COMPLEX NETWORKS AS AN ANALYTICAL FRAMEWORK: SCIENTIFIC COLLABORATION NETWORKS AND PERSISTENCE OF ENDEMIC DISEASE IN HETEROGENEOUS POPULATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Daniel T. Citron

July 2017

COMPLEX NETWORKS AS AN ANALYTICAL FRAMEWORK: SCIENTIFIC COLLABORATION NETWORKS AND PERSISTENCE OF ENDEMIC DISEASE IN HETEROGENEOUS POPULATIONS

Daniel T. Citron, Ph.D.

Cornell University 2017

Complex networks have proven to be useful as a versatile framework for understanding different systems across many disciplines. This dissertation will use networks in two different contexts for the purposes of answering a variety of questions.

The first chapter will focus on data-driven studies of scientific publishing practices. The recent availability of large electronic publication data sets has made it possible to perform large-scale empirical studies of science. The first section of this chapter will discuss patterns of text re-use among articles in the arXiv, a large scientific corpus. We show how habitual text re-use is restricted to a minority of authors, and that articles containing large quantities of re-used text tend to be cited less frequently.

The second section of the first chapter will study the assembly of scientific co-authorship networks. Previous studies of co-authorship networks have found topological transitions in which co-authorship networks coalesce to form a densely connected community. Such studies have relied on manual annotation of publishing data sets, which has restricted their size and scope to covering only a handful of disciplines. We overcome these limitations using techniques from natural language processing and machine learning to generate a large population of co-authorship networks representing many different disciplines. Consistent with earlier findings, we observe a similar

global topological transition across many different scientific disciplines, suggesting that this is a general property of the development of scientific communities.

The second chapter will use mathematical models to study the persistence of endemic disease in a heterogeneous population. Endemic disease occurs when infection continues to affect a population over an extended period of time instead of dying out following the initial outbreak. Infectious disease modeling can provide important insights into understanding what factors contribute to the persistence of endemic disease. In particular, what role does population heterogeneity play in the persistence of endemic disease? Since the propagation of infectious disease relies on transmission of a pathogen through direct or indirect contact, networks provide an intuitive mathematical framework for modeling the connections between different hosts in a population.

Here, we use the stochastic SIRS model to explore the properties of the endemic disease state, and to understand how a population's underlying contact network affect the persistence of endemic disease. Using a combination of computer simulations and analytical techniques, we find how different model parameters affect the properties of the endemic state. We also uncover a simple phenomenological relationship between the statistical properties of the endemic state and the persistence lifetime that appears to remain robust for a wide range of model parameters and contact networks.

BIOGRAPHICAL SKETCH

The author was born and raised in the San Francisco Bay Area. He entered college at the University of Chicago as a Religious Studies major, with a particular interest in learning about the early history of Christianity in the Roman Empire, but very quickly discovered an interest in subjects where it was possible to use mathematics as a tool for understanding. He switched to majoring in Physics, but scheduled other courses requiring mathematical applications such as economics and statistics.

The author first discovered his interest in complex systems and emergent phenomena when he learned about STARFLAG, an ambitious interdisciplinary research project aimed at imaging, modeling, and understanding the three-dimensional flocking behavior of starlings. Suddenly, Physics was not just the study of subatomic particles or solid state matter as presented in undergraduate coursework - all of nature was open to be studied.

As an undergraduate, the author worked on a wide variety of research projects, including modeling cell division and jamming in disordered systems. After graduation, he spent a brief time working as an intern at a biomedical engineering startup before returning to Chicago to work at the Advanced Photon Source at Argonne National Laboratory. As a research support scientist, the author designed hardware and software for X-ray tomography experiments.

During the prospective graduate student weekend at Cornell University, the author first met his future advisor Professor Christopher R. Myers. Their first meeting was scheduled to last only twenty minutes, but the conversation was so engaging and fun that it lasted over an hour. The discussion featured a wide variety of topics open for study - infectious disease modeling, systems biology, metabolic networks - all of which sounded like interesting, challenging, and worthwhile topics to work on as a graduate student. Enrolling at Cornell in the summer of 2011, the author began working in

the Myers group, investigating ways of characterizing agricultural landscape data for the purposes of disease modeling. This was the author's first exposure to the field of infectious disease dynamics, and he immediately became fascinated by the subject. This fascination has lasted throughout the author's graduate studies, and the author intends on continuing to learn how to use mathematical tools to help contribute to the global fight against infectious disease.

For my loving parents, without whom none of this would have been possible.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my graduate advisor Professor Christopher R. Myers for six years of patience and kindness, and for giving me the freedom and opportunity to explore the topics that I found interesting. Not every graduate student has this type of opportunity, and I am extremely grateful for it. Additionally I would like to thank Professor Paul Ginsparg for the opportunity to explore the arXiv and for guiding me through my first exposure to computational social science.

For the project investigating patterns in text re-use among articles on the arXiv, I would like to thank Professor Paul Ginsparg for designing and leading the project. For the project investigating the assembly of co-authorship networks, I would like to primarily thank my co-author Samuel F. Way for initially suggesting the use of machine learning to sort through the data, and sticking with the project through to the end despite being very far away. The author would also like to thank Brent Schneeman, Laurence Brandenberger, Professor Michael H. Macy, Professor Paul Ginsparg, Haofei Wei, and Alexandra Schofield for supporting the project with their interest and helpful discussions.

The disease modeling project would not have been possible without the guidance and support of my advisor Christopher Myers. Sarabjeet Singh and Drew Dolgert were also very helpful for teaching me how to develop and implement the computer code that made the computer simulations possible. Thanks also to Jason Hindes, Kevin O'Keeffe, Dave Schneider, and Oleg Kogan.

I would like to thank all of the attendees of the 2015 Complex Systems Summer School at the Santa Fe Institute for being such a singularly and astonishingly energetic and curious community of scientists, and for inspiring me to look beyond disciplinary boundaries.

Looking further back, I would like to thank my undergraduate advisor Professor

Sidney Nagel, who first taught me to pursue the research that makes me smile, and who never let me give up. And even further than that, I would like to thank Dr. David Enelow, for being the first to instill in me the idea that interesting topics deserve rigorous scrutiny, and a sense of how rewarding academic work can be.

And lastly, for Team Judge, who see that the world is strange and work hard to keep it that way.

Grant Support

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144153, and NSF. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A network is any system that may be expressed mathematically as a graph consisting of nodes (vertices) and links (edges). This definition is vague, but its lack of specificity leaves the study of networks open for applications in a wide variety of areas. Traditionally, networks have been the focus of graph theory. The study of "complex networks" is a more recent development, brought about by new prospects of conducting data-driven studies of large and complicated systems. In this context, the word "complex" is used to describe networks that lie outside the usual purview of traditional graph theory. The networks that appear in nature are often are neither regular nor symmetric. They are usually sparsely connected, meaning that there is a very low density of edges connecting the different nodes. They may incorporate lots of different modes of heterogeneity - heterogeneity by node degree, heterogeneity of path lengths between nodes, heterogeneity in the types of edges connecting the nodes. Complex networks may also include modular components, or communities of nodes that connect more strongly to one another than to nodes outside. Additionally, many naturally-occurring networks display structure across multiple scales [11, 42, 79].

Complex networks have proven to be extremely versatile as a framework for understanding many different systems across a wide variety of disciplines. After all, many different academic fields, from physics to biology to sociology, involve the study of large systems of components that interact together to produce some emergent behavior. Complex networks provide a generalizable, global approach to analyzing and understanding such systems.

The use of complex networks as an analytical framework can be categorized into two main (but not necessarily non-overlapping) categories. In the first category, the

structure of the network is itself the object of study. To give one example, characterizing how different members of a community communicate with one another [22, 65, 95] requires knowing who it is that interacts with whom in a social network. Using this insight, one can investigate how the structure of those interactions enable or hinder the sharing of information within the community. To give another example, societies depend on power grids and other infrastructure networks to operate consistently. Understanding the structural properties of infrastructure networks, therefore, is crucial for designing them to be robust so that they continue to function in the event of component failure [27, 28, 40]. In both of these examples, understanding the structure of the network allows one to understand how all of the different connections between different components together enable the network's functionality.

In the second category, the network provides a substrate for a dynamical process [11]. Usually in this instance the network is considered as a static parameter that defines the interactions between the different system components, while the behavior of the dynamical process is the object of study. In this case, the question is whether the network's connectivity contributes in some non-trivial way to the outcome of the dynamical process. For example, models for consensus formation, such as the voter model, have been explored on networks for the purposes of investigating how network structure influences a community's ability to reach an agreement [87].

This dissertation focuses on the use of complex networks as frameworks for understanding two very different problems: analyzing collaborative communities of scientific researchers, and incorporating interaction heterogeneity into models of infectious disease dynamics. The first chapter will focus on data-driven studies of scientific publishing practices. The first section will explore patterns of text re-use among authors who submit articles to the arXiv. The second section will characterize the network

structure of communities of scientific collaborators, showing how co-authorship networks assemble and evolve over time. In both cases, network analysis will provide insight into the global patterns of how authors and articles connect to one another.

The second chapter will discuss the persistence of endemic disease in heterogeneous communities. Infectious disease modeling represents the transfer of infection between hosts in a population using dynamical models. In this chapter, networks will be used to mathematically represent the heterogeneous connections between the different hosts. This makes it possible to explore the question of how a contact network's structure affects the persistence of endemic disease.

CHAPTER 2

DATA-DRIVEN STUDIES OF SCIENTIFIC PUBLISHING

2.1 Introduction

The sociological study of science explores questions of how individual researchers interact with, compete with, and collaborate with one another in order to make scientific advances [38]. Scientific progress relies on the dissemination of knowledge of discoveries, methods, and theories through the research community. One example of this is the adoption of the use of Feynman diagrams as a method for calculation by the particle physics community. Without the strong social ties between colleagues, mentors, and students, it would have been very difficult for such a tool to come into general use [58]. The importance of such social ties, and how they support the spread of ideas and information, suggests that networks may be useful for understanding how the scientific community operates.

Indeed, networks have become an important framework for studying science. One of the earliest examples of using networks to understand the practice of science was undertaken by de Solla Price in 1965, who drew from a painstakingly indexed data set of more than a million citations between articles in the field of genetics [45]. de Solla Price explicitly outlined a network representation for how new articles cite older articles from related fields. He used the citation network to document an exponentially decreasing distribution in the number of citations an article receives, as well as the apparent exponential decrease in the number of times older articles are cited compared to newer articles. Given the relative importance of recently published articles, de Solla price came to the conclusion that an "alerting service" that would disseminate news of important articles would be of assistance to mid-century scientists [37].

The introduction of electronic publishing and online repositories of scientific articles has enabled large-scale studies of scientific research practices. Not only do the venues for electronic publishing act as an important resource for scientific researchers, but also they themselves act as data sets for studies of science [21, 23, 47, 91]. Without these data sets, such studies would require painstaking and time-consuming work by researchers to sort through and categorize millions of existing scientific articles and their relationships to one another. Such demanding labor requirements would make it impractical to pursue an empirical understanding of the practices and research output of the scientific community.

Electronic publishing has greatly expanded the possibilities for empirical studies of the sociological aspects of science. It is now possible to trace many different sorts of connections and relationships that exist between different scientific articles, researchers, and academic institutions. For example, a co-authorship network may be used to describe the relationships between authors who have collaborated together on one or more articles. Generally, co-authorship networks use nodes to represent authors and links to represent collaborations between pairs of authors, but one may also use a bipartite author-article network in which author nodes link to articles they have written and article nodes link to their authors. Co-authorship networks reflect social reality, as the creation of a link only appears of two authors have worked together and produced at least one scientific article [77, 78].

The static properties of co-authorship networks have been described in great detail. Measurements of the degree distributions of numbers of authors per paper, number of papers per author, or total number of collaborators per author in these networks reflect the typical sizes of scientific collaborations. The size of the largest connected component represents the extent to which a community is connected to

itself [75, 77, 78]. Exploring the paths connecting pairs of nodes allows one to characterize how "distant" authors are from one another, quantifying how closely different members of a research community work with one another[76]. Examining these paths further, one may also identify particular authors as being particularly important in that they serve as bridges between otherwise disconnected clusters [29, 76, 79]. Community detection algorithms [44] may also be used to detect groups of authors who publish together frequently than with others, revealing structural divides that may reflect geographical, institutional, or disciplinary separation between groups of collaborators [49]. Citations data has also been combined with co-authorship network data in order to evaluate how an author's placement in the co-authorship network affects which articles he or she chooses to cite, including the probability of self-citations or citations of close collaborators as opposed to citations of more distant authors [66]. Together, all of these different analyses provide important quantitative insight into the activities of the scientific community.

One may also use networks to study more complicated questions, such as the importance of teamwork in scientific research. Scientific research collaborations have been increasing in size over time [66] and it has become increasingly common to cross institutional and disciplinary boundaries [22, 96]. One of the apparent reasons why research collaborations grow in size is that these larger groups can combine multiple types of expertise, making them more effective at addressing complicated research problems, and making it possible to have greater impact [22, 96]. Network models are particularly important for understanding how these new collaborations form [53], and so provide an understanding the incentives and organizing mechanisms that lead to successful research strategies.

2.1.1 Chapter Summary

This chapter will discuss two data-driven studies of the practice of science, and demonstrate how networks can be useful tools for understanding complicated data sets. The first section (Section 2.2) will focus on the problem of authors who produce scientific articles that contain large amounts of text re-used from previously published articles. The study uses a new technique for comparing the textual content of a large number of scientific articles from the arXiv. All pairs of articles are compared this way, making it possible to explore questions about the prevalence, frequency, and distribution of text copying behavior among authors who submit articles to the arXiv. This section reports basic statistics of patterns of text re-use, as well as measures the correlation between how many citations an article receives and how much re-used text is present in that article. This work was originally completed under the direction of Paul Ginsparg [31].

The second section will discuss the formation and assembly of scientific collaboration networks. Many previous studies have used publication data sets of scientific articles to explore the formation and evolution of networks of co-authors. These studies often focus on a small number of scientific fields, analyzing each field individually and characterizing the development of the research community by measuring the properties of its corresponding collaboration or citation network. Section 2.3 will go further, using a large publication data set from the arXiv in conjunction with tools from machine learning and natural language processing to algorithmically identify a large population of scientific fields. Each field is represented by a group of articles with similar content. Such a large set of fields makes it possible to perform a large-scale comparison across many different fields of varying size and specificity, making it possible to test whether there are general rules to the development of scientific

co-authorship networks. This work was completed in collaboration with Samuel F. Way ([32], in review).

2.1.2 About the ArXiv

The arXiv serves as the data set for both of the sections presented in this chapter. The arXiv is an open-access repository of scientific preprints accessible online atwww.arxiv.org. The site was founded in 1991 and, as of the end of 2016, hosts over 1.1 million articles, primarily in the areas of Physics, Mathematics, and Computer Science [2]. By December of 2016, the arXiv was growing at a rate of around 9450 new articles submitted per month [2].

The arXiv has proved an invaluable resource for researchers to share their research output with one another[48], and also contains in aggregate a large amount of data that makes it possible to study patterns in research practices on a global scale. The arXiv data set include articles' full texts as well as relevant metadata (article titles, author names, date of submission, etc.). Additionally, arXiv has been well studied from a scientometric perspective (e.g. [47, 63]), and so is known to be useful for understanding the scientific research community.

2.2 Text Overlaps in a Scholarly Corpus

2.2.1 Introduction

This first project serves as an example of how one can use a complicated data set to understand patterns of behavior that occur in the scientific research community. In this particular case, the patterns that occur - the frequent re-use and subsequent presentation of text already used in a previously published article - serve as evidence of a pattern of unhealthy publishing behavior in certain sectors of the scientific research community.

As discussed in a previous paper [88], the "text winnowing" methodology of [86] was adapted to evaluate the amount of text shared in common between two articles. This adapted methodology is extended to systematically compare the textual content in all pairs of articles contained in the data set. This makes it possible to look for global patterns in text re-use across all of the arXiv. "Text re-use" here refers to the practice of submitting an article for publication that copies verbatim text that has been published elsewhere. The data set used for the current analysis consisted of over 760,000 articles submitted to arXiv between mid-1991 to mid-2012, towards the end of which time it was receiving roughly 80,000 new submissions per year[1].

Before implementing the systematic comparisons between articles, the administrators of arXiv had no method for detecting text re-use. The only evidence for the existence of this publishing practice came from individually reported cases of plagiarism. To give a few anecdotal examples, the authors of [72] pointed out unattributed use of their text in a series of four arXiv articles in 1999. Second, a news article from 2003 [46] described the case of an otherwise unknown person who tried to establish

research credentials for career advancement by submitting texts largely copied from other sources. Third, a news article in 2007 [43] noted that at that time known cases of text re-use spanned a wide range, from 27 pages of lecture notes by another author used verbatim in a thesis, to re-use of common introductory material, to text overlaps of benign common phrases. Lastly, as reported in another news article [26], a large number of articles from a group of coauthors was withdrawn from arXiv due to re-use of text copied from a variety of sources. Collecting many such cases, the administrators were motivated by these anecdotal reports to systematically study the prevalence of text re-use on the arXiv.

Knowing the prevalence of text re-use was also important for improving administration of the arXiv, since authors who habitually re-use text present an inconvenience to readers [17]. Problematic authors include those who (intentionally or otherwise) artificially inflate their publication count by reusing large blocks of text in each submission. We make no attempt to interpret the motivations of authors who engage in this practice. There are many possible reasons why someone might re-use text, not all of which are necessarily pernicious. Previously, screening for these had been haphazard, and moreover there was no systematic baseline to identify outliers or to provide a principled response to the claim that conspicuous re-use of text was common practice and therefore accepted by the community. The current work provides a more systematic assessment of the statistics of text re-use in the full arXiv dataset and enables arXiv administrators to identify extreme cases of text re-use. Indeed, it is now possible to immediately detect articles containing excessive amounts of re-used text, and as of May 2012 arXiv administrators now publicly flag these articles accordingly.

Publishing Norms and Text Re-use

While there is no universal standard regarding the reuse of text in scientific publications, many universities and publishers have established explicit guidelines regarding publishing articles that contain text reused from a previous source. Universities, including Cornell [4], typically point to materials at the Federal Office of Research Integrity [5], stating that "Substantial unattributed textual copying of another's work means the unattributed verbatim or nearly verbatim copying of sentences and paragraphs which materially mislead the ordinary reader regarding the contributions of the author." Policies in other countries, where available, are similar. The US Federal materials clarify that use of common phrases within a community is not considered misleading, and a finding of misconduct generally requires a "significant departure from accepted practices of the relevant research community." Similarly, the American Physical Society's guidelines regarding the content submitted to its journals are unequivocal regarding text reuse: "Authors may not ... incorporate without attribution text from another work (by themselves or others), even when summarizing past results or background material. If a direct quotation is appropriate, the quotation should be clearly indicated as such and the original source should be properly cited" [6]. These guidelines do permit "material previously published in an abbreviated form" to provide the basis for a more detailed article, as long as reproduction of previously used material is minimized and properly referenced.

To be clear, this analysis is restricted to detecting text overlaps and does not attempt to detect plagiarism in its most general form, which includes unattributed use of ideas. That is to say, not all cases of plagiarism are detected with our methods, as it is possible to copy an idea without copying the original text. Furthermore, the analysis is restricted to simple factual statements regarding the observed patterns of

text overlap of materials included in our data set. No attempt is made to detect text copied from sources outside of arXiv (e.g. Wikipedia or he rest of the WWW), so attention is restricted to a simple factual statements regarding textual overlap of materials only within arXiv.

2.2.2 Methodology

Pre-processing

Before performing our analysis of the arXiv, the collection of articles are first pre-processed and sorted to avoid the inclusion of false positives. That is to say, there are features specific to the arXiv dataset that cause our text reuse methods to over-estimate the amount of text reused and the frequency of text re-use. Each text is processed to remove the reference section, since text overlaps among the references are not of interest. Author names from very large experimental collaborations (e.g. ATLAS or CMS) are also excluded, since these can masquerade as authors reusing their own text. Whenever possible, block quotes are also identified and ignored (but find in any event that these are a very tiny fraction of the text re-use detected in the corpus).

Winnowing

Text re-use is detected by using an index database to quickly compute the text overlap between any pair of papers. This database is constructed using a representative subset of the text to characterize each article. For fast comparison between articles, this database should fit in RAM, so its size is reduced using the following winnowing

methodology. A text winnowing methodology (described in [88], as adapted from [86]) is employed to quickly compare the text of all pairs of articles in the corpus.

Each article can be effectively "fingerprinted," with its content represented by a set of hashes stored in a database that resides in memory (RAM) for rapid lookups. The hashes are determined by sequences of seven words in the article, called 7-grams, eliminating sensitivity to commonly used shorter sequences (e.g., "this article is organized as follows"). The number of hashes retained for each document are "winnowed" [86], which reduces their number by a factor of 3.6 (at a small loss of sensitivity to word sequences of fewer than 12 words), and further reduced (by another 4%) by eliminating "common" 7-grams [88]. The resulting hash database requires about 12 Gb of RAM and permits many hundreds of lookups per second on inexpensive hardware. (A more detailed explanation of this methodology may be found in the supplemental material for [31].)

