L* and *FPL* were used). The code listings are typeset in *Bera Mono*, originally developed by Bitstream, Inc. as "Bitstream Vera". (Type 1 PostScript fonts were made available by Malte Rosenau and Ulrich Dirr.)

The typographic style is the brainchild of Dr. André Miede and was inspired by Robert Bringhurst's *The Elements of Typographic Style* [19]. It is available for LaTeX via CTAN as "`classicthesis`", ©André Miede 2011.

All figures were created by the thesis author using the open source vector graphics editor *Inkscape* [`http://inkscape.org/`], the GUI-based network analysis software package *Pajek* [`http://pajek.imfm.si/doku.php`], and the numerical computing and 4GL programming environment *MATLAB* [`http://www.mathworks.com/`].

Final Version as of October 3, 2011 at 10:01.

DECLARATION

I hereby certify that this thesis, which is approximately 40,000 words in length, has been written by me, that it is the record of work carried out by me, and that it has not been submitted in any previous application for a higher degree. Wherever the contributions of others were involved, every effort has been made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

Limerick, October 2011

Adam W. Hackett