Detecting text reuse

Having constructed the index database of papers and hashes representing uncommon 7-grams, the textual content of each pair of articles in the data set is compared. If two articles have at least one hash in common then there is overlap between the two papers, indicating that the later paper has re-used text from the earlier paper. For typical amounts of text overlap, the number of overlapping words is roughly six or seven times the number of such overlapping 7-grams. Thus, two articles with 100 overlapping 7-grams can be thought of as having roughly 35 sentences in common.

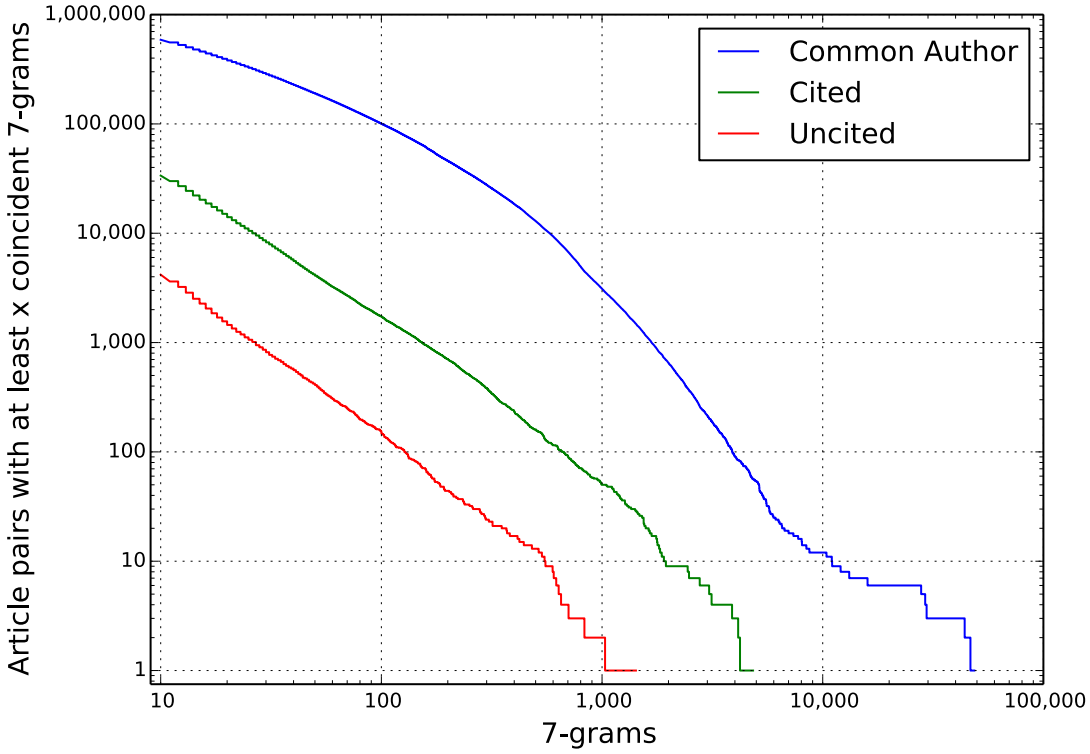

Figure 2.1: **Cumulative Distribution of Text Overlaps**: Cumulative distribution of the number of overlapping 7-grams across all article pairs with Common Author in blue, Cited in green, and Uncited in red. The vertical axis is the number of article pairs with at least the number of overlapping 7-grams given on the horizontal axis (starting with a minimum of at least 10). Both horizontal and vertical axes are logarithmic.

2.2.3 Global Prevalence Text Reuse

In the following analyses, we distinguish between three distinct modes of text reuse, in increasing order of severity: "Common Author" (AU) designates a pair of overlapping articles with at least one author in common; "Cited" (CI) designates a pair with no common authors but at least one article cites the other; and "Uncited" (UN) designate a pair of articles with neither common authors nor citation of the earlier article.

Fig. 2.1 shows the frequency with which incidents of text overlap between papers are detected in the dataset. Each curve represents the cumulative number of article pairs with at least the number of coincident 7-grams specified on the horizontal axis. The three curves represent the three different modes of text reuse, with Common Author, Cited, and Uncited colored in blue, green, and red, respectively. For example, the Common Author curve in blue, there were roughly 100,000 cases with at least 100 7-grams in common, 3000 with at least 1000 in common, and only about 10 such pairs with as many as 10,000 in common. The logarithmic scale on the y-axis shows that Common Author text reuse is approximately an order of magnitude more frequent than Cited text re-use and approximately two orders of magnitude more frequent than Uncited text re-use.

At first glance, the data represented in Fig. 2.1 suggests cause for concern: is the literature really so replete with text re-use? Are there truly so many authors who repurpose their own text and that of other authors, with or without attribution? Before jumping to conclusions, there are perhaps other various mitigating circumstances related to the re-use of textual content in the context of arXiv. In the case of authors reusing their own past material, it may be that such recycling is sometimes acceptable practice. For example, doctoral theses in physics once consisted largely of original materials, but graduate students are now expected to publish multiple articles, and it is a common practice for the thesis to incorporate some of these articles in their entirety, without changes. Similarly, in most disciplines it is also considered acceptable to have separate short and in-depth versions of the same work, with the former incorporated into the latter. There is also the case of review articles, in which the acceptability of reusing text is somewhat more contentious. Some authors take it for granted that review articles should be original syntheses of past work, whereas others feel free to use large blocks of material from earlier articles. Attitudes towards reusing text in

conference proceedings vary widely, differing between authors and across fields. In Physics publication, for example, conferences are a secondary publication venue, and it is accepted that authors will re-use earlier material. In Computer Science, on the other hand, conference publication is a primary venue, and significant self-copying by authors is not the norm. Lastly, lecture notes, book contributions, and other popularizations constitute another form of publication in which liberal re-use of earlier material could be considered acceptable.

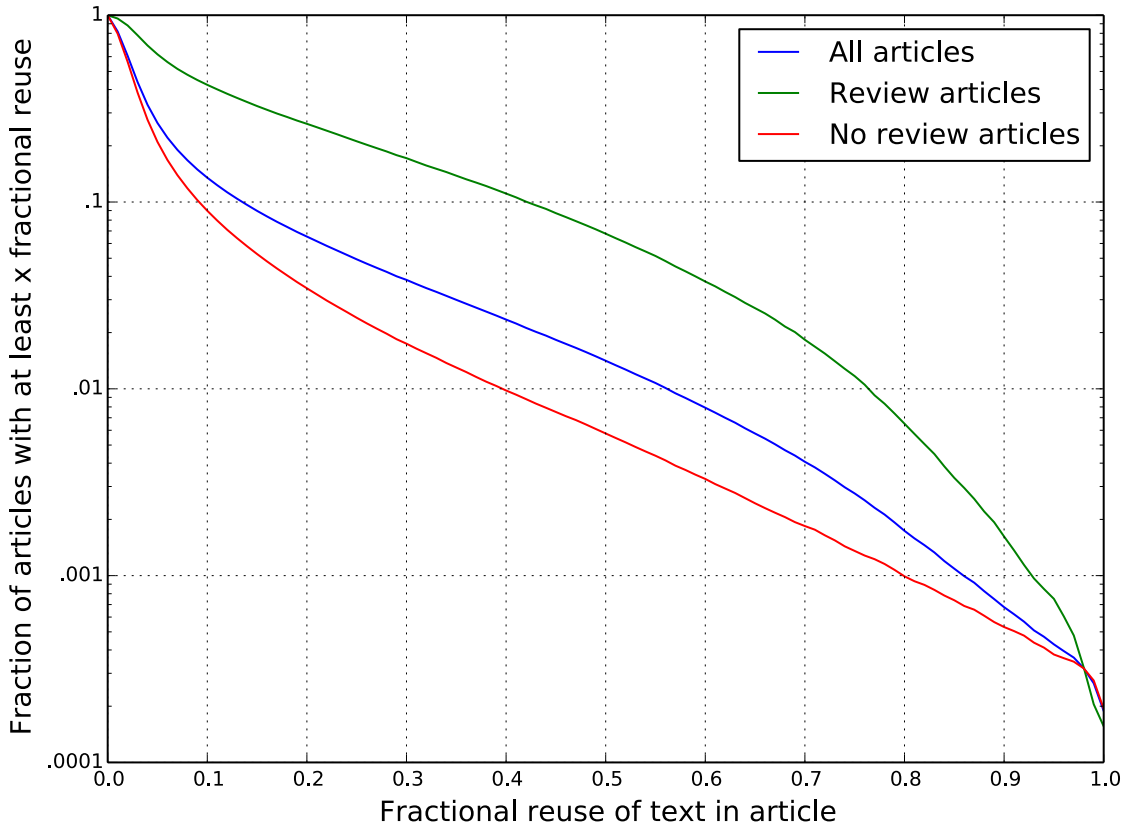

Figure 2.2: **Text Overlap Distributions s in Review and Non-Review Articles**: the vertical axis gives the fraction of articles with at least the indicated fraction of reused 7-grams on the horizontal, where green (upper) signifies Review articles, red (lower) signifies non-Review articles, and blue (middle) combines both. The vertical is plotted on a log scale to permit seeing the full range; the dropoff in fraction of articles with given amount of reuse would be much steeper on a linear scale.

To assess the extent to which text reuse is concentrated among articles in the

aforementioned classes (review articles, conference proceedings, dissertations, and so forth), a subset of articles is denoted as "Review." The results from Fig. 2.1 are partitioned according to this new categorization. Articles are designated as belonging to the Review category if the article metadata (abstract, keywords, etc) includes keywords such as "review", "proceedings," or "thesis" in order to detect articles that were self-identified by submitters as review-type. Fig. 2.2 shows how this partition changes the results from before. The horizontal axis shows the fractional text reuse within the article (given by the fraction of 7-grams in an article that appear in some other article) and the vertical axis indicates the fraction of articles in the database with that percentage of reuse. The middle solid line (blue) shows the fraction of all articles (Review and non-Review) with at least the indicated fractional reuse. For example, articles in which 50% of 7-grams appear elsewhere comprise roughly 2% of our dataset. The upper solid line (green) isolates from that set the fraction of articles self-identified in the Review category, and shows the fraction of those articles with the indicated fractional reuse. Roughly 7% of those articles contain at least 50% reuse, whereas less than .6% of the non-Review articles (solid red line) have as much text reuse. Thus, Fig.2.2 shows that the vast majority of the common author text reuse seen in Fig.2.1 occurs in contexts generally regarded as acceptable by the community. What remains problematic and will be discussed further is the group of occurrences, represented by the red line, with a non-negligible percentage of text reuse that does not occur in those contexts.

2.2.4 Text Reuse by Individual Authors

Given the prevalence of text reuse, it is natural to wonder how these texts are distributed between the authors. That is to say, is text reuse concentrated among a few

serial offenders, or whether most authors reuse text some of the time? The following analysis shows the distribution of cases of text re-use across all authors in the dataset. This will establish the extent to which text re-use is "normal" behavior by quantitatively identifying behaviors that stand out as abnormal.

Text Overlap Networks

To illustrate the distribution of text re-use by authors, we construct and examine text overlap networks. In a text overlap network, each node represents an article and each edge represents a pairwise textual overlap between two articles. Because articles published later in time copy from earlier ones (and not vice versa), all edges in the network are directed forward in time to represent the transfer of text. Each edge is weighted according to the number of 7-grams that the two connected articles have in common. Again, the different modes of text overlap are distinguished and colored differently (AU in Blue, CI in Green, UN in Red).

Given all articles written by a particular author, the author's text overlap network illustrates whether or not the author habitually re-uses text. The density of connections for a specific author's network is proportional to the amount of text reused by that author, so the text overlap network provides a useful framework for visualizing the extent of text reuse within a set of articles and for examining how articles by a particular author or group of authors overlap with one another. Fig. 2.3 shows the text overlap networks of two authors with vastly different patterns of text reuse. Articles by Author A have few overlaps: of 217 co-authored articles, only 6 contain previously published text; whereas Author B's text overlap network is far more densely connected. The blue edges reveal clusters of articles by Author B with material copied from one another. Furthermore, in contrast to Author A, Author B

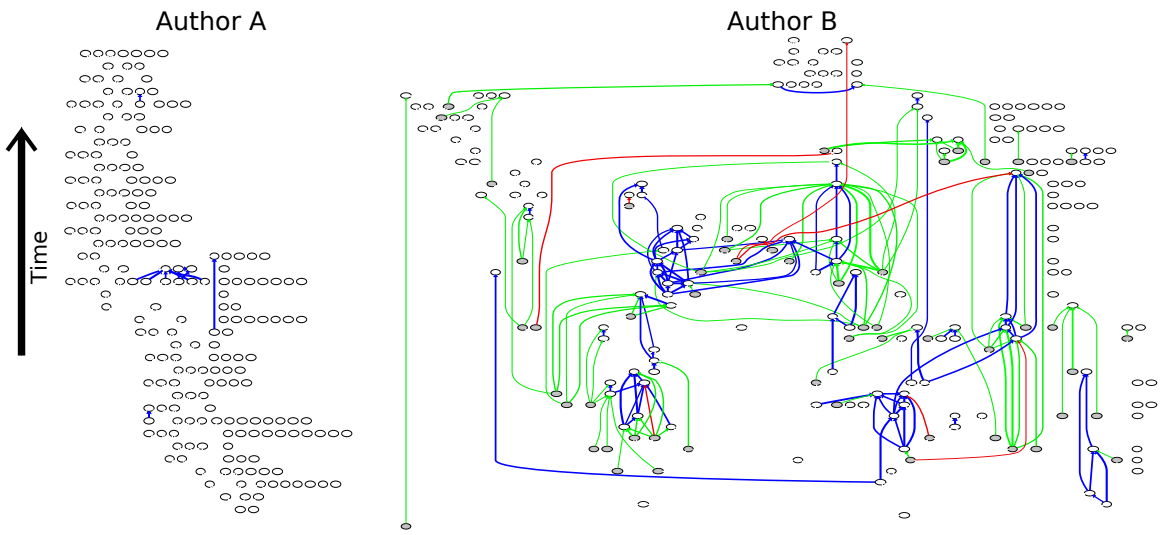

Figure 2.3: **Example Text Overlap Networks**: Visualizations of the text overlap networks of two authors, A and B. The blue, green, and red edges represent Common Author, Cited, and Uncited text overlaps, respectively. The edge thickness increases with the amount of overlap between the two articles. Articles are arranged in the diagram by time of submission, with the earliest articles grouped near the bottom and more recent articles at the top. Uncolored nodes indicate texts coauthored by the author of interest, and gray nodes represent texts by other authors, included where the author of interest has reused text therefrom.

has also reused text from articles written by other authors (represented by green- and red- colored edges.)

While the example of Author B demonstrates that it is possible to produce large numbers of articles more quickly by copying from prior content, the example of Author A demosntrates that it is not necessary to copy or self-copy to generate a large number of publications. Author A submitted 177 articles and Author B submitted 174 articles between January 2000 and June 2012, each averaging about 1.2 articles submitted per month in that period. While both authors are prolific, only the latter habitually copied previous text. Prolific authors should not automatically be suspected of habitual text re-use (nor are text re-users necessarily as prolific as author B). While many or most authors have little desire to retread the same material more than once, there are also authors whose publications tend to consist largely of

previously published material, with minimal new content.

Detecting serial copiers

These qualitative observations suggest a quantitative measure of an author's tendency to reuse text: namely, the fraction of an author's articles that include significant amounts of copied material. In general, small overlaps are not of interest. To focus on the more significant occurrences, we consider only cases of at least 100 7-grams in the case of Common Author overlaps and at least 20 shared 7-grams in the case of Cited or Uncited overlaps. Recalling the winnowing procedure, these thresholds correspond approximately to 50 and 10 sentences of copied text, respectively. For the Cited and Uncited overlaps, we choose a lower threshold because these modes of copying are more problematic than the case of self-copying. (Fortuitously, Cited and Uncited overlaps are far rarer than the case of Common Author copying, so our lower threshold does not yield a surfeit of detected cases.) The results are insensitive to the choice of thresholds, as slightly changing the thresholds does not change the group of authors whose copying behavior is considered outside the norm. The thresholds filter out insignificant instances of text reuse. Additionally, implementing thresholds aids in reducing the number of false positives stemming from pdf to text conversion errors, author or citation lists, restatement of theorems, or an occasional block quotation of text. To restrict attention to habitual and frequent reuse of text, we include only authors who have submitted at least 4 articles.

Fig. 2.4 shows the cumulative histogram of the number of authors whose articles contain a given fraction of significant AU, CI, and UN text overlaps. Most importantly, the number of authors with articles flagged for each of the three types of overlaps drops significantly as the fraction of problematic articles increases from 0%.

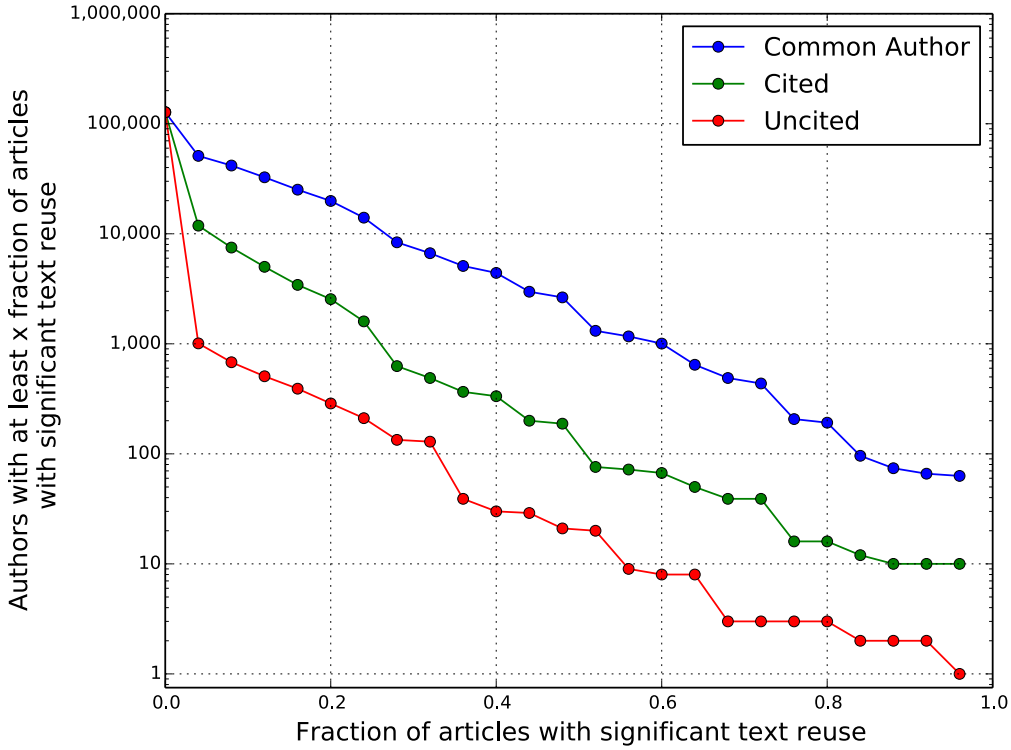

Figure 2.4: **Cumulative Histogram of Authors vs. Fraction of Articles Containing Significant Text Re-Use**: Cumulative histogram of the number of authors (vertical axis) having at least a given fraction of their articles with significant text overlaps (horizontal axis).For example, roughly 1720 authors have significant AU text overlap in at least 50% of their articles. Common Author (AU), Cited (CI), and Uncited (UN) overlaps are plotted in blue, green, and red, respectively. Articles with "significant" text overlaps have at least 100 7-grams re-used(AU) or 20 7-grams re-used (CI or UN). Note that the vast majority of authors rarely re-use a significant amount of text from other sources.

Of the total 392,850 authors in the dataset, only 49,830 have at least 1% of their articles contain AU text overlaps; only 8990 contain CI text overlaps; and only 1630 contain UN text overlaps. The vast majority of authors, therefore, either never or only rarely reuse significant amounts of text in new publications. In the problematic region, there are only 10,550, 1130, and 130 authors with at least 25% of their articles containing significant AU, CI, and UN overlaps, respectively. It is clear that the practice of reusing text is uncommon and is restricted to a minority of serial offenders, responsible for the heavy tail in Fig. 2.1.

Text Overlap and Citations

Knowing now that the excessive re-use of previously published material is restricted to a small minority of authors, the next step is to investigate their standing in the global scientific community. Are serial copiers influential or not, and do their articles have an impact on the research community? To assess the impact that serial copiers have, we use the number of citations that each article has received as a measure of its influence, and investigate possible correlation with the amount of copied content in the article. This stage of the analysis focuses on a subset of 116,490 articles for which there exists relatively clean citation data, primarily in Astrophysics and High Energy Physics (provided by Alberto Accomazzi from the Astrophysics Data System).

The fraction of copied content contained in an article is estimated by dividing the number of 7-grams that have appeared previously by the total number of 7-grams from the article, without removing the common 7-grams. All articles containing 95% or more copied content are excluded from this analysis, since these are typically articles erroneously submitted more than once to arXiv after minor revisions and do not represent the phenomenon of interest. All articles with less than 5% copied content are also excluded, because often these articles contained errors in the pdf to text conversion, for example due to font issues, making the estimate of fraction of copied content unreliable. (Note that including common 7-grams means that all properly converted texts will exhibit some reused content.)

Fig. 2.5 shows the number of citations plotted against the fraction of copied content contained in each article. The wedge of points at the left of the scatter plot shows that there is a higher variance in the number of citations for papers containing low amounts of copied content. Qualitatively speaking, it is more likely for papers with a low fraction of copied content to receive very many citations, whereas it is relatively

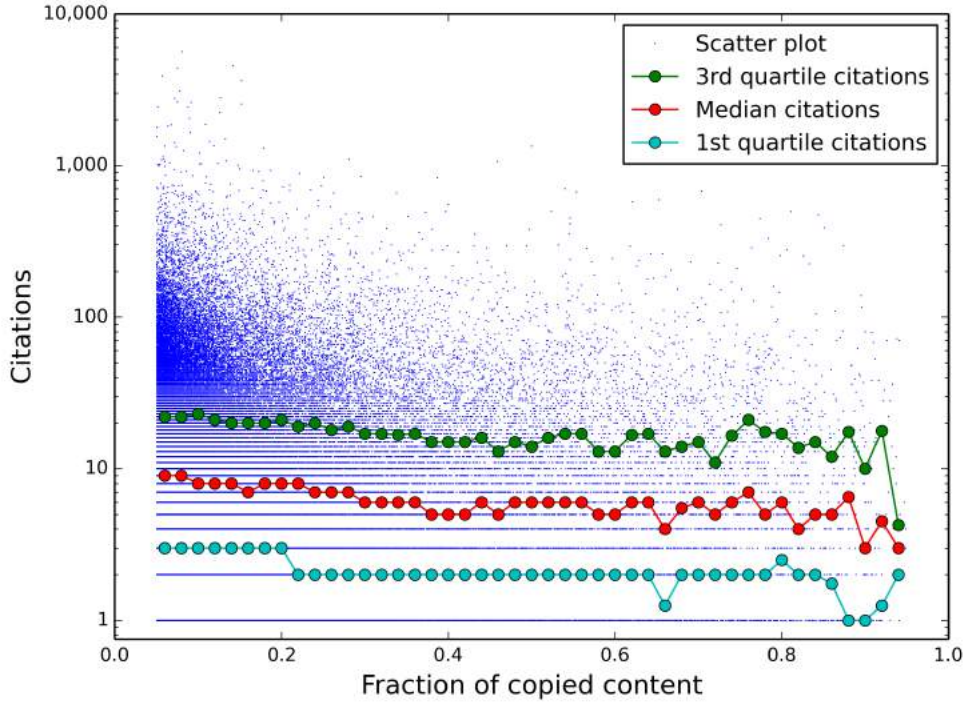

Figure 2.5: **Number of Citations vs. Fraction of Copied Content in Each Article**: Scatter plot of the number of citations vs. fraction of copied content (blue). The median number of citations vs. fraction of copied content is shown in red, indicating a negative correlation between the number of citations and the amount of copied content. The y-axis is logarithmic, and the plot also shows 1st and third quartiles for the citations. The Spearman correlation coefficient for the median is $r = -.739$ ($p = 6.76 \cdot 10^{-9}$), meaning that text re-use is negatively correlated with citations received.

rare for papers featuring a high fraction of copied content to receive the same number of citations. To quantify this, the figure also shows the median number of citations as a function of fraction of copied content in red, which has a Spearman correlation coefficient of $r = -.739$ ($p = 6.76 \cdot 10^{-9}$). This illustrates a strong decreasing trend of citations for articles with increasing copied content. The presence or absence of reused text in an article thus serves as a quality flag, since articles with large amounts of copied content tend to be cited less frequently by other research groups, and are considered less important.

2.2.5 Discussion

An efficient method for detecting text overlaps between pairs of articles has been applied to a large corpus of scientific articles from the arXiv. Analyzing the patterns in text re-use across all articles, one can establish a baseline standard for "common practice." It is clear that, although text re-use is common, instances of text re-use are common only among a small minority of the authors on arXiv. This project serves as one example of how the availability of large scientific publications data sets has made it possible to analyze the practices of the scientific community on a global scale.

The text overlap networks that appear in the above analysis are a useful framework for illustrating how different articles borrow text from one another. In a way, these networks represent the transfer of information between different members of the scientific community, although the mode of information transfer does not appear to lead to healthy or impactful research practices. The following section uses network analysis to explore another type of network from scientific publishing - co-authorship networks. These networks represent the different collaborations that occur between different researchers. These collaborations involve the sharing of knowledge and skills, and so, in contrast to the text overlap networks, do appear as a the result of practices that benefit the scientific community.

2.3 Assembly of Co-Authorship Networks

2.3.1 Introduction

A scientific field of study is defined not only in terms of its research questions, but also in terms of the institutions, conferences, journals, and other formal and informal professional networks through which researchers communicate with one another [38]. Such communities allow for the transfer of knowledge, skills, and other resources required for researching complex problems [22, 38, 53, 58]. A co-authorship network is a conservative representation of a research community, that outlines the professional connections between scientific researchers. Co-authorship networks are important objects of study, as they are an empirically measurable representation of the communities that assemble in order to work in an area of research.

Previous Work

Two recent studies have investigated the development of 9-12 research fields by measuring the assembly of each field's co-authorship network using a large electronic collection of articles [15, 16]. They search for patterns in the growth and development of co-authorship networks across different scientific fields. These studies argue that while each field differs in size and publishing practices (differing in rate of publication, size of collaborations, etc.), nevertheless there appear to be common patterns in how each field's co-authorship network develops. Specifically, each co-authorship network undergoes a topological transition in which a densely connected giant component of researchers forms over time. This dramatic structural change is similar to a percolation transition [79], and serves as an empirical indication that the research

community undergoes large-scale social reorganization as more researchers join and collaborate with others [15, 16, 53].

Another study [64] takes three example fields (complex networks research; ADS/CFT; Randall-Sundrum model) and describes three stages of development characteristic to co-authorship network assembly in science. The co-authorship network begins as a set of disconnected groups, which then join together to form a large tree-like component. As the research community grows and mixes further, the largest component becomes densely connected to itself through the formation of long-range ties. This general pattern is consistent with what was reported in [15, 16], which also emphasized how the long-range ties between authors created a densely connected community with very short distances between different authors.

Together, these previous studies suggest the existence of common patterns in how scientific communities assemble over time. However, they rely on manual annotation of their data, which requires a great deal of labor in order to assemble a co-authorship network. This in turn limits the number of examples studied and reported on, making it difficult to justify the claim that the patterns observed for a few examples are universal across all scientific fields.

Machine Learning for a Larger-Scale Survey of Communities

The present study proposes a framework for analyzing a large population of examples in order to verify that the development of co-authorship networks, as characterized by earlier studies, is robust across many scientific fields. Specifically, we use techniques from natural language processing and machine learning to generate a larger set of example co-authorship networks from the arXiv, a large scientific corpus. Topic modeling is employed to cluster articles together based on their semantic content,

and we interpret the clusters of related articles as representing different fields of science.

Measurements of the algorithmically-generated co-authorship networks can show whether they develop in a manner similar to the manually-annotated co-authorship networks studied previously. With this methodology, we aim to facilitate a larger survey of co-authorship networks across scientific fields first by testing the efficacy of topic modeling as a way to rapidly detect a large number of fields, and then by comparing the assembly behavior of each field's co-authorship network for the purposes of testing whether their growth patterns remain consistent for a large set of fields of varying size and specificity.

2.3.2 Data Set

The data set used for the present analysis includes the 189,000 articles categorized as Condensed Matter Physics by the submitting author ("cond-mat" on the arXiv), beginning in April of 1992 and ending in June 2015. The following data from each article are used: a list of author names; the date the article was added to arXiv; the title; and the abstract.

In addition, a subset of condensed matter articles from the Web of Science (WoS) is also employed for the purposes of validating the results obtained using the arXiv data set. WoS is a database of peer-reviewed scientific articles curated by Thomson Reuters. To compliment the arXiv data set, we also use the 660,000 articles classified as Condensed Matter Physics published between April 1992 and June 2015. Each of these articles has a title, abstract, and list of author names available in the Web of Science database [3]. The set of articles from Web of Science partially overlaps

with the arXiv data set and represents a complementary data set with non-uniform coverage of the subfields contained on arXiv [63]. The set of arXiv articles is only a sample of all published works, and, due to differences in the site's adoption across communities, arXiv's coverage varies from one subfield to the next. Using a second data set makes it possible to verify that any results obtained using the articles from the arXiv reflect a truly representative sample, and are not caused by the arXiv's incomplete coverage of certain scientific subfields.

To track the contributions of individual authors, we adopt the convention of labeling each author with "[First initial] [surname]" (e.g. "Lindsay M. Barnes" becomes "L Barnes") in order to address variation in author naming conventions (e.g. Jim vs. James; or inconsistent inclusion of middle names and initials). This convention errs on the side of fewer rather than more individual authors, and it does create the possibility of two different authors' names overlapping. For the present study, however, the possibility of names overlapping is mitigated by restricting analyses to the set of authors publishing within a particular subfield of physics. Larger-scale analyses involving a broader reach of disciplines will require additional steps to disambiguate author identities. After preprocessing author names in this way, the arXiv data set includes 96,000 unique authors.

For the purposes of text mining and topic modeling, scientific content of an article is represented by its title and abstract under the assumption that authors write titles and abstracts with the intention of concisely summarizing an article's contents. Past studies have argued that focusing analyses on article abstracts has the additional benefit of minimizing the amount of "structural" text processed by the topic model, allowing the inferred topic structures to focus on field-specific content, rather than commonalities in presentation of the English language [47, 57].

2.3.3 Methods

Topic Modeling

Past studies exploring the formation of co-authorship networks have relied on manual annotation to determine which authors contribute to and are therefore considered part of a scientific field [15, 16, 64]. This approach, however, requires a great deal of human effort and, consequently, has been applied to only a few disciplines and with somewhat arbitrary definitions of which publications and authors belong to the community in question. It remains unclear how robust past results are to varying the criteria for selecting communities, and for varying levels of specificity governing the breadth and size of such communities.

To address these limitations, we introduce an approach that uses topic modeling to automate the process of identifying groups of semantically-related documents and partitioning their authors into fields corresponding to their areas of expertise [24]. As a consequence of the number of documents belonging to a given subfield and the commonality of its language, the topics and therefore the fields extracted by this technique will vary in terms of size and specificity, yielding a population of corresponding co-authorship networks. That is, it becomes possible to test whether the reported structural patterns are robust to varying definitions of sub-community. At the same time, we explore the usefulness of topic modeling as an automated, scalable means for partitioning the global network of all researchers into co-authorship networks organized around specific fields.

Topic modeling is an unsupervised machine learning technique that characterizes the underlying thematic content of a given corpus by identifying groups of semantically-related, co-occurring words—the "topics"—while simultaneously iden-

tifying the proportion of each topic present in each document in the corpus. Here, we use latent dirichlet allocation (LDA) [20, 52], a popular topic model that produces static definitions for topics, formalized as probability distributions over all words in a given vocabulary. Accordingly, for each document the model infers a distribution over these topics. In summary, the LDA algorithm takes as input a set of documents, each of which contains a group of words, and yields two main outputs: a probability distribution of each word's occurring in a particular topic, and a probability distribution of each topic representing a particular document.

Prior to applying topic modeling, several common natural language processing techniques are used to preprocess the corpus text. For each article, the text from the title and abstract is combined into a single document, all non-alphabetic characters are removed, and all letters are converted to lowercase. Common English stop words ("the," "and," "of," etc.) are also removed, as well as certain words that appear very commonly in the arXiv data set but that contain no scientific content (numbers, names of publishers, "thank you," etc.). Lemmatization is applied in order to increase increase the likelihood of discovering overlaps in the word usage within and between documents. For example, this process converts "wolves" and "Wolf" to "wolf," combining the counts of each of a word's possible forms into a single count captured by its lemma.

After preprocessing all articles, MALLET [68], an open-source implementation of LDA, is used to train a series of topic models, varying the number of topics between $k = 25$ and $k = 100$. As expected, for small k, LDA produces broadly-defined topics, and for large k, more narrowly-defined topics. For purposes of the present study, $k = 50$ provides sufficient resolution for the model to recover topics that resemble established subfields within condensed matter physics. We emphasize that we do

not intend for this topic model to represent the optimal or definitive partition of arXiv according to subject matter. Rather, the model provides a large set of readily-interpretable topics, varying in both size and specificity, making it possible to test the robustness of past claims against a heterogeneous population of fields and their corresponding authors. We present our analysis of the $k = 50$ topic model below. Note that the results presented here tare robust to small changes in k, meaning that the results reported below do not change significantly if the analyses are repeated using a model with $k = 45$ or $k = 55$ topics.

After training our topic model, we manually inspect each topic to determine whether it resembles a field of condensed matter physics. We consider the highest probability words representing each topic and judge whether those words uniquely describe an established field of condensed matter physics. As an example, the most probable words associated with Topic 28 include keywords such as "dynamic," "glass," "liquid," "temperature," and "relaxation." The set of articles with high probability ($P(\text{Topic} = 5) > 0.6$) of belonging to Topic 28 includes "Evidence of growing spatial correlations during the aging of glassy glycerol" (1209.3401) and "New conserved structural fields for supercooled liquids" (1312.3503). This suggests that articles strongly associated with Topic 28 are related to the physics of glassy systems.

To give a second example, the most probable words associated with Topic 5 include keywords such as "quantum," "state," "qubit," "entanglement," and "decoherence." The set of articles to which the topic model assigns a high probability includes articles such as "Demonstration of Two-Qubit Algorithms with a Superconducting Quantum Processor" (0903.2030) and "Controllable coupling between flux qubits" (cond-mat/0507496). Together, these observations suggest that articles strongly associated with Topic 5 are related to quantum computing and quantum information.

For this latter example, we further check the validity of the trained model by verifying that the articles identified by the topic model do not merely reflect clusters of articles specific to arXiv by inferring topics on the articles belonging to the Web of Science (WoS) data set. The topic model infers that the Web of Science articles "Flexible two-qubit controlled phase gate in a hybrid solid-state system" and "Two-electron coherence and its measurement in electron quantum optics" both belong to Topic 5 with high probability. This confirms that articles associated with Topic 5 appear to be related to quantum computing on both data sets.

In addition to quantum computing and glassy physics, LDA identifies topics resembling other established subfields of condensed matter physics, including spin glasses (Topic 1); Bose-Einstein condensates (Topic 3); magnetic materials (Topic 19); topological phases (Topic 30); and cuprate superconductors (Topic 43). (Refer to Appendix A to see each topic's interpretation.)

The topic model also appears to identify review articles as a group distinguished not by scientific content but by stylistic content. Topic 8 captures standard research terminology and includes words such as "review," "comment," "important," "discuss," and "phenomenon." For this reason, Topic 8 becomes an important point of comparison to contrast the topics that do identify clusters of articles with common scientific themes.

Co-Authorship Network Generation

The topic model is now used to construct a set of co-authorship networks, where each network represents the set of authors that produced the articles strongly associated with one of the topics discovered by the topic model. Note that the topic modeling algorithm is only given information related to the textual content of the articles

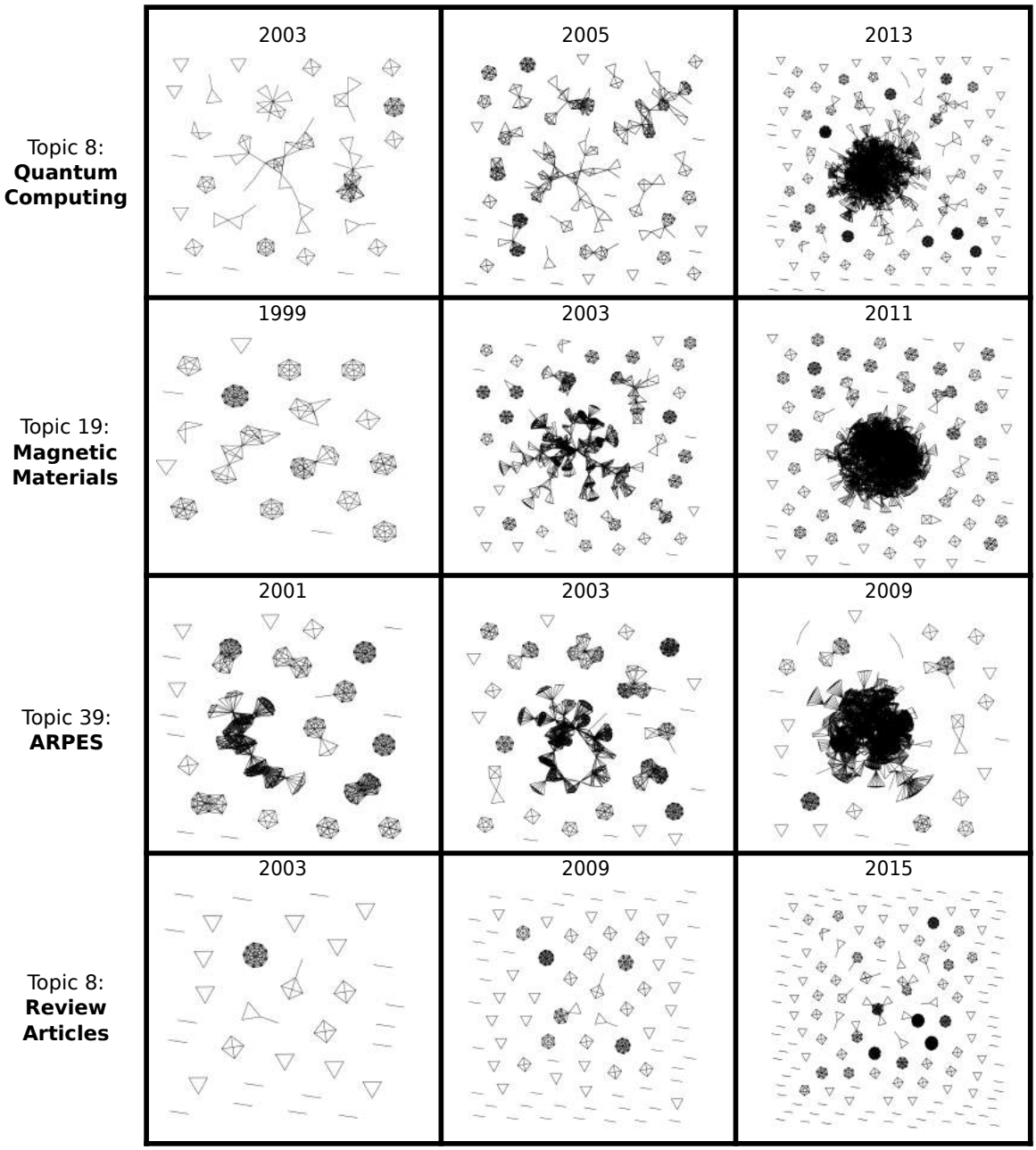

Figure 2.6: **Visualizations of Network Assembly**: Each row shows a co-authorship network's development over time, with network snapshots labeled by the year observed. The three uppermost rows correspond to three different scientific fields, and illustrate the three stages of assembly from a disjointed group of cliques, to a tree-like connected cluster of cliques, to a densely connected giant component that dominates the network. The bottom row corresponds to the review articles, which do not form a giant component.

and receives no information about authorship, authors' collaborative relationships, or publication dates. While there are topic modeling algorithms that do take into

account other links between documents (such as [54, 85]), by using an ordinary topic modeling algorithm it becomes possible to determine whether textual content is sufficient to reproduce patterns in how groups of researchers in the same related form a collaborative community.

The articles that are primarily associated with each topic t are selected by finding the subset of articles assigned a probability weight $P(t) > 0.6$. We chose 0.6 as the threshold in order to select articles that are strongly associated with one particular topic, without making the cutoff so strict that it excludes too many articles. With the cutoff set as $P(t) > 0.6$, each topic contains has between 100 and 3000 arXiv articles. (For the sake of being thorough, we also use an alternative thresholding criterion to check whether the choice of thresholding biases our results, and repeat all subsequent analyses. In this second scheme, each article is assigned to the smallest set of topics that account for 50% of its probability weights across all topics. For example, an article with 40% in Topic 1, 20% in Topic 2, and 10% in Topic 3 would be assigned to Topics 1 and 2. All reported results are robust to varying the thresholding scheme.)

For each topic, the co-authorship network is constructed by identifying the list of authors who contributed to each of the articles associated with the topic. Within the co-authorship network, each node represents an author that has contributed to at least one relevant article. Each edge represents a collaboration between two authors, meaning that they have written at least one article together [75, 77, 78]. Hence, a group of authors who collaborated on an article together appears in the network as a fully connected clique, and two articles with multiple authors in common will appear in the network as overlapping cliques that share nodes. For this reason, the co-authorship networks discussed here have very high clustering coefficients, much higher than for random networks with the same size and degree distributions.

The assembly and growth of each co-authorship network is reconstructed over time using each month of arXiv's operation from April 1992 through June 2015 as a discrete time step. At each time step the network includes all author nodes that have written articles at or prior to the current time step. Each pair of author nodes is connected by a single edge if that pair has collaborated on one or more articles at or prior to the current time step.

2.3.4 Results

Co-Authorship Network Measurements

Figure 2.6 shows three stages of network growth for four different example topics: quantum computing (Topic 5), magnetic material properties (Topic 19), electronic spectra (Topic 39), and review articles (Topic 8). The three first three topics (top three rows) have co-authorship networks that appear to transition from a set of disjointed cliques to a giant connected component. For the review articles (Topic 8, lowest row) very few of the cliques overlap or join together and no giant component forms. This is consistent with the interpretation that the group of "review articles" represents a set of authors writing the same *type* of article, not a group of authors with similar research interests. As such, the authors associated with Topic 8 do not have enough in common with one another to invite collaborations.

For the first three topics in Figure 2.6, there appear to be three separate stages through which the giant component develops. Each network begins as a disjointed set of cliques, as the authors who share a field publish in separate groups. Next, a few of the cliques join together, forming a loosely connected, almost tree-like backbone of connected cliques as authors begin to collaborate across cliques. In the final stage,

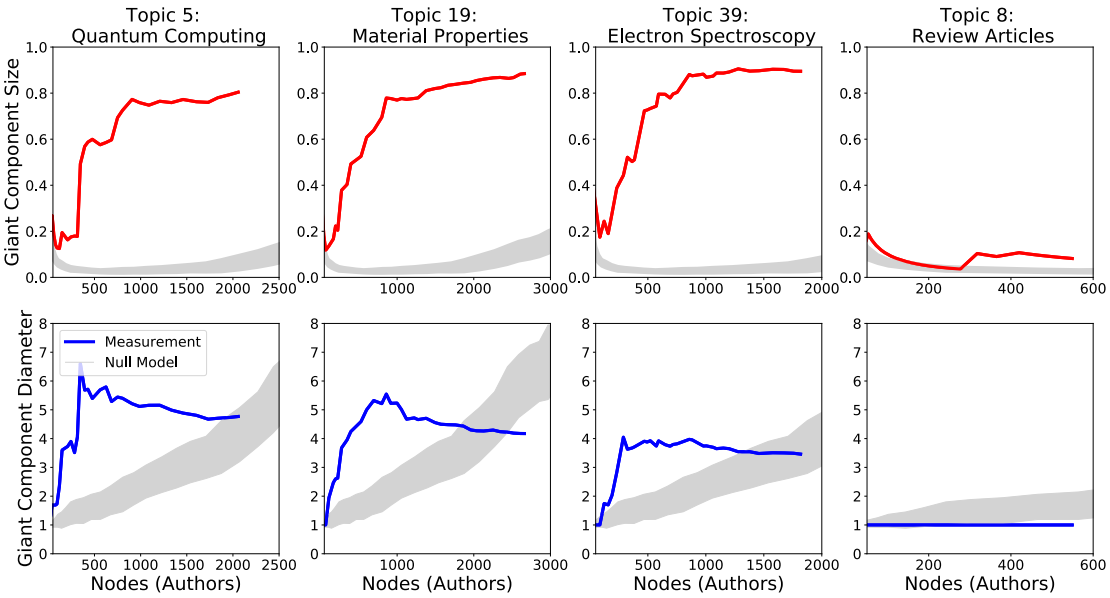

Figure 2.7: **Quantitative measurements of co-authorship networks**: The top row shows the fraction of nodes belonging to the largest component as a measure of network size, plotted vs. the total number of nodes in the network. The bottom row shows the mean geodesic path length of the largest component (diameter) vs. the total number of nodes in the network. The three leftmost columns correspond to three example topics (5, 18, 39) visualized in Figure 2.6. In each of these cases, the relative size of the largest component grows steadily and encompasses a large majority of the nodes. At the same time, the network diameter behaves non-monotonically, first increasing and then decreasing, suggesting that long-range ties are being added to the network. For comparison, the column on the right shows these same measurements for the review articles (Topic 8), which do not form a giant component. The gray region represents the average behavior of a null model that generates co-authorship networks that do not use the LDA topic model to group articles together.

enough cliques overlap with one another such that the largest connected component becomes densely connected. This characteristic three-stage pattern is consistent with what has been reported previously [64].

This interpretation of the network visualizations is quantitatively confirmed by measuring various properties of each topic's co-authorship network. The fraction of nodes belonging to the largest connected component ("giant component size") quantifies the relative size of the largest component. The giant component's mean geodesic

path (network "diameter"), the mean path length between all pairs of nodes belonging to the largest component, quantifies the separation distances between different authors. The diameter ranges between a minimum for fully connected networks and a maximum for treelike networks, and so serves as a measure of how closely connected the individuals belonging to the giant component are to one another [16, 93].

Figure 2.7 shows two measurements of the giant component for each of the co-authorship networks shown previously in Figure 2.6. For Topics 5, 19, and 39 (three leftmost columns), the largest component's size increases steadily as more and more nodes are added to the network. Thus, for each of these topics, the largest component grows to dominate the rest of the network. At the same time, the diameter first increases as the giant component grows initially and then peaks and decreases. The diameter's non-monotonic behavior suggests two stages in the development of the giant component: initial growth as cliques first start to connect to one another, and densification when enough "long-range" edges form to reduce the average distance between authors [64, 79, 93]

These two growth stages are consistent with the growth of a treelike cluster of cliques that becomes a densely connected cluster as more long-range edges form between distant parts of the network. The long-range ties that appear are clearly an important aspect of co-authorship network development, and they may be interpreted as the transfer of researchers between different, previously disconnected research groups who focus on similar topics. It has also been suggested that such long-range ties are created as a result of postdoctoral researchers who transfer between different research groups [16, 58]. The network growth behaviors of Topics 5, 9, and 39 in Figure 2.7 differ from the co-authorship network of the review articles (Topic 8, rightmost column), as no large component forms to connect the mostly unrelated review articles

to one another.

This characteristic development of co-authorship networks is not merely the result of sampling a large number of articles that join together by chance. The gray regions seen in Figure 2.7 show the results of a null model that constructs co-authorship networks from a collection of articles selected uniformly at random from the condensed matter corpus. The shaded gray regions of each plot represent the mean ± 1 standard deviation of 100 instances of the null model. The average behavior of each of these regions contrasts dramatically with the measurements of the scientific co-authorship networks identified using the topic model. Not also that the review articles' co-authorship network behavior is far closer to that of the randomly assembled articles. These results strongly suggest that the aggregation of authors to form a giant, densely connected component is not merely the result of sampling an arbitrary subset of arXiv. Rather, it appears that the topic model, which was given no information about authorship or other such links between documents, was able to identify clusters of researchers based on their textual content alone. The nonrandom grouping of authors further validates the topic model's meaningful clustering of articles: the articles represent the output of an association of researchers with similar interests.

The pattern in the development of the co-authorship networks illustrated in Figure 2.6 and Figure 2.7 characterizes a large number of the co-authorship networks identified by the topic model. 24 topics have co-authorship networks that undergo the transition from a scattered collection of cliques; to an extended, treelike connected group of cliques; to a densely connected giant component. These results are qualitatively consistent with those obtained earlier for groups of articles annotated by human experts [15, 16]. 13 of the remaining topics appear to form a single large component but have not yet formed enough long-range ties that the network diameter

has stopped growing monotonically. The remaining 13 topics show little or no sign that they form any giant connected component. This last group includes the review articles (Topic 8). (Refer to Appendix B for a summary of all co-authorship networks' behavior.)

Finding that a topic's corresponding co-authorship network does not form a densely connected GCC does not necessarily suggest that the research field is not well-established. There are several possible reasons why a dense giant component does not form in all cases. The existence of a giant component only indicates that there are a great many researchers that have collaborated with one another. Inter-group collaborations may be more frequent or larger in some fields than in others, and a giant component is only likely to form when there are a large number of collaborations between research groups. Additionally, the arXiv does not represent a comprehensive sampling of articles from all subfields of science, and its coverage of some fields may be incomplete, such as microscopy (Topic 15) and surface chemistry (Topic 47).

Validation Across Corpora

The characteristic growth patterns seen for the co-authorship networks of authors from arXiv can be shown to be consistent across corpora. The topic model trained on the arXiv data set is employed to infer topics for the condensed matter physics articles from the Web of Science. The same procedures for generating and measuring the co-authorship networks for the Web of Science articles reveals that the topic model trained on the arXiv is still able to identify large connected clusters of articles in the Web of Science. Figure 2.8 compares the behavior of the co-authorship networks that occur within both arXiv and Web of Science.

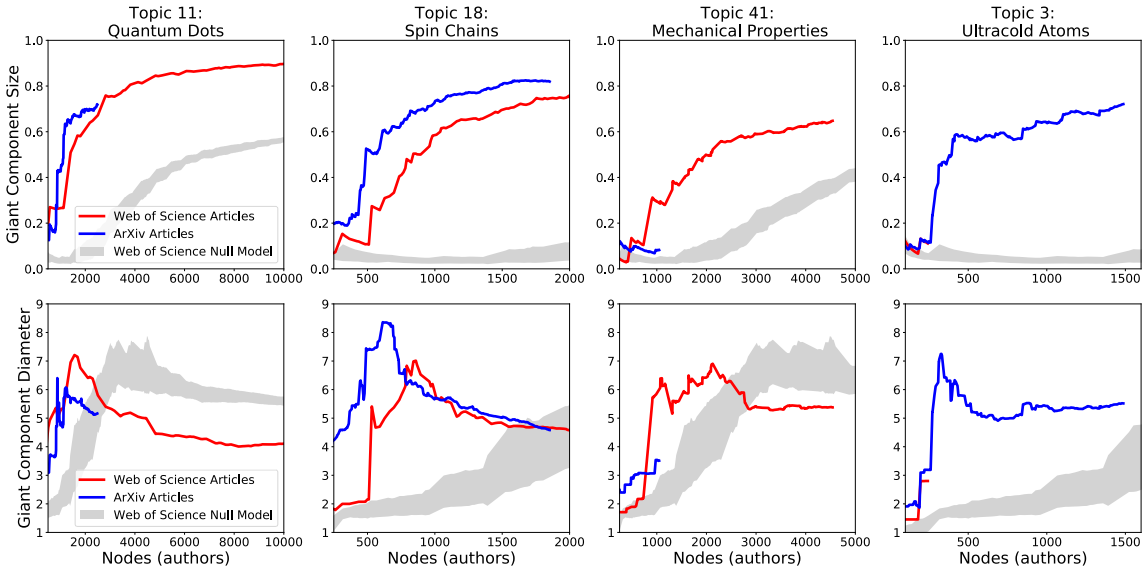

Figure 2.8: **Comparison Between Co-authorship Networks From arXiv and Web of Science**: Each column corresponds to a different topic. The top row shows the fraction of nodes belonging to the largest component as a measure of network size vs. the total number of nodes in the network. The bottom row shows the mean geodesic path length of the largest component, "diameter," vs. the total number of nodes in the network. Each plot shows the measurements made of the co-authorship network from the Web of Science (in red), from arXiv (in blue), as well as co-authorship networks generated from randomly chosen articles from Web of Science (null model, in gray). For 24 topics, the Web of Science co-authorship networks develop similarly as compared to arXiv (e.g. Topic 11 and Topic 18, first and second columns). In 11 cases, the Web of Science co-authorship networks undergo a topological transition even if the arXiv networks do not (e.g. Topic 41, third column). In 8 cases, the Web of Science co-authorship networks fail to develop in the same way as on arXiv (e.g. Topic 3).

In the majority of cases, the co-authorship networks identified from the Web of Science articles behave similarly to the ones identified on arXiv. For example, the co-authorship networks for research on quantum dots and spin chains (Topic 11 and Topic 18, first and second columns of Figure 2.8) form a dense giant component for both arXiv and for Web of Science. In several other cases, there are co-authorship networks that do not undergo a topological transition on arXiv but do for the Web of Science articles. Mechanical properties of materials (Topic 41) is shown in Figure 2.8, but other topics include electronic transport measurements (Topic 12); nanoscale devices (Topic 16) inelastic scattering experiments (Topic 33). That these topics have

an experimental focus, which is noteworthy as experimental research subjects are known to have less coverage on arXiv, but are covered more comprehensively on the Web of Science [63]. There are also a few topics whose corresponding co-authorship networks do transition for arXiv but do not undergo a measurable transition for the Web of Science. For example, the co-authorship network for ultracold atoms (Topic 3, rightmost column of Figure 2.8) contains so few authors that no network forms.

Overall, 34 out of 50 topics have co-authorship networks that behave similarly for the Web of Science data and for the the arXiv data (Appendix B). Additionally, 9 experimentally-focused topics have co-authorship networks have more densely connected giant components on account of having better coverage on the Web of Science compared to arXiv. Another 3 topics (Topics 9, 10, and 42), have very low coverage on the arXiv (fewer than 100 associated articles) and do not form giant connected components with either the arXiv or the WoS. Given that, across both corpora, none of these topics has many strongly associated articles, they may be interpreted as not reflecting the output of a coherent scientific community and so are not useful for the purposes of the present study. The consistency of the behavior of these co-authorship networks measured across different corpora suggests that the collaborative communities identified using the model are reflected in multiple data sets.

Robustness to Edge Removal

Many of the co-authorship networks identified using the topic model form densely connected giant components, but how robust are these results if edges are removed? The co-authorship network development patterns seen in the data are constructed under the assumption that the relationships that edges represent are maintained indefinitely once they are established. Similarly, much of the previous work on co-

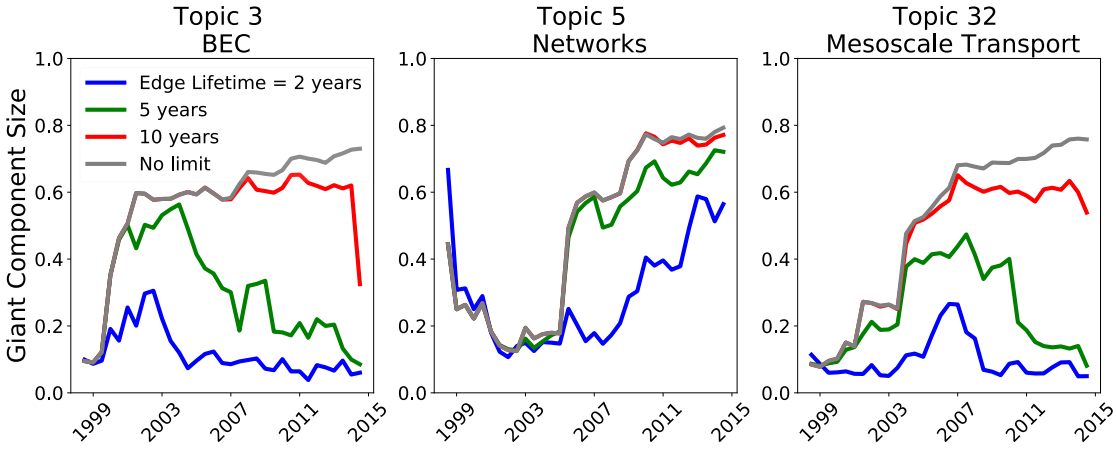

Figure 2.9: **Network Robustness to Edge Removal**: Each plot shows how the network assembly changes when edges only remain in the network for a limited amount of time. Each plot shows the network's giant component size over time for four different edge lifetimes. For short edge lifetimes (2 years in blue; 5 years in green), the giant connected component fails to develop or develops much more slowly compared to the permanent edge ("no limit," gray) case. For longer edge lifetimes (10 years, red), the giant component approaches the no limit case.

authorship networks assumes that collaborative ties, once established, are maintained forever [15, 16, 64]. In practice, when such a collaborative relationship requires real effort to maintain, this assumption is not necessarily valid.

Each topic's co-authorship network is re-assembled, this time allowing edges to expire after a fixed number of months. That is to say, if two authors do not repeat a collaboration after a certain amount of time, the edge representing their relationship is removed from the network.

In Figure 2.9, the uppermost curve (gray; "no limit") shows how the giant component grows if edges survive indefinitely, while the lower curves show how those measurements change if the edges are removed after 2 (blue), 5 (green), or 10 (red) years. For short edge lifetimes, edges are removed relatively quickly after they are added, meaning that the co-authorship network is more likely to fall apart. Each of the three example topics in Figure 2.9 forms a densely connected giant component if

edges are never removed, but shortening the lifetime of edges to a few years causes the giant component to fall apart or delays the amount of time before the component forms. In some cases (such as for the field studying networks, Topic 5, middle column of Figure 2.9), the network measurements for 5 and 10 years are very close to the indefinite lifetime limit. This suggests that this co-authorship network is particularly robust to edge removal, reflecting a very densely connected giant component where edges are frequently renewed[64].

Currently, it is unknown what criteria for including and excluding nodes and edges from co-authorship networks best reflect the reality of authors entering and exiting different fields. What is clear, however, is that the assumption that the relationships represented by edges between authors last forever is important for obtaining the quantitative results that reflect a topological transition in the co-authorship network. Shortening the lifetime of edges can dramatically change a co-authorship network's evolution over time.

2.3.5 Discussion

This study expands upon previous research exploring the growth and development of co-authorship networks using topic modeling to algorithmically identify and study a large population of scientific fields, along with their associated articles and authors. The results show that a majority of the algorithmically identified co-authorship networks undergo a topological transition to form a densely-connected giant component characterized by three stages of development. These patterns corroborate findings from earlier studies that focused on small numbers of (often manually assembled) co-authorship networks. This suggests that the characteristic topological transition is robust to variations in the definition of a scientific field, in terms of both size and

specificity. Additionally, this methodology employ algorithmic clustering and require little input from human experts, yet the results are largely consistent with previous studies.

Additionally, the patterns in co-authorship network development are consistent across corpora, which is demonstrated by repeating the analysis using data from both arXiv and the Web of Science. One notable difference between the two corpora is reflected in how arXiv's selections of articles related to certain experimentally-focused topics are under-populated: in these cases, the co-authorship networks drawn using the arXiv data are not consistent with the larger Web of Science data set. For the other topics, however, the arXiv contains co-authorship networks that do appear to sufficiently sample and qualitatively represent the full collaborative communities.

First Pass in Topic Modeling

Topic modeling is a rich and actively growing area of research within the statistical modeling and natural language processing communities. The present methodology employs Latent Dirichlet Allocation, one of the most popular yet simplest forms of topic modeling. This model assumes a static definition for topics and thus scientific communities, which are known evolve with time. Additionally, the model does not directly incorporate other, non-semantic relationships between documents (such as co-authorship or citations), which may signal alternate forms of cohesion within a scientific community. Future work in this area, however, should explore more sophisticated algorithms that consider topic dynamics (e.g. [19, 92]) and additional measures of community cohesion in order to more thoroughly address the co-evolution of scientific fields.

Contributions to scientometric studies

This method for algorithmically generating and analyzing a large number of fields can also be used as a framework for further exploring the claims made in a wide variety of bibliometric contexts. For example, one could also perform a comparison of the micro-scale dynamics of individual authors many different fields. Recent studies have used agent-based models of author behavior to explain the patterns in publishing behavior that one sees in different fields of science (e.g. [24, 90]). Again, most of these studies have relied on manually annotated data sets, and as such, they have historically been limited to only a handful of fields. The approach that developed in this study, however, enables future work, in conjunction with comprehensive data sets like the arXiv or Web of Science, to further test the accuracy of these models of author behavior across a large and diverse population of scientific fields.

CHAPTER 3

PERSISTENCE AND STOCHASTIC EXTINCTION OF INFECTIOUS DISEASES ON NETWORKS

3.1 Introduction

Infectious diseases are spread by the passing of a disease-causing pathogen between individual hosts. This transmission may occur through direct contact (e.g. measles, HIV, influenza) or indirectly as mediated through the environment or through a host vector (e.g malaria, Lyme disease) [8, 60, 67]. The field of infectious disease modeling strives to use mathematical models to understand how epidemics progress through host populations.

Of course, in practice it is the biomedical expertise of doctors and other public health workers who discover and implement vaccinations and other interventions that reduce the rate of infection, morbidity, mortality from infectious disease. Where modeling is useful is that it can perform population-level experiments *in silico* for the purposes of forecasting the progression of an epidemic, or for evaluating the expected efficacy of different strategies for combating an outbreak of disease [11, 34]. To give one example, Bozzette et al. use simulations to model an outbreak of smallpox and to test the potential trade-offs between the benefits of vaccination against the potential harm caused by smallpox vaccine side effects for different vaccine deployment strategies [25]. Additionally, mathematical models of disease dynamics can provide analytical insight into how certain parameters affect the outcome of an epidemic [60].

One challenge facing disease modeling is to understand the phenomenon of en-

demic disease. A disease is considered endemic if it persists in a population over a long period of time rather than dying out following a single outbreak. To assist public health workers in combating endemic diseases, there are many important questions that can be answered using modeling tools. What factors lead some populations to be able to sustain endemic disease for long periods of time? Can the endemic disease be expected to go extinct spontaneously? If so, for how long will it persist before extinction? It also remains an open question of how population heterogeneity influences the persistence of endemic disease. In this context, network models of contacts between individual hosts are particularly useful, as networks provide a natural framework for modeling a community where some hosts have many contacts while other hosts have very few.

In the present chapter, we focus on one model of endemic disease, the stochastic SIRS model, and explore the answers to these questions. Using a combination of analytical methods and computer simulations, we study the statistical properties of the endemic disease state. We extend our analysis to include contact network heterogeneity in the population, in order to discover the effect that a population's heterogeneity has on the persistence of endemic disease. In particular, we highlight the relationship between the statistical properties of the endemic disease state and the characteristic persistence lifetime of the endemic disease state, and show that changing the network heterogeneity does not appear to change this relationship.

3.2 Infectious Disease Modeling

3.2.1 History of Mathematical Disease Modeling

Mathematical models have been used to understand infectious disease dynamics for almost a century [62]. The SIR compartmental model goes back nearly a century [62], and has proven to be useful for understanding some key aspects of disease modeling - first became aware of threshold effects where small changes in parameters can lead to large change in disease outcome - large and small epidemics

Simple mathematical models of disease dynamics have proven to be versatile and useful across many types of pathogens [8, 60, 67]. Additionally, they may also be used to model other types of non-epidemiological processes, such as the spread of computer viruses [61, 82, 83].

3.2.2 Compartmental Models

Population-level modeling of infectious disease dynamics begins with the simplifying assumption that individual hosts can be found in one of a possible set of disease states. In the most basic formulation begins hosts begin **susceptible** to infection, as in no pathogen is yet present. Hosts are considered **infected** once they have contracted the disease and are capable of spreading it to others. Hosts are considered **recovered** once the host is no longer infected - the disease has run its course and they are no longer infectious, or their immune system has cleared them of infection, or they have died - in this stage, hosts are no longer susceptible to the disease and are no longer able to spread it to others. [60]. At the population model, individuals are grouped

together according to their disease state, such that the population is divided into disease state "compartments."

SIR-type compartmental models of disease dynamics have proven useful in that they are highly versatile. One may adapt such a model by further subdividing the population and adding additional compartments to more closely reflect the specific population or type of infectious disease that one wants to study [60].

Another advantage of compartmental models is that they can provide analytical insight into the complicated, nonlinear problem of how infectious disease spreads through a population. In this way they are more effective than agent-based models, which track individual hosts as they move around their environment. Agent-based models are able to incorporate a great deal more details about the behavior and interactions of individual hosts, but they also may require an intractable number of assumptions about parameters. In the end, a very detailed agent-based model provides little analytical insight. Unlike in compartmental models, it can be difficult to determine how the outcome of the agent-based model depends on the input parameters without extensive simulations [11, 60].

The following section will review the essential properties of the most basic versions of SIR-type compartmental models.

SIR model

The Susceptible-Infected-Recovered model is used to describe the spread of acute infections, such as influenza and chicken pox, that leave the host immune to future infection [60]. In their pioneering paper, Kermack and McKendrick used this model to fit to data of an outbreak of plague in Bombay. They found that the dynamics of

the SIR model were useful for describing the increase and subsequent decrease in the number of reported cases (number infected) that constituted the outbreak of disease [62].

To describe this model mathematically, let N be the population size, and let (X, Y, Z) be the number of susceptible, infected, and recovered individuals in the population. Let $(S, I, R) = (X/N, Y/N, Z/N)$, the respective fractions of susceptible, infected, and recovered individuals. For constant population, $S(t) + I(t) + R(t) = 1$ for all $t > 0$. The SIR model may be expressed with a set of ordinary differential equations describing the time evolution of (S, I, R):

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{3.1}$$

The term $\beta Y/N = \beta I$ is the force of infection, or the per capita rate at which susceptible individuals become infected through contact with their infected neighbors. Thus, the nonlinear term (βSI) reflects the transmission of infection to susceptible individuals. Infected individuals are assumed to recover at a constant rate γ. [60, 11]. The initial conditions of Eq. 3.1 begin with zero recovered individuals, the majority of the population susceptible, and a few infected individuals.

The deterministic SIR model cannot be solved analytically, but it is possible to determine a condition in which a large outbreak can occur through some analysis. Dividing the first equation in Eq. 3.1 by the third equation yields a new differential equation for S parametrized by R:

$$\frac{dS}{dR} = -\beta/\gamma S \tag{3.2}$$

Integrating with respect to R yields:

$$S(t) = S(0)e^{(-\beta/\gamma R(t))} \tag{3.3}$$

It is now possible to solve for the total number of individuals who are affected by the epidemic and pass through the infected state into the recovered state. As $t \to \infty$, there are no more infected individuals left in the population, so $1 = S(t \to \infty) + R(t \to \infty)$. Taking the $t \to \infty$ limit in Eq. 3.3 yields a transcendental equation that relates $R(t \to \infty)$ to the combination of parameters β/γ:

$$1 - R(t \to \infty) = S(0)e^{(-R_0 R(t \to \infty))} \tag{3.4}$$

Fig. 3.1 shows the numerical solutions to Eq. 3.4. $R(t \to \infty) = 0$ when $\beta\gamma < 1$, and $R(t \to \infty) > 0$ when $\beta\gamma > 1$ [11, 55, 60, 79]. Here, we can define $R_0 \equiv \beta/\gamma$, which is also known as the "basic reproductive ratio." In an epidemiological context, R_0 is the average number of secondary cases caused by the introduction of a single infected individual over its infection period: γ^{-1} is the period of infection, while β is the rate per contact at which neighbors become infected through contact.

Numerical integration of Eq. 3.1 yields the solutions shown in Fig. 3.2. Note that above the epidemic transition $R_0 > 1$, the number of infected individuals grows and then dies out, such that that a non-zero fraction of individuals pass through the infected state into the recovered state.

SIS model

The Susceptible-Infected-Susceptible model describes the spread of infections that do not impart lasting immunity, such as gonorrhoea [51, 60]. After being infected, individuals return to the susceptible state. Because the number of susceptible individuals

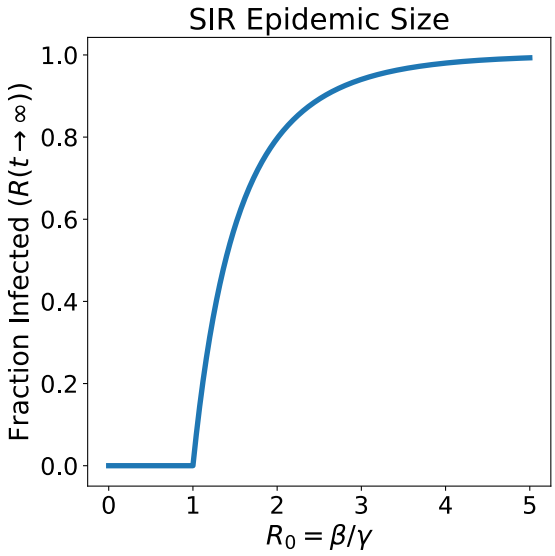

Figure 3.1: **SIR Epidemic Transition**: Numerical solutions to Eq. 3.4, showing the fraction of individuals affected by an epidemic as a function of $R_0 \equiv \beta/\gamma$. Below $R_0 = 1$, there is no outbreak, but above $R_0 = 1$ the outbreak grows to affect a nonzero fraction of the full population.

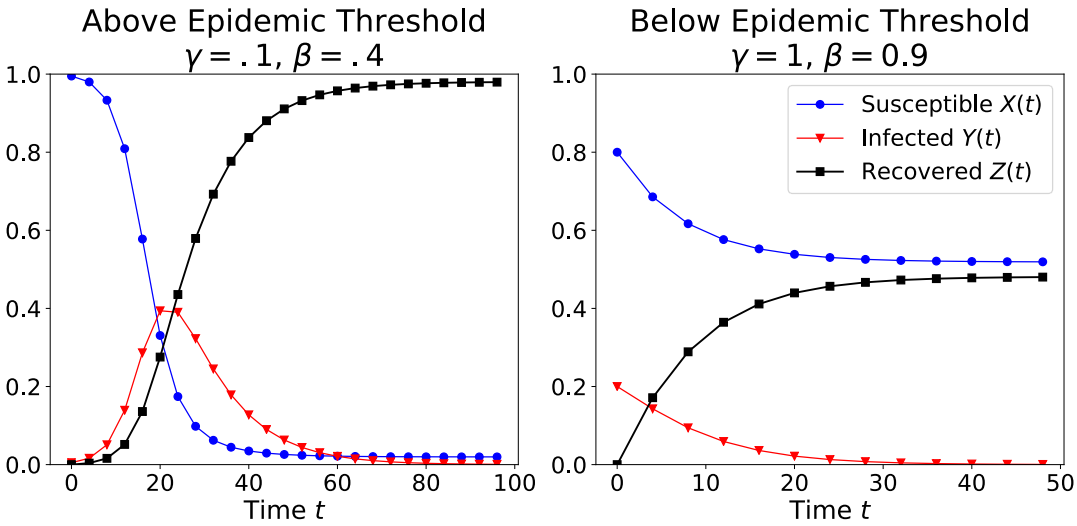

Figure 3.2: **Deterministic SIR Dynamics**: Left hand plot shows the solution to the deterministic SIR equations (Eq. 3.1) above the epidemic threshold ($R_0 = \beta/\gamma = 4$). Note the peak in the number of infected individuals, representing the outbreak of disease. Almost, but not all of the initially susceptible individuals are affected by the infection and end in the recovered state. Right hand plot shows the solution to the deterministic SIR equations below the epidemic threshold ($R_0 = \beta/\gamma < 1$), where the number of infected individuals quickly dies out before it can affect the majority of the population.

is constantly replenished, the deterministic SIS model can produce an endemic state where the number of infected individuals remains finite and never dies out.

Using the same notation as for the SIR model, the SIS model may be expressed as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI + \gamma I \\ \frac{dI}{dt} &= \beta SI - \gamma I \end{aligned} \tag{3.5}$$

These two equations may be simplified into a single one-dimensional ODE:

$$\frac{dI}{dt} = \beta(1 - I)I - \gamma I \tag{3.6}$$

Setting the left hand side of Eq. Eq. 3.6 to zero yields the number infected in the long-term steady-state [11, 60]:

$$\begin{aligned} I^* &= 1 - \gamma/\beta = 1 - R_0^{-1}, \qquad \beta > \gamma \\ I^* &= 0, \qquad \beta < \gamma \end{aligned} \tag{3.7}$$

Again, the solutions depend on the constant $R_0 = \beta/\gamma$. Linear stability analysis [89] of the Eq. 3.7 shows that for $R_0 > 1$ there is a finite number of infected individuals in the steady-state. This is interpreted as an endemic disease state, where the disease is continually spreading to individuals after they recover from the infection such that there remains a finite amount of infection. This contrasts to the behavior of the SIR model, in which there is a single outbreak of infection that dies out. For $R_0 < 1$, the number of infected individuals drops to 0. Numerical solutions to Eq. 3.5 are shown in Fig. 3.3.

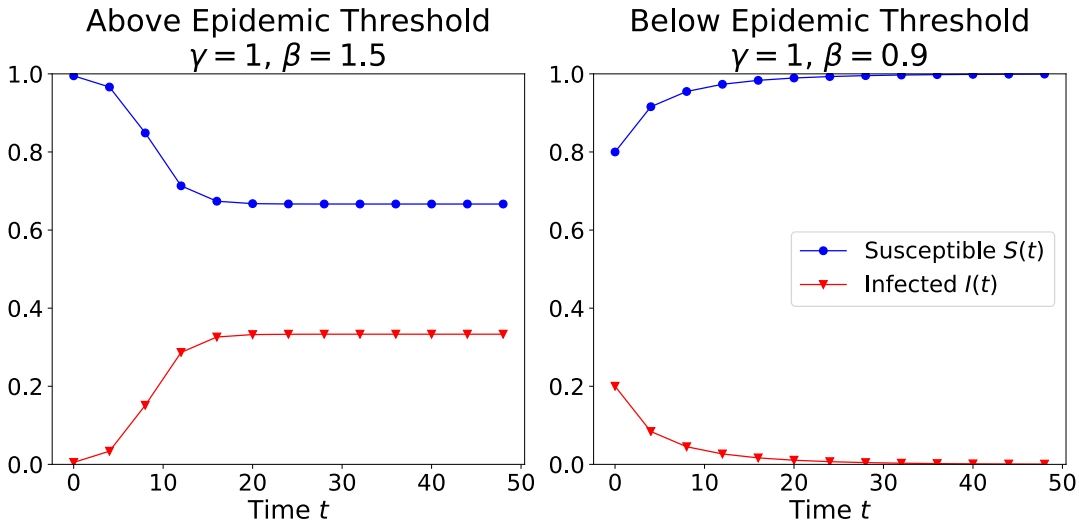

Figure 3.3: **Deterministic SIS Dynamics**: Left hand plot shows the solution to the deterministic SIR equations (Eq. 3.5) above the endemic threshold ($R_0 = \beta/\gamma = 1.5$). Note how the number of infected individuals converges to a fixed value, and then remains at that value. This behavior represents the endemic state, where a finite number of individuals remain infected indefinitely. Right hand plot shows the solution to the deterministic SIS equations below the endemic threshold $R_0 = \beta/\gamma < 1$, where the number infected dies out rather than reach an endemic state.

SIRS model

The SIR model with waning immunity, or SIRS model, describes a disease in which recovered individuals lose immunity over time, meaning that infected individuals are temporarily recovered before becoming susceptible again. SIRS can be used to describe epidemics in which immunity is lost over time [80], for example because the disease-causing pathogen evolves quickly enough that hosts' immune systems no longer responds to it [7, 50].

The SIRS model combines elements from both the SIR model and SIS model [39]. Similar to the SIR model, when individuals recover they enter the recovered state and can no longer be infected. Effectively, they are removed from the population during this stage. Similar to the SIS model, however, individuals are eventually recycled

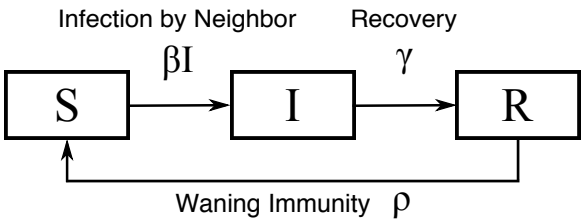

Figure 3.4: **Schematic of SIRS model**: Individuals begin in the susceptible state. Through contact with infected individuals, they become infected. Over time, they recover and acquire immunity. Once that immunity is lost, they are returned back into the susceptible state.

back into the susceptible state because the acquired immunity is not permanent.

The deterministic SIRS model (transition schematic shown in Fig. 3.4) can be described with the following ODEs, where ρ is the rate of waning immunity.:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI + \rho R \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I - \rho R \end{aligned} \tag{3.8}$$

Note that in the limit that $\rho = 0$, Eq. 3.8 reduce down to Eq. 3.1 [39].

Assuming the population is constant over time ($S+I+R = 1$), the three equations can be reduced down to two:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI + \rho\left(1 - S - I\right) \\ \frac{dI}{dt} &= \beta SI - \gamma I \end{aligned} \tag{3.9}$$

The steady-state behavior of Eq. 3.9 is found by setting the left hand side to zero. Similar to the endemic state in the SIS model, the nontrivial solution, with nonzero infected individuals, is stable above the endemic threshold $R_0 = \beta/\gamma > 1$.

$$\begin{aligned} S^* &= \gamma/\beta = R_0^{-1} \\ I^* &= \frac{\rho}{\rho + \gamma}\left(1 - R_0^{-1}\right) \\ R^* &= \frac{\gamma}{\rho + \gamma}\left(1 - R_0^{-1}\right) \end{aligned} \tag{3.10}$$

Below the endemic threshold, the solution $(S, I) = (1, 0)$ is stable. The endemic state solution (Eq. 3.10) is very similar to that of the SIS model (Eq. 3.7), except that the endemic level is now modulated by an additional factor that depends on the rate of waning immunity ρ. For $\rho > \gamma$, this factor is large and approaches 1 as ρ increases, meaning that for a shorter time spent immune the population has a higher number infected in the endemic state [33]. For $\rho < \gamma$, this factor becomes small and suppresses the endemic level.

Further linear stability analysis also shows that unlike the 1-dimensional SIS model, the 2-dimensional SIRS model can show damped oscillations as it converges towards the endemic state [51, 60]. These occur above the endemic threshold ($R_0 > 1$) when $4(R_0 - 1)(1 + \rho/\gamma)^2 > \rho/\gamma(R_0 + \rho/\gamma)^2$, or for high values of R_0 and low values of ρ/γ as in the upper right-hand corner of Fig 3.5 D.

Further analysis of the SIRS model's endemic state will be the focus of the remainder of this chapter.

3.3 Stochastic Models of Disease Dynamics

3.3.1 Recurrent Epidemics and Spontaneous Extinction

Although deterministic models capture many essential aspects of infectious disease dynamics, they do fail to reflect other important empirically observed features. Mid-century epidemiological studies of measles showed that outbreaks could occur repeatedly [13]. For each outbreak, the number of reported cases would increase and then die off as in a single SIR outbreak. These outbreaks occurred and re-occurred re-

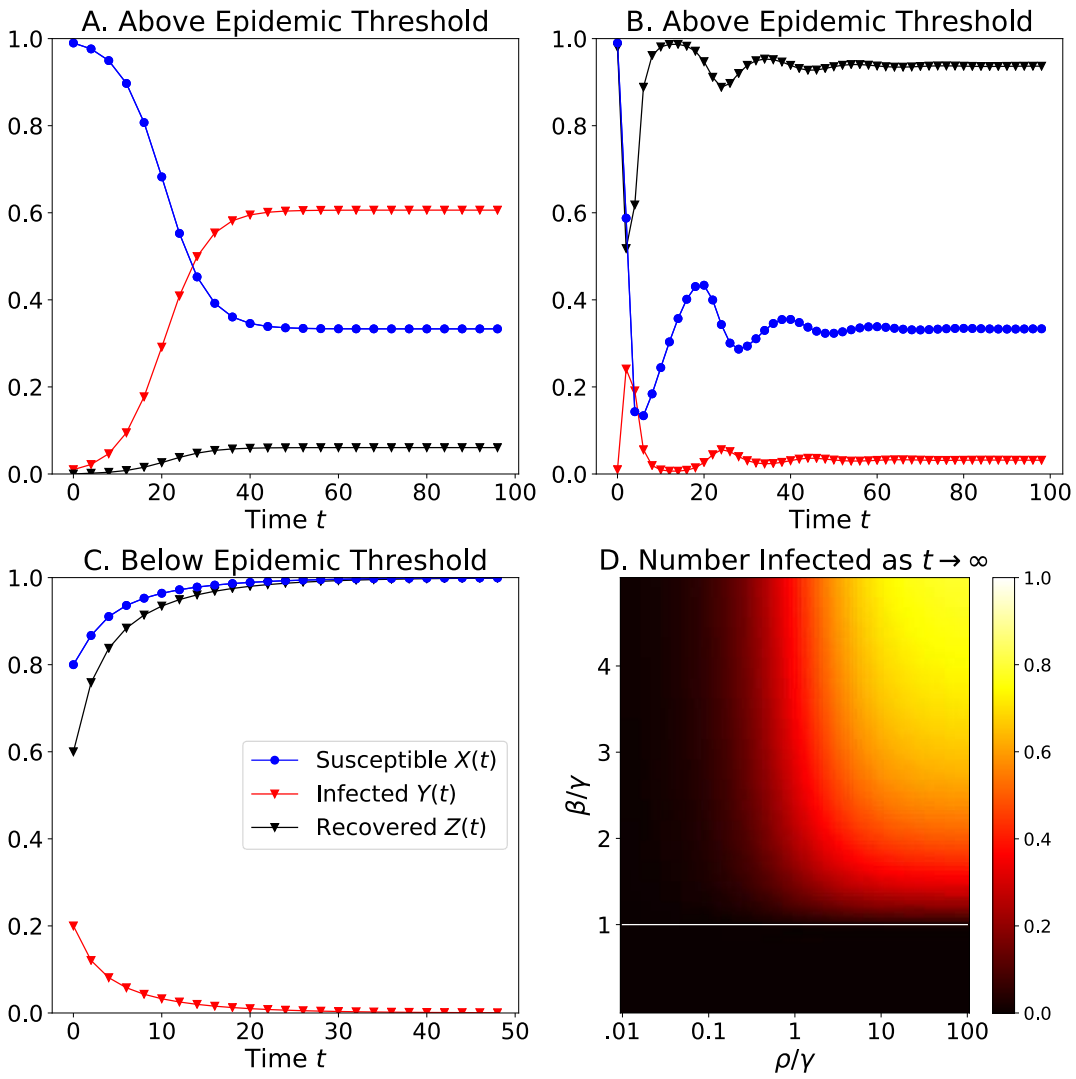

Figure 3.5: **Deterministic SIRS Dynamics**: A. The solution to the deterministic SIR Sequations (Eq. 3.1) above the endemic threshold ($\beta = .4$, $\gamma = .1$, $\rho = 1.$). The model's behavior is very similar to that of the SIS model, where the trajectories converge to an endemic state with the number of infected individuals remaining finite for all time. B. The solution to the deterministic SIRS equations above the endemic threshold ($\beta = 3.$, $\gamma = 1.$, $\rho = 0.05$), this time with ρ chosen such that the damped oscillations appear. Again, after the oscillations are damped away, the trajectories still converge to an endemic state. C. The solution to the deterministic SIRS equations below the endemic threshold ($R_0 = \beta/\gamma < 1$). D. Heat map showing the endemic infection level (Eq. 3.10) for varying values of parameters β/γ and ρ/γ. Note that for $\beta/\gamma < 1$ (below the white line) the number infected die out and the endemic level is 0. Above the endemic threshold $\beta/\gamma = 1$ there is always a finite amount of infection remaining in the population, although the endemic infection level is much higher for $\rho/\gamma > 1$.

peatedly over time. From a modeling perspective, the deterministic models used for measles could not account for the apparent extinction and recurrence of infection.

To solve this problem, Bartlett proposed that stochastic disease models could be used to describe the problem of recurrent epidemics, as stochastic models do include a mechanism for the spontaneous extinction of an outbreak of disease [13]. In particular, Bartlett observed, that outbreaks of measles were more likely to die out spontaneously in small and isolated populations than in large populations [12, 13]. In particular, there appeared to be a relationship between the size of a community and the duration of the disease outbreak, leading to the notion of a critical community size. Above the critical community size, infection was likely to remain endemic in the population, while below the critical community size, infection was likely to die off spontaneously within a few years. This critical community size could be measured and verified for measles across many different communities [12, 14, 18]. Furthermore, stochastic versions of models of disease dynamics were able to reproduce the spontaneous extinction and recurrence of disease outbreaks [12].

In this context, the properties of stochastic models provide a key insight into the spontaneous extinction of disease. Even if a deterministic model predicts that a disease should remain endemic in a population forever, stochastic models predict that spontaneous extinction is possible. A stochastic model cannot just be thought of as a deterministic model with noise added. Rather, adding stochasticity to models of infectious disease dynamics adds important features that deterministic models cannot reproduce.

3.3.2 The SIRS Master Equation

Randomness is introduced into the SIRS model by treating it as a stochastic process that describes the behavior of an ensemble of trajectories. Instead of expressing the model as a system of deterministic ODEs, the stochastic SIRS model is defined as a continuous time Markov chain with three types of transitions: infection, in which an susceptible individual becomes infected; recovery, in which a recovered individual becomes recovered; and loss of immunity, in which a recovered individual becomes susceptible again. These transitions are expressed mathematically in Table 3.3.2, with (m, n) as the number of susceptible and infected individuals respectively. Again, the transitions depend on the parameters β, γ, and ρ.

Event	Transition	Rate
Infection	$(m, n) \longrightarrow (m-1, n+1)$	$\beta mn/N$
Recovery	$(m, n) \longrightarrow (m, n-1)$	γm
Loss of Immunity	$(m, n) \longrightarrow (m+1, n)$	$\rho\,(N-m-n)$

Table 3.1: **Transitions in Stochastic SIRS Model**: m is the number of susceptible individuals, and n is the number of infected individuals.

This stochastic process can be simulated using Gillespie's Direct algorithm [60, 41]. Fig. 3.6 shows a comparison between the output of the deterministic SIRS model (from numerically integrating Eq 3.9) and a stochastic simulation. While the number of infected individuals in the deterministic model converges to and remains at the endemic level, the number of infected individuals in the stochastic model fluctuates about that endemic level. Eventually, the fluctuations lead to a spontaneous extinction, which is not predicted by the deterministic model. The spontaneous extinction seen here is an example of the phenomenon that originally motivated Bartlett to use

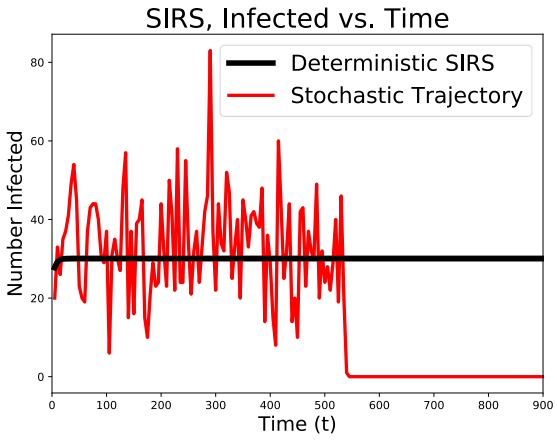

Figure 3.6: **Spontaneous Extinction**: A comparison between the output of the deterministic and stochastic versions of the SIRS model. While the number of infected individuals in the deterministic model converges to and remains at the endemic level, the number of infected individuals in the stochastic model fluctuates about that endemic level. Eventually, the fluctuations lead to a spontaneous extinction, which is not predicted by the deterministic model.

stochastic models to explain recurrent epidemics [12, 13].

It is also possible to gain some intuition from Fig. 3.6 for how spontaneous extinctions depend on the properties of the endemic state. The stochastic trajectory fluctuates about a mean endemic level. The likelihood of a spontaneous extinction occurring depends on the relative size of the fluctuations compared to the mean endemic level: if the fluctuations are small compared to the mean, then spontaneous extinctions are less likely to take place. The following analysis of the stochastic SIRS model will analyze the behavior both of the mean endemic level as well as the characteristic fluctuation sizes in order to estimate the rate at which spontaneous extinctions occur.

Using these transition rates, it becomes possible to derive a master equation (Kolmogorov forward equation) describing the behavior of the probability distribution of

trajectories $p(t) = \mathbb{P}(X(t) = m, Y(t) = n)$ [10, 33]:

$$\begin{aligned}\frac{\partial}{\partial t}p_{m,n} &= \beta(m+1)(n-1)/N p_{m+1,n-1}(t) \\ &+ \gamma(n+1)p_{m,n+1}(t) + \rho(N-(m-1)-n)p_{m-1,n}(t) \\ &- (\beta mn/N + \gamma n + \rho(N-m-n))p_{m,n}(t)\end{aligned} \tag{3.11}$$

The initial conditions in Eq. 3.11 start with n_0 infected and $N - n_0$ susceptible, so $p(t=0) = \delta_{m,N-n_0}\delta_{n,n_0}$.

Quasi-static Distribution

To understand the behavior of Eq. 3.11, it is convenient is to rewrite the probability distribution in terms of its quasi-static distribution (QSD) [33, 73]. The SIRS model includes an absorbing state at $(m, n) = (N, 0)$. When the number of infected individuals goes to zero (a spontaneous extinction), the trajectory can never leave this point. Focusing analysis on the QSD is the same as focusing on the ensemble of trajectories that are active and have not yet reached the absorbing state [36]. Let the QSD for p(t) be $q(t) = \mathbb{P}(X(t) = m, Y(t) = n | Y(t) \neq 0)$, meaning that it is a probability distribution that conditions on $Y(t) > 0$:

$$q_{m,n}(t) = \frac{p_{m,n}(t)}{1 - p_{\cdot,0}(t)} \tag{3.12}$$

where the dot notation denotes the marginal probability of having zero Infected at time t: $p_{\cdot,0} = \sum_{m=0}^{N} p_{m,0}(t)$.

Setting $n = 0$ and summing equation 3.11 over all m yields an expression for the rate at which active trajectories transition into the absorbing state [33, 74, 73]:

$$\frac{\partial}{\partial t}p_{\cdot,0}(t) = \gamma p_{\cdot,1}(t) \tag{3.13}$$

Taking the time derivative of Eq. 3.12:

$$\frac{\partial}{\partial t}q_{m,n} = \frac{1}{1-p_{\cdot,0}\left(t\right)}\left(\frac{\partial}{\partial t}p_{m,n}\left(t\right) + \frac{\partial}{\partial t}p_{\cdot,0}\left(t\right)q_{m,n}\left(t\right)\right) \tag{3.14}$$

Combining the time derivative of the QSD Eq. 3.14 with the master equation for the probability distribution (Eq. 3.11) and Eq. 3.13 yields the master equation for the QSD, conditioning on the trajectories remaining active and avoiding the absorbing state:

$$\begin{aligned}\frac{\partial}{\partial t}q_{m,n} &= \beta\left(m+1\right)\left(n-1\right)/N q_{m+1,n-1}\left(t\right) \\ &+ \gamma\left(n+1\right)q_{m,n+1}\left(t\right) + \rho\left(N-\left(m-1\right)-n\right)q_{m-1,n}\left(t\right) \\ &- \left(\beta mn/N + \gamma n + \rho\left(N-m-n\right)\right)q_{m,n}\left(t\right) \\ &+ \gamma q_{\cdot,1}\left(t\right)q_{m,n}\left(t\right)\end{aligned} \tag{3.15}$$

Note that this master equation is very similar to that of the probability distribution except for a single nonlinear term ($\gamma q_{\cdot,1}\left(t\right)q_{m,n}\left(t\right)$). This nonlinear term represents the fact that some trajectories are leaving the QSD by entering the absorbing state, and is smaller than all other terms that appear in Eq. 3.15. Assuming that $\gamma q_{\cdot,1}\left(t\right) \ll 1$ is the same as assuming that the rate of spontaneous extinction is small, and that the QSD remains stable over long periods of time [33].

3.3.3 Cumulant Equations

Eq. 3.15 is very complicated and it is very difficult to solve for $q(t)$ exactly, but it is possible to solve approximately for the cumulants of the QSD. This is accomplished by using a change of variables to express the QSD in terms of its cumulants, and then by truncating the expansion to obtain a set of coupled ODEs that may be solved and analyzed using standard methods.

We define a probability generating function (PGF) for the QSD [10, 94]:

$$P(x,y,t) \equiv \sum_{m,n=0}^{\infty} q_{m,n}(t)\, x^m y^n \tag{3.16}$$

Multiplying both sides of equation 3.15 by $x^m y^n$ and summing over all m, n yields the following partial differential equation, with the initial condition given by $P(x,y,0) = x^{(N-n_0)} y^{n_0}$:

$$\begin{aligned} \frac{\partial P}{\partial t} &= \beta\left(y^2 - xy\right)\frac{\partial^2}{\partial x \partial y} P(x,y,t) \\ &+ \gamma(1-y)\frac{\partial}{\partial y} P(x,y,t) \\ &+ \rho(x-1)\left(N - x\frac{\partial}{\partial x} - y\frac{\partial}{\partial y}\right) P(x,y,t) \end{aligned} \tag{3.17}$$

Eq. 3.17 has no known solution (and in fact cannot be solved because of incompletely-defined boundary conditions), but it can be simplified by performing a change of variables and expressing the probability generating function in terms of its cumulants. Letting $x \equiv e^{\theta}$ and $y \equiv e^{\phi}$, the moment generating function is $M(\theta,\phi,t) = P(x,y,t)$, and the cumulant generating function is $K(\theta,\phi,t) = \log\left(M(\theta,\phi,t)\right)$ [10, 94].

At this point, we use a simplifying assumption for the QSD with a bivariate Gaussian distribution with means $(\mu(x), \mu(y))$ and variances $(\sigma(xy), \sigma^2(x), \sigma^2(y))$. All higher-order cumulants are assumed to be zero. (There are other sophisticated assumptions that one may employ for simplifying the cumulant expansion, but in this context assuming a Gaussian distribution gives sufficient understanding of the model behavior.)

$$K(\theta,\phi,t) = \mu(x)\theta + \mu(y)\phi + \sigma(xy)\theta\phi + \frac{1}{2}\sigma^2(x)\theta^2 + \frac{1}{2}\sigma^2(y)\phi^2 \tag{3.18}$$

Each of these cumulants of the probability distribution depends on time (e.g. $\mu(x) = \mu_x(t)$), allowing the quasi-stationary distribution to change over time [33].

Applying these changes of variables to Eq. 3.17 and collecting terms in powers of θ and ϕ yields a set of nonlinear ODE's for each of the cumulants.

$$\begin{aligned}
\frac{\partial}{\partial t}\mu(x) &= \rho\left(N-\mu(x)-\mu(y)\right)-\beta\left(\sigma(xy)+\mu(x)\mu(y)\right)/N \\
\frac{\partial}{\partial t}\mu(y) &= -\gamma\mu(y)+\beta\left(\sigma(xy)+\mu(x)\mu(y)\right)/N \\
\frac{\partial}{\partial t}\sigma(xy) &= -\beta\left(\mu(x)\mu(y)+\sigma(xy)\right)/N-\gamma\sigma(xy)-\rho\left(\sigma(xy)+\sigma^2(y)\right) \\
&\quad +\beta\left(\mu(y)\sigma^2(x)+\mu(x)\sigma(xy)-\mu(x)\sigma^2(y)-\mu(y)\sigma(xy)\right)/N \\
\frac{\partial}{\partial t}\sigma^2(x) &= \beta\left(\mu(x)\mu(y)+\sigma(xy)\right)/N+\rho\left(N-\mu(x)-\mu(y)\right) \\
&\quad -2\beta\left(\mu(x)\sigma(xy)+\mu(y)\sigma^2(x)\right)/N-2\rho\left(\sigma(xy)+\sigma^2(x)\right) \\
\frac{\partial}{\partial t}\sigma^2(y) &= \beta\left(\mu(x)\mu(y)+\sigma(xy)\right)/N+\gamma\mu(y) \\
&\quad +2\beta\left(\mu(y)\sigma(xy)+\mu(x)\sigma^2(y)\right)/N-2\rho\sigma^2(y)
\end{aligned} \tag{3.19}$$

The full probability distribution for the ensemble of trajectories is assumed to begin at a single point, with zero variance in the distribution. $(\mu(x),\mu(y),\sigma(xy),\sigma^2(x),\sigma^2(y))=(N-n',n',0,0,0)$ (As time progresses, the trajectories stochastically diverge from one another and the probability distribution widens such that the variances become nonzero.)

Using Gaussian moment closure to simplify the master equation in this context has some clear benefits. The approximation has dramatically reduced the difficulty of the problem from a PDE with incompletely-defined boundary conditions to a set of ODE's with well-defined initial conditions. The trade-off is relying on the assumptions that the rate at which trajectories leave the QSD and die out is small, and that the QSD is approximately distributed according to a Gaussian distribution. As shown below, these latter assumptions fail in some important cases, such as when the QSD becomes close to zero and the decay rate is high.

There is alternative way to derive Eq. 3.19 using the diffusion approximation

[59, 60, 73]. A summary of this can be found in Appendix C.

3.3.4 Endemic State Analysis

Similar to the solution to the SIRS endemic state for the the ODEs in the deterministic model (Eq. 3.10), the long-term steady state behavior of Eq. 3.19 can be approximately solved for by setting the left hand side to 0. Adopting the notation $R_0 \equiv \beta/\gamma$ and let $\alpha \equiv \rho/\gamma$, expanding the solution in powers N yields the following expressions for the stochastic SIRS endemic state [33] :

$$\begin{aligned}
\mu(x)^* &= N\frac{1}{R_0} + \frac{1+\alpha}{\alpha}\frac{1}{R_0-1} + O\left(N^{-1}\right) \\
\mu(y)^* &= N\frac{\alpha}{1+\alpha}\left(1-\frac{1}{R_0}\right) - \frac{1}{R_0-1} + O\left(N^{-1}\right) \\
\sigma(xy)^* &= -N\frac{1}{R_0} + O\left(1\right) \\
\sigma^2(x)^* &= N\frac{\alpha\left(R_0-1\right)+\left(1+\alpha\right)^2}{\alpha\left(\alpha+R_0\right)R_0} + O\left(1\right) \\
\sigma^2(y)^* &= N\frac{\alpha\left(\alpha+R_0\right)^2+\left(1+\alpha\right)\left(R_0-1\right)}{\left(1+\alpha\right)^2\left(\alpha+R_0\right)R_0} + O\left(1\right)
\end{aligned} \tag{3.20}$$

There are some important similarities between the long-term endemic state described by Eq. 3.10 and the endemic state described by Eq. 3.20. Dividing both sides by N and taking the limit in which $N \to \infty$, Eq. 3.20 reduces to the deterministic model's solution, where the mean fractions of susceptible and infected $(\mu(x)/N, \mu(y)/N)$ become (S, I), and the variances about the mean go to zero. In other words, the deterministic model is the same as the stochastic model in the infinite population limit.

One of the additional advantages that analyzing the full stochastic model has over analyzing the deterministic model is that it accounts for the finite size of the popu-

lation. For example, in the solution for $\mu(y)$ the term $\frac{-1}{R_0-1}$ becomes large when R_0 is very close to the endemic threshold $R_0 - 1 \ll 1$. This additional term represses the mean number infected infected in the endemic state, particularly for small populations, and cannot be accounted for in the deterministic (infinite population) limit.

Fig. 3.7 shows comparisons between the cumulant equations' predictions and the simulations for four different values of ρ and four different values of R_0. The population size is $N = 500$. The simulations were measured over 10^4 trajectories and prepared such that the simulations began in the QSD (predicted by using the cumulant equations). QSD measurements are made during its steady state, after transients have ended but before the trajectories have died out. In the simulation, the mean endemic level increases with R_0, and also increases with ρ. Qualitatively, the cumulant equations agree with how the mean endemic level ($\mu(y)$, top row) and the QSD standard deviation ($\sigma(y)$, bottom row). Quantitatively, however the moment closure approximations are only accurate in regimes where the mean endemic level is high, meaning $R_0 > 1$ and $\rho/\gamma > 1$.

For the left hand side of Fig. 3.8 (A.), there is good quantitative and qualitative agreement between the simulation data and the cumulant equations' endemic steady state. The cumulant equations predicts $(\mu(x), \mu(y)) = (424.9, 57.0)$, while the means estimated from the simulation are $(424.8 \pm .3, 57.1 \pm .3)$. The cumulant equations also predicts the standard deviations $(\sigma(x), \sigma(y)) = (24.8, 19.4)$, while in the simulation they are $(24.9 \pm .3, 19.5 \pm 2)$. For the right hand side of Fig. 3.8 (B.), the cumulant equations predict $(\mu(x), \mu(y)) = (417.2, 7.52)$, while the means estimated from the simulation are $(385.6 \pm 1.8, 9.7 \pm 1.3)$. In this regime, with $\mu(y)$ close to zero and $\sigma(y)$ large enough such that the distribution of trajectories overlaps with the absorbing state, there is no longer any agreement between the simulations and the cumulant

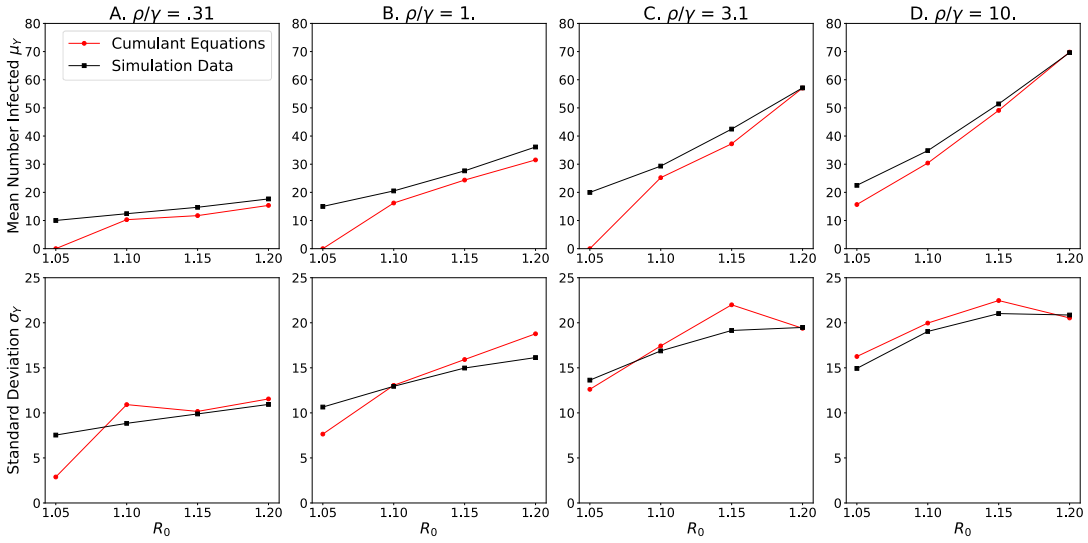

Figure 3.7: **Accuracy of Cumulant Equations**: Comparing the cumulant equations with stochastic simulation results in a population with $N = 500$. The simulations were measured over 10^4 trajectories. Each column represents a different value of ρ/γ, where for $\rho/\gamma < 1$ the mean number infected is suppressed. The top row shows comparisons of the mean number infected $\mu(y)$, plotted vs. increasing values of R_0. The bottom row shows comparisons of the standard deviation in the number infected $\sigma(y)$.

equations.

Fig. 3.8 gives a phenomenological understanding of why the Gaussian approximation fails when $\mu(y)$ close to zero. Referring back to Eq. 3.15, it was assumed that the term $\gamma q_{\cdot,1} \ll 1$. Clearly, from Fig. 3.8 B. $q_{\cdot,1}$ is no longer small enough to justify this assumption. $\gamma q_{\cdot,1}$ is the rate at which trajectories leave the QSD and enter the absorbing state. And so, when spontaneous extinctions occur at a high rate the above analysis of the QSD's master equation is no longer expected to be accurate. Additionally, the simulation QSD is clearly non-symmetric and non-Gaussian, as the distribution is cut off near $Y = 1$. It makes sense, then, that the assumption of a Gaussian-distributed QSD fails to quantitatively account for the behavior in this regime.

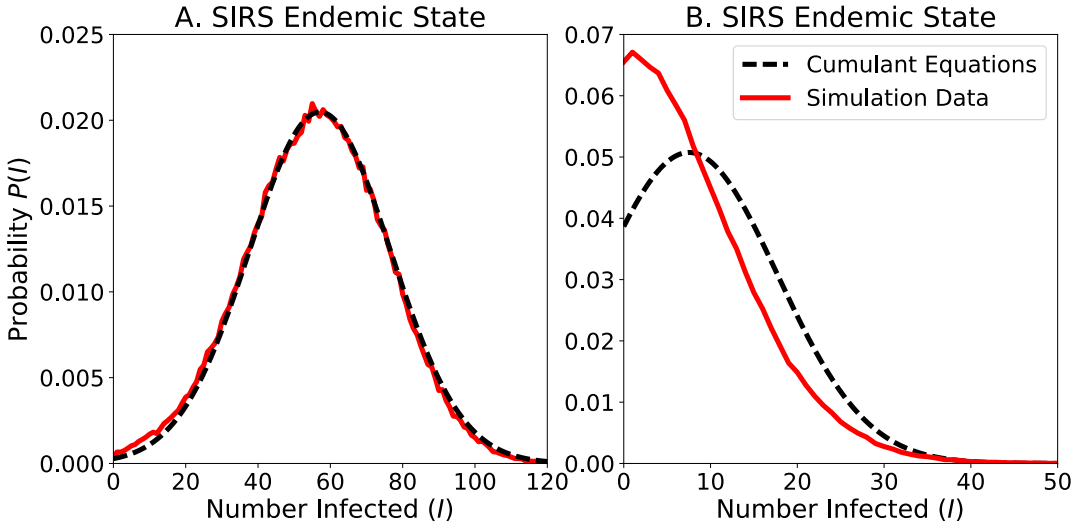

Figure 3.8: **Cumulant Equations' Approximation to Quasi-static Distribution**: Comparisons between simulations of the stochastic SIRS model quasi-static distribution and the cumulant equations (Eq. 3.20) A. QSD for an ensemble of 10^4 simulated trajectories in a population of $N = 500$ with $R_0 = 1.2$, $\gamma = 1.0$, $\rho = 3.2$. There is good quantitative agreement between the cumulant equations' approximation and the QSD measured using the simulations, particularly near the peak of the distribution. B. QSD for an ensemble of 10^4 simulated trajectories in a population of $N = 500$ with $R_0 = 1.2$, $\gamma = 1.0$, $\rho = 0.01$. In this regime, where the mean of the QSD is much closer to the absorbing state such that the QSD overlaps with 0, it is no longer accurate to approximate the QSD using a Gaussian distribution.

3.3.5 Mean Time to Extinction

To further explore the properties of the endemic disease state, we now turn to the question of how long the endemic disease state is expected to persist. The spontaneous extinction event illustrated by Fig. 3.6 suggests that understanding the rate of spontaneous extinctions requires knowing both the mean endemic level as well as the distribution of fluctuation sizes. The cumulant equations provide estimates of both of these quantities.

Combining Eq. 3.12 and Eq. 3.13 and assuming that the QSD is constant makes it possible to integrate Eq. 3.13 with respect to time to obtain an expression showing how in the number of trajectories in the QSD exponentially decays as trajectories

reach the absorbing state [9, 73, 74]:

$$p_{\cdot,0}(t) \approx 1 - e^{-q_{\cdot,1}t} \tag{3.21}$$

$\gamma q_{\cdot,1}$, the QSD evaluated at $Y = 1$, gives the exponential decay rate. The quantity $q_{\cdot,1}$ may be estimated using the cumulant equations' predictions for the mean and standard deviation of the QSD:

$$\begin{aligned} q_{\cdot,1} &\approx e^{-(1-\mu(y))^2/(2\sigma^2(y))}/\sqrt{2\pi\sigma^2(y)}/A \\ \text{where} \quad A &= \sum_{i=1}^{N} e^{-(i-\mu(y))^2/(2\sigma^2(y))}/\sqrt{2\pi\sigma^2(y)} \end{aligned} \tag{3.22}$$

This suggests that there is a phenomenological relationship between the properties of the QSD ($\mu(y)$ and $\sigma^2(y)$) and the exponential rate $q_{\cdot,1}$ at which trajectories in the QSD go extinct.

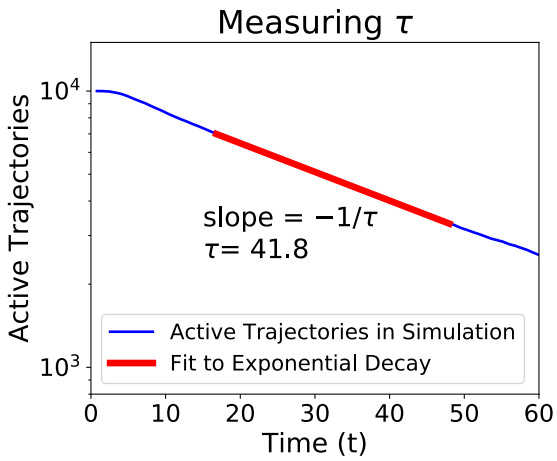

Figure 3.9: **Decay Rate Measurement**: Measuring the rate at which at which trajectories go extinct. The plot shows the number of active trajectories plotted vs. time, for an ensemble of 10^4 simulations of a population with $N = 500$ and model parameters $\beta = 1.1$, $\rho = 1.0$, and $\gamma = 1.0$. The measured slope is $q_{\cdot,1} = -0.0239$, with correlation coefficient $r = -.99993$. In this case, the mean time to extinction $\tau = 41.8$.

For quantitative comparison, the rate $q_{\cdot,1}$ can also be measured from simulations by counting the rate at which trajectories go extinct. Figure 3.9 illustrates how this is done, by fitting to the exponential decay rate. We also introduce the notation

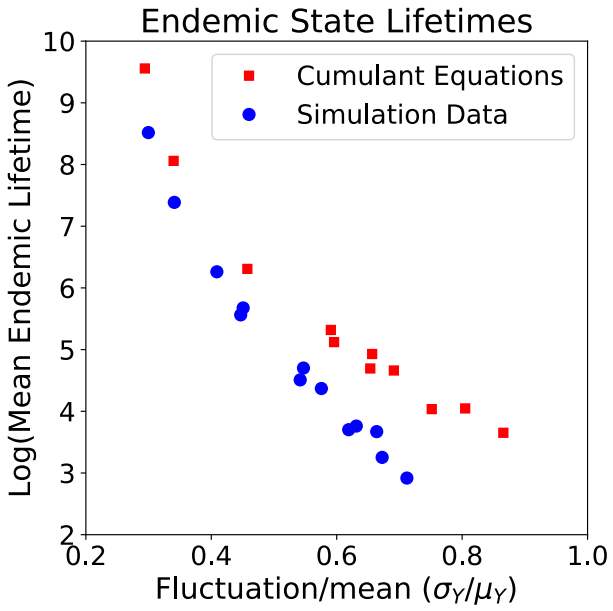

Figure 3.10: **Endemic State Lifetimes**: Comparison between the endemic state lifetimes measured in the simulations and the endemic state lifetimes predicted using the value of $\gamma q_{\cdot,1}$ from the cumulant equations (Eq. 3.22). The x-axis shows $\sigma(y)/\mu(y)$, a measure of how large the fluctuations are relative to the mean. The data points correspond to the simulations plotted in Fig. 3.7, ignoring all points where the cumulant equations predicts $\mu(y) = 0$. The cumulant equations' predictions (red squares) consistently over-estimate the endemic state lifetime for the simulation data (blue circles), although there is qualitative agreement in that the cumulant equations do predict longer lifetimes for more peaked distributions (small $\sigma(y)/\mu(y)$).

$\tau \equiv 1/q_{\cdot,1}$, the mean time to extinction, or characteristic lifetime of the endemic state.

Fig. 3.10 shows the relationship between the endemic state lifetimes and the relative size of fluctuations $\sigma(y)/\mu(y)$ as measured in the simulations (blue circles, same as the data plotted in Fig. 3.7, with $N = 500$, $R_0 = [1.05, 1.1, 1.15, 1.2]$, $\rho/\gamma = [0.31, 1., 3.1, 10.]$). Matching the intuition stated previously, small fluctuations correlate with longer endemic states and large fluctuations correlate with shorter endemic states. Fig. 3.10 also shows estimates of the mean time to extinction calculated using the cumulant equations together with Eq. 3.22. These estimates consistently over-estimate the endemic state lifetimes measured in the simulations, but neverthe-

less agree qualitatively with the relationship between $\sigma(y)/\mu(y)$ and τ.

What is remarkable here is how each set of data points appears to collapse onto a single curve. There are two mappings at work here. First, Eq. 3.20 describes the relationship between the SIRS model parameters (N, R_0, ρ/γ) and the statistical properties of the QSD ($\mu(y)$, $\sigma(y)$). Second, Eq. 3.22 predicts that there is a straightforward relationship between the statistical properties of the QSD and the mean time to extinction τ. All together, for both the simulation data and the cumulant equations, the model behavior throughout all of parameter space appears to be restricted to one low-dimensional curve. We interpret this as a phenomenological relationship between the relative size of fluctuations ($\sigma(y)/\mu(y)$) and the mean time to extinction τ.

3.4 Population Heterogeneity and Network Effects

The previous analysis of the SIRS model only considered the endemic state in a fully mixed, homogeneous population in which all individuals and contacts are identical. The simplifying assumption of a homogeneous population is mathematically convenient but not particularly realistic when it comes to representing how diseases affect real-world populations. There are many different ways in which heterogeneity enters into disease modeling. Populations may be subdivided into communities (metapopulations), in which individuals interact frequently with their neighbors within the community but interact infrequently with individuals who belong to other communities. For certain diseases transmission may depend strongly on the age of an individual, such as if young children have not yet been vaccinated and so are more at risk of becoming infected [60]. Some diseases may have high variability in transmission across

different individuals, meaning that most people do not have a high transmission rate but a small number of people have very high transmission rate [69]. Similarly, there can be variability in the number of contacts that different individuals have, such that some individuals have few opportunities to spread an infection to others while other individuals have a very large number of such opportunities [69, 84]. Other diseases feature the transfer of infection between different species, such that understanding zoonotic spillover requires consideration of the different modes of how an infection spreads between animals and humans [71].

Incorporating population heterogeneity into a model allows for a more detailed description of how different individuals interact with one another [11, 81]. It is also important to understand how adding new features to disease models affect the outcome of an outbreak of disease, and to know how heterogeneous populations are affected differently from homogeneous populations [34, 35]. The purpose of the following sections will be to extend the analysis of the stochastic SIRS model in homogeneous populations to include population heterogeneity, and then to explore how the heterogeneity affects the properties of the SIRS endemic disease state.

3.4.1 SIRS in Heterogeneous Populations

For a homogeneous population, with a single type of individual, the term describing transmission through contact in Eq. 3.8 is $\beta XY/N$. A heterogeneous population includes multiple classes of individuals, where members of each pair of classes may interact differently. To account for this, the single interaction parameter β is now replaced by a "who-is-infected-by-whom" matrix $\boldsymbol{B}$ [60]. If K is the number of classes, then $\boldsymbol{B}$ a $K \times K$ matrix where $B_{i,j}x_i y_j/N$ is the force of infection between infected members of class j and susceptible members of class i.

To give an example of how to use a matrix $\boldsymbol{B}$ to describe population heterogeneity, one might imagine a population divided into three separate communities, where members belonging to each community interact strongly with each other but weakly with members of the other communities. To model such a population, the who-is-infected-by-whom matrix becomes:

$$\boldsymbol{B} = \begin{bmatrix} \beta & x & x \\ x & \beta & x \\ x & x & \beta \end{bmatrix}$$

where $\beta > x$ in order to account for stronger force of infection within each community.

The analysis of the heterogeneous SIRS model proceeds in the same way as for the homogeneous model, starting with the deterministic version of the model. If the population is divided into K classes, each class makes up a number N_i of the total population (where $\sum_i^K N_i = N$) and contains different fractions of susceptible and infected individuals ($(S_i, I_i) = (X_i/N_i, Y_i/N_i)$):

$$\begin{aligned} \frac{dS_i}{dt} &= -\sum_{j=1}^{K} B_{i,j} S_i I_j + \rho\,(1 - S_i - I_i) \\ \frac{dI_i}{dt} &= \sum_{j=1}^{K} B_{i,j} S_i I_j - \gamma I_i \end{aligned} \tag{3.23}$$

For K classes, the QSD is approximated to be a $2K$-dimensional multivariate Gaussian distribution, with $2K$ means $(\mu(x_i), \mu(y_i))$ and $2K$ variances ($\sigma^2(x_i)$ and $\sigma^2(y_i)$). Covariances $(\sigma(x_i, x_j), \sigma(x_i, y_j), \sigma(y_i, y_j))$ form a $2K \times 2K$ matrix. Applying the same analysis used to account for the stochastic effects, it is possible to derive a new set of ODEs for the cumulants of the QSD for a heterogeneous population.

$$
\begin{aligned}
\frac{\partial}{\partial t}\mu(x_i) &= \rho\left(N_i - \mu(x_i) - \mu(y_i)\right) - \sum_{j=1}^{K} B_{i,j}\left[\mu(x_i)\mu(y_j) + \sigma(x_i, y_j)\right] \\
\frac{\partial}{\partial t}\mu(y_i) &= \sum_{j=1}^{K} B_{i,j}\left[\mu(x_i)\mu(y_j) + \sigma(x_i, y_j)\right] - \gamma\mu(y_i) \\
\frac{\partial}{\partial t}\sigma(x_i, y_j) &= -\rho\sigma^2(y_i) - (\gamma + \rho)\,\sigma(xy) \\
&\quad - \sum_{m=1}^{K} B_{i,m}\left[\mu(x_i)\sigma(y_j, y_m) + \mu(y_m)\sigma(x_i, y_j)\right] \\
&\quad + \sum_{m=1}^{K} B_{j,m}\left[\mu(x_j)\sigma(x_i, y_m) + \mu(y_m)\sigma(x_i, y_j)\right] \\
&\quad - \delta_{i,j}\sum_{m=1}^{K} B_{j,m}\left[\mu(x_i)\mu(y_m) + \sigma(x_i, y_m)\right] \\
\frac{\partial}{\partial t}\sigma(x_i, x_j) &= -2\rho - \rho\sigma(x_j, y_i) - \rho\sigma(x_i, y_j) \\
&\quad - \sum_{m=1}^{K} B_{i,m}\left[\mu(x_i)\sigma(x_j y_m) + \mu(y_m)\sigma(x_i, x_j)\right] \\
&\quad - \sum_{m=1}^{K} B_{j,m}\left[\mu(x_j)\sigma(x_i, y_m) + \mu(y_m)\sigma(x_i, x_j)\right] \\
&\quad + \delta_{i,j}\left(\rho\left(N_i - \mu(x_i) - \mu(y_i)\right) + \sum_{m=1}^{K} B_{i,m}\left[\mu(x_i)\mu(y_m) + \sigma(x_i, y_m)\right]\right) \\
\frac{\partial}{\partial t}\sigma(y_i, y_j) &= -2\gamma\sigma(y_i, y_j) \\
&\quad + \sum_{m=1}^{K} B_{i,m}\left[\mu(x_i)\sigma(y_j y_m) + \mu(y_j)\sigma(x_i, y_j)\right] \\
&\quad + \sum_{m=1}^{K} B_{j,m}\left[\mu(x_j)\sigma(y_i, y_m) + \mu(y_m)\sigma(x_j, y_i)\right] \\
&\quad + \delta_{i,j}\left(\gamma\mu(y_i) + \sum_{m=1}^{K} B_{i,m}\left[\mu(x_i)\mu(y_m) + \sigma(x_i, y_m)\right]\right)
\end{aligned}
\tag{3.24}
$$

Eq. 3.24 appears complicated, but nevertheless can be integrated numerically. Also, the steady-state behavior of Eq. 3.24 can be solved for directly by setting the left hand sides to zero. In practice, both numerical integration of Eq. 3.24 as

well as root-finding algorithms converge fastest for parts of parameter space where the mean numbers infected ($\mu(y_i)$) are larger than 0, which occasionally leads to numerical difficulties when close to the critical point. This is to be suspected, as the assumption that the QSD is Gaussian-distributed is no longer valid close to the endemic threshold.

3.4.2 Heterogeneous Mean Field for Annealed Networks

Networks are particularly useful for incorporating heterogeneity in contacts between individuals [11]. Not every person interacts with the same number of other people. For example, the sexual contacts traced in [84] reveal sparse networks with a great deal of degree heterogeneity - most individuals have only a few sexual partners, while a small number of individuals have very many sexual partners. These two types of individuals, those with few and those with many contacts, have different amounts of risk when it comes to contracting a sexually transmitted infection. For the purposes of creating a more realistic model, It can be crucial to incorporate variation in the amount of risk of exposure or transmission across different individuals. For a heterogeneously connected population, the initial conditions for the start of an epidemic can strongly depend on who is infected first, as person with more contacts is more likely to allow the disease to spread to others than a person with fewer contacts [69].

One way to model a network's heterogeneity is to use the uncorrelated annealed network approximation [11]. This approximation assumes that a node's degree is its most important property, such that all nodes with the same degree are identical and can be grouped together into degree classes. This is convenient for the purposes of analysis, since it is no longer necessary to track the state of each node independently. Instead, the nodes belonging to each degree class form a single compartment in the

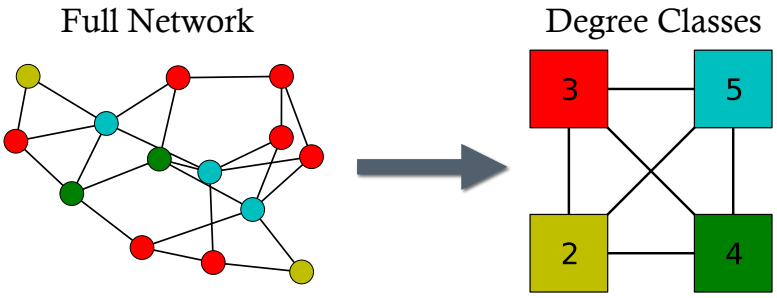

Figure 3.11: **Heterogeneous Mean Field Schematic**: Each node in the network has a particular degree. Rather than treat each node separately, each node is categorized according to its degree, such that each group of nodes constitutes a degree class. Each degree class interacts differently with each of the other degree classes. In the context of compartmental disease modeling the who-is-infected-by-whom matrix $\boldsymbol{B}$ defines how strongly the different degree classes interact with one another.

model, and each degree class interacts differently with each other degree class. In the context of compartmental disease modeling the who-is-infected-by-whom matrix $\boldsymbol{B}$ defines how strongly the different degree classes interact with one another [60].

The who-is-infected-by-whom matrix for an uncorrelated annealed network is derived by determining the force of infection for the nodes belonging to each degree class. Suppose each of the nodes in degree class i have degree k_i. Each of the k_i edges has a probability $k_j\mathbb{P}(k_j)/\langle k\rangle$ of connecting to another node with degree k_j, where the probability of connecting to a node with degree k_j [79]. The probability of connecting to a node with degree k_j is proportional to k_j and the normalizing factor $\langle k\rangle$ is the mean degree. Thus, the probability of connecting a node with degree k_i to an infected node with degree k_j is:

$$\frac{k_i k_j \mathbb{P}(k_j)}{\langle k\rangle}\frac{Y_j}{N_j} = \frac{k_i k_j}{\langle k\rangle}\mathbb{P}(k_j)I_j$$

The total probability of connecting to any infected node, therefore, requires a sum over all degree classes, and the transmission term in Eq. 3.23 becomes

$$\beta\sum_{j=1}^{K}\frac{k_i k_j}{\langle k\rangle}\mathbb{P}(k_j)S_i I_j$$

meaning that the who-is-infected-by-whom matrix is

$$B_{i,j} = \beta \frac{k_i k_j}{\langle k \rangle} \mathbb{P}(k_j) \tag{3.25}$$

Incorporating Eq. 3.25 into Eq. 3.23, the deterministic SIRS model for annealed networks becomes

$$\begin{aligned} \frac{dS_i}{dt} &= \rho\left(1 - S_i - I_i\right) - \beta k_i S_i \sum_j^K \frac{k_j}{\langle k \rangle} \mathbb{P}(k_j) I_j \\ &= \rho\left(1 - S_i - I_i\right) - \beta k_i S_i \Theta \\ \frac{dI_i}{dt} &= -\gamma I_i + \beta k_i S_i \sum_j^K \frac{k_j}{\langle k \rangle} \mathbb{P}(k_j) I_j \\ &= -\gamma I_i + \beta k_i S_i \Theta \end{aligned} \tag{3.26}$$

where $\Theta \equiv \sum_j^K \frac{k_j}{\langle k \rangle} \mathbb{P}(k_j) I_j$ [11, 82, 81]. Θ is an effective mean field, calculated by taking a weighted average over the fraction of infected nodes in each of the network's degree classes. The strength of each node's (or rather, each degree class's) interaction with Θ depends on the degree ($\sim k_i \Theta$).

The next step is to solve for the steady-state behavior of Eq. 3.26. Setting the left hand side of Eq. 3.26 to 0, the endemic level becomes

$$\begin{aligned} S_i^* &= \frac{1}{1 + k_i \beta / \gamma \left(1 + \gamma / \rho\right) \Theta} \\ I_i^* &= \frac{k_i \beta / \gamma \Theta}{1 + k_i \beta / \gamma \left(1 + \gamma / \rho\right) \Theta} \end{aligned} \tag{3.27}$$

Plugging the expression for I_i^* from Eq. 3.27 into the definition of Θ yields a self-consistency equation for Θ:

$$\Theta = \frac{k_j}{\langle k \rangle} \mathbb{P}(k_j) \frac{k \tilde{\beta} / \gamma \Theta}{1 + k \tilde{\beta} / \gamma \left(1 + \gamma / \rho\right) \Theta} \tag{3.28}$$

The trivial solution is $\Theta = 0$, which corresponds to a disease-free state with $I_i^* = 0$. Dividing both sides of Eq. 3.28 yields condition on the parameters above which $\Theta > 0$

and $I_i^* > 0$:

$$\frac{\beta}{\gamma} \geq \frac{\langle k \rangle}{\langle k^2 \rangle} \tag{3.29}$$

where $\langle k^2 \rangle \equiv \sum_j^K \mathbb{P}(k_j) k_j^2$ [11, 81, 82]. This, for the deterministic heterogeneous SIRS model, is the endemic threshold above which there is a sustained endemic state. In contrast to the condition derived for the homogeneous deterministic SIRS model previously, Eq. 3.29 now depends explicitly on the contact heterogeneity of the population. The expression $\langle k^2 \rangle$ is a property of the degree distribution.

For more heterogeneous networks, with widely varying degree distributions, $\langle k^2 \rangle$ can become very large (or even diverge to infinity in the limit of infinitely large networks with heavy-tailed degree distributions) [11, 79, 81, 82]. In this way, this analysis of the deterministic model has shown how heterogeneity in the distribution of contacts in a population can affect the outcome of an epidemic, where highly heterogeneous populations may have vary low endemic thresholds compared to more homogeneous populations.

3.5 Stochastic SIRS on Annealed Networks

The next step of this discussion will focus on analyzing the endemic state for the stochastic version of the SIRS model in a heterogeneous population.The notion of critical community size, introduced earlier, relates to the relationship between population size and the lifetime of the endemic disease state: in stochastic models of endemic disease, the average lifetime of an endemic disease state is longer in larger populations than in smaller populations.

Population size is only one parameter of models of endemic disease, and it remains an open question how population heterogeneity contributes to the lifetime of the

endemic disease state. To explore this question, we consider a set of four annealed networks with varying levels of heterogeneity. Each network contains 500 nodes that have been partitioned into two degree classes - this way, the model separately tracks nodes with high degree and nodes with low degree. Each network has the same mean degree $\langle k \rangle = 10$, and the same low degree $k_{\text{low}} = 5$. The degree of high degree nodes k_{high} as well as the proportion of high degree nodes $\mathbb{P}(k_{\text{high}}) = 1 - \mathbb{P}(k_{\text{low}})$ is allowed to vary such that the breadth of the degree distribution $\langle k^2 \rangle$ also varies between networks. For each of the four networks, the ratio of the first two moments of the degree distribution $\langle k^2 \rangle / \langle k \rangle$ is different, and serves as a measure of each network's heterogeneity. (This quantity is also important from a modeling perspective, as it defines the location of the endemic threshold for each network.)

Label	k_{low}	$\mathbb{P}(k_{\text{low}})$	k_{high}	$\mathbb{P}(k_{\text{high}})$	$\langle k \rangle$	$\langle k^2 \rangle$	$\sigma_k^2 = \langle k^2 \rangle - \langle k \rangle$
A	5	0.50	15	0.50	10	125	25
B	5	0.80	30	0.20	10	200	100
C	5	0.941	90	0.059	10	500	200
D	5	0.985	330	0.015	10	1700	400

Table 3.2: **Network Statistics**: Basic properties of the four heterogeneous networks analyzed. Each network contains $N = 500$ nodes. The mean degree is held constant ($\langle k \rangle = 10$) across all four networks, but the second moment in the degree distribution ($\langle k^2 \rangle$), a measure of degree heterogeneity, increases from A to D. The fraction of low degree nodes increases and the fraction of high degree nodes decreases from A to D.

In principle, it is possible to analyze networks with any degree distribution using Eq. 3.24 or other tools that compartmentalize the network into degree classes. For the purposes of the present study, however, it is far more straightforward to focus on networks with binary degree distributions. These networks is more tractable to analyze but still have controllable heterogeneity.

3.5.1 Endemic State Phase Diagrams

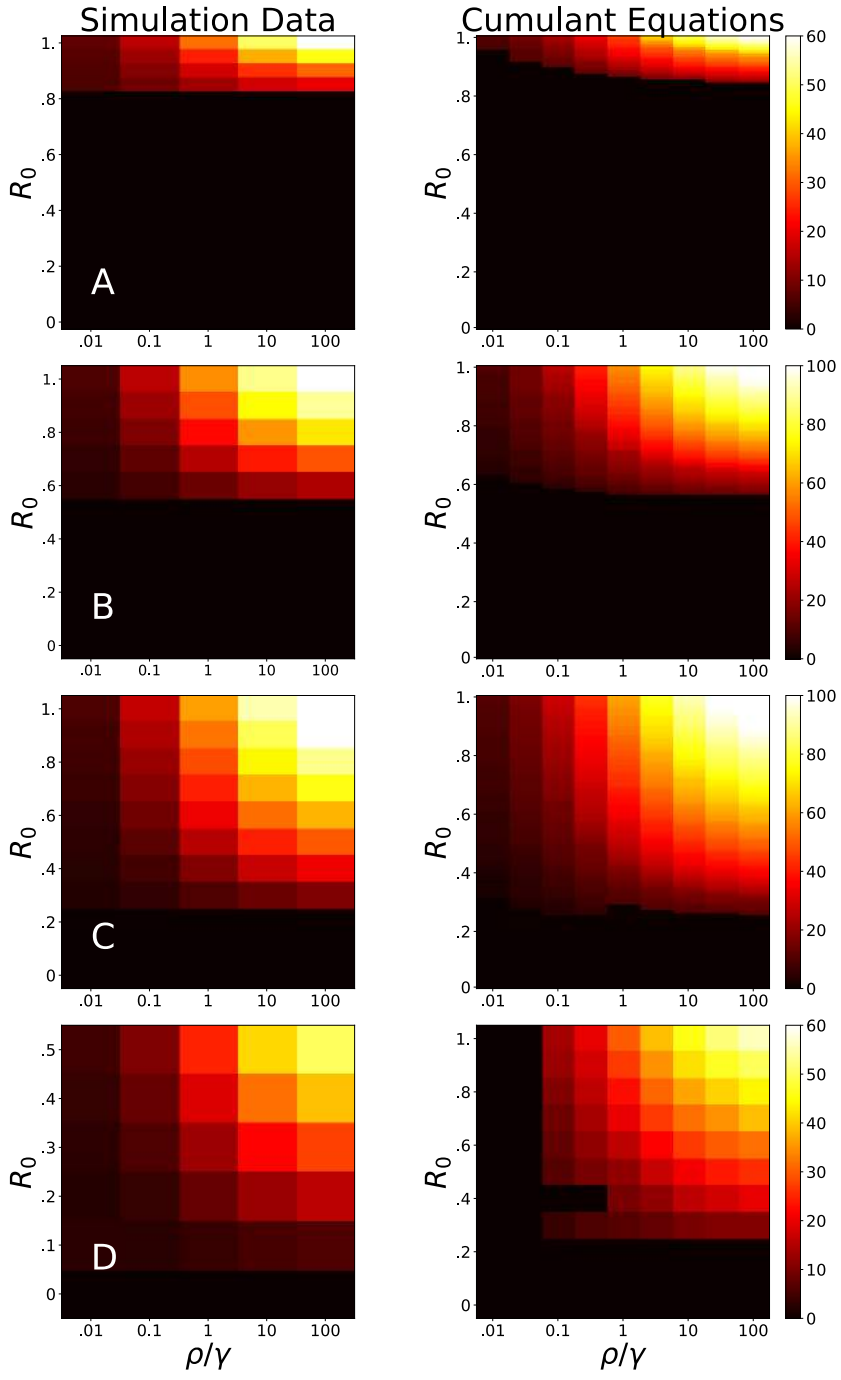

Figure 3.12: **Mean Endemic Infection Level**: The left hand column shows the results of simulations of the SIRS model on four networks with differing heterogeneity, plotting the total mean infection level $\mu(y_{total})$. The right hand column shows the same quantity predicted by the cumulant equations (Eq. 3.24).

Fig. 3.12 and Fig. 3.13 show quantitative comparisons between properties of the QSD measured using stochastic simulations of the SIRS model and predicted using the endemic steady-state behavior of the cumulant equations (Eq. 3.24). For each network (A, B, C, D), 10^4 trajectories of the SIRS model were simulated in order to measure the properties of the SIRS dynamics. For each network, the SIRS model parameters ρ/γ, $R_0 = \beta/\gamma$ (with $\gamma = 1.0$) were varied in order to survey a range of the model's behavior.

Fig. 3.12 plots the total mean infection level $\mu(y_{total}) = \mu(y_{\text{low}}) + \mu(y_{\text{high}})$ for networks A, B, C, and D. The left hand column shows the results of the simulations, and the right hand column shows the numerical predictions of the cumulant equations. For each network, the dependence of $\mu(y_{total})$ on parameters $(R_0, \rho/\gamma)$ is qualitatively similar to that of the deterministic SIR model (Fig. 3.5 D.), with $\mu(y_{total})$ increasing with R_0 above the endemic threshold, as well as increasing with ρ/γ. One key quantitative difference between the different networks, however, is how the endemic threshold level changes depending on the heterogeneity of each network according to Eq. 3.29. The quantitative agreement between the simulation results and the cumulant equations varies across the different points in parameter space. Similar to the pattern seen in Fig. 3.7, the cumulant equations are most accurate for parameter values where the mean infection level is high.

Being able to predict the size of fluctuations in the model is an important feature of the cumulant equations. Fig. 3.13 plots the standard deviation about the mean infection level $\sigma(y_{\text{total}})$, where $\sigma^2(y_{\text{low}} + y_{\text{high}}) = \sigma^2(y_{\text{low}}) + \sigma^2(y_{\text{high}}) + 2\sigma(y_{\text{low}}, y_{\text{high}})$. This quantifies the characteristic fluctuations about the mean seen in stochastic simulations. Again, the cumulant equations are most accurate for parameter values where the mean infection level is high.

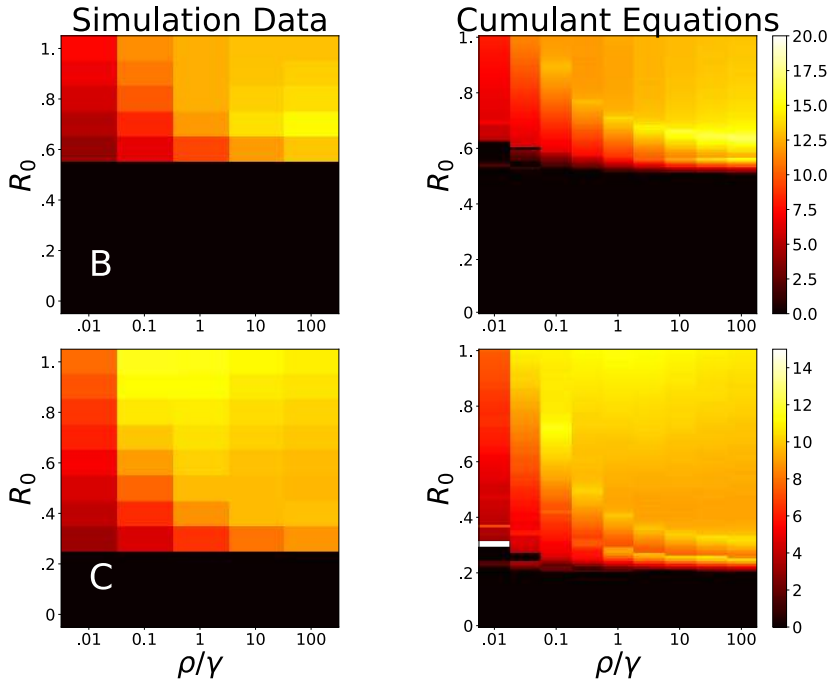

Figure 3.13: **Fluctuation Size of Endemic Infection Level**: The left hand column shows the results of simulations of the SIRS model on four networks with differing heterogeneity, plotting the total mean infection level $\sigma(y_{\text{total}})$. The right hand column shows the same quantity predicted using the long-term behavior of the cumulant equations (Eq. 3.24).

In contrast to the behavior of the mean endemic level's dependence on model parameters, the fluctuations remain mostly constant above the endemic threshold. Comparing Fig 3.13 to Fig. 3.12, there is a noticeable lack of variation in the size of the fluctuations across the upper region of parameter space. Only in the region very close to the endemic threshold does there appear to be a rapid change in the fluctuation size. For example, in the $\rho/\gamma = 10.$ column of the plot showing the mean endemic level in network B, the mean endemic level increases steadily from $\mu(y_{\text{total}}) = 26$ at $R_0 = 0.6$ to 105 at $R_0 = 1.0$, almost a factor of 4. The fluctuation sizes about the mean endemic level change by less than 2% across the same range of parameters.

Fig. 3.14 shows the relative size of the fluctuations $\sigma(y_{\text{tot}})/\mu(y_{\text{tot}})$, plotted as a

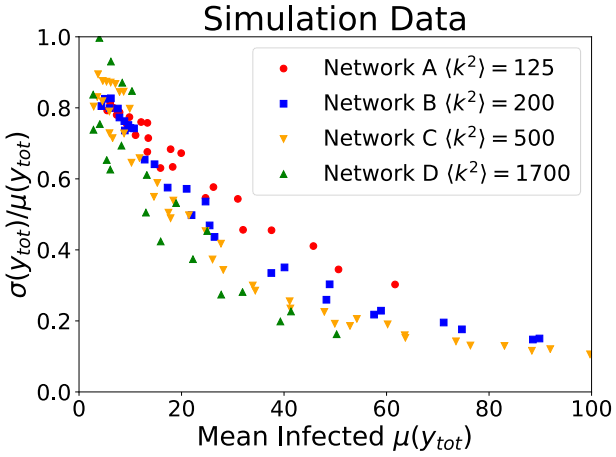

Figure 3.14: **How Variance Depends on Graph Heterogeneity**: Simulation results for networks with varying heterogeneity, plotting the relative size of the variance $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ vs. $\mu(y_{\text{total}})$ for four different graphs. There appears to be a monotonic relationship such that $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ decreases as $\mu(y_{\text{total}})$ increases. For fixed $\mu(y_{\text{total}})$, the fluctuations tend to decrease gradually as the network heterogeneity increases.

function of the $\mu(y_{\text{tot}})$. There appears to be a roughly monotonic relationship between the relative size of fluctuations and the mean endemic level, such that fluctuations decrease. Note that for a fixed value of the mean endemic level, relative fluctuation size decreases as the network heterogeneity increases.

3.5.2 Mean Times to Extinction

Figure 3.15 shows plots of the logarithm of the lifetime of the endemic state $\log \tau$ for both the simulations and the cumulant equations. Eq. 3.22, with $\mu(y_{\text{total}})$ and $\sigma^2(y_{\text{total}})$ from Eq. 3.24, was used to estimate τ at each point in parameter space. Comparing Fig. 3.12 to Fig. 3.15, it appears that the mean lifetime is high in the parts of parameter space where $\mu(y_{\text{total}}) > 0$. This is intuitive because when the mean endemic level is high, it is less likely for a fluctuation to spontaneously cause the extinction of the endemic state. In contrast to the behavior of the mean endemic

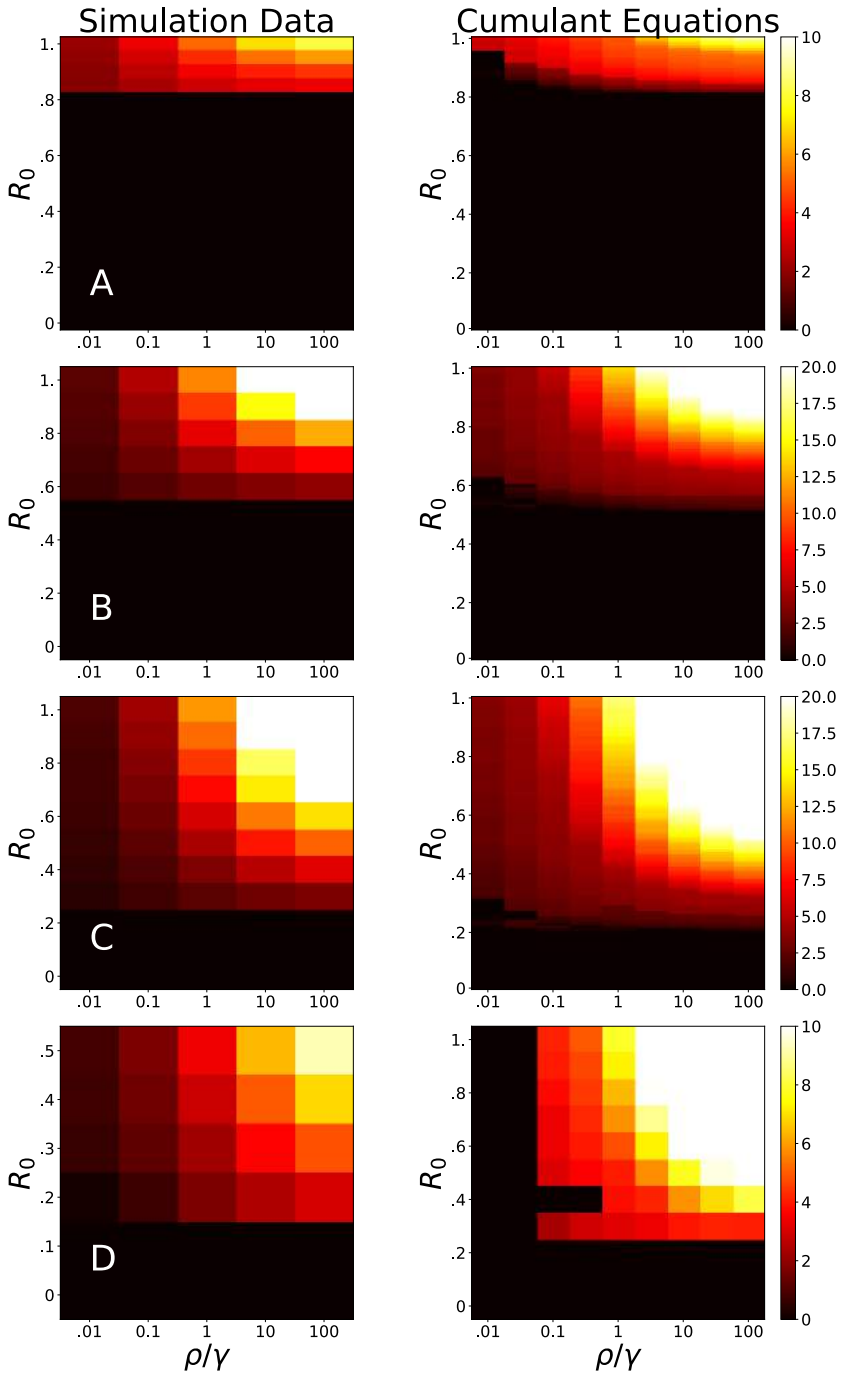

Figure 3.15: **Endemic State Lifetime**: The left hand column shows the results of simulations of the SIRS model on four networks with differing heterogeneity, plotting the endemic state lifetime throughout parameter space. The right hand column shows the mean infection level predicted using the long-term behavior of the cumulant equations (Eq. 3.22, using the results from Eq. 3.24). Note the similarity in the active regions between these plots and the plots of the mean endemic level - τ is high when endemic level is high. Note also the nonlinear behavior of the lifetime, as it appears to diverge for large values of ρ/γ and R_0.

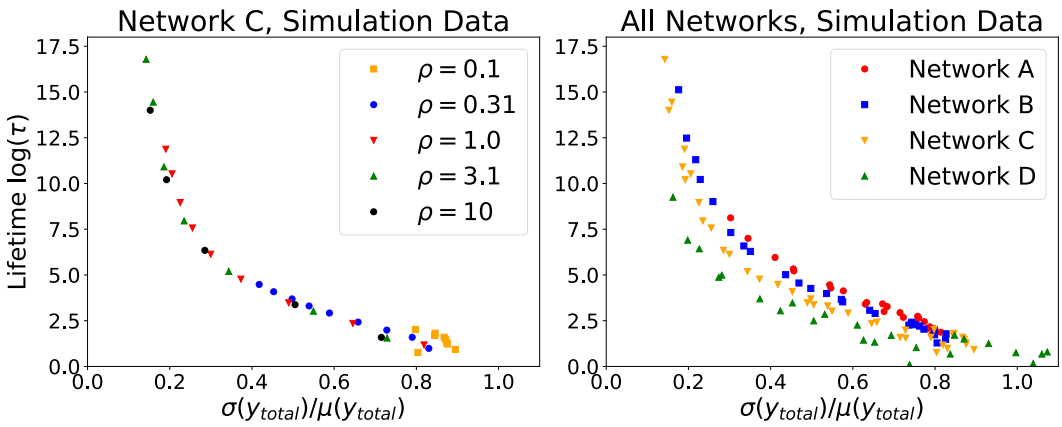

Figure 3.16: **Lifetimes vs. Relative Fluctuation Sizes**: Left side shows the relationship between τ and $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ for Network C only. The data have been partitioned according to the value of ρ for the purposes of illustrating how data from different regions of parameter space all collapse together onto the same curves. Right side plots the relationship between τ and $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ for SIRS simulations for four different graphs. For constant values of $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$, the lifetime decreases as the heterogeneity decreases, suggesting that focusing on the statistics of y_{total} misses some details about what y_{low} and y_{high} might be doing separately.

level, however, the endemic state lifetime shown in Fig. 3.12 varies nonlinearly with the model parameters, and in fact appears to begin to diverge in the upper right hand corners of the plots.

Recalling the relationship between the mean times to extinction and the relative size of fluctuations suggested in Fig. 3.10, the plots in Fig. 3.16 plots τ as a function of $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ for different networks. The left hand plot shows the simulation data for network C only, where the data have been partitioned according to the value of ρ. This illustrates how data from different regions of parameter space collapse together onto the same curves. The right hand plot juxtaposes the same plot for networks A, B, C, and D. For all four networks, the curve of data shows that the endemic state lifetime τ increases and begins to diverge as $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ decreases. (In Appendix D, the data points of τ vs. $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ are plotted separately for each network, juxtaposed with the corresponding cumulant equation predictions.)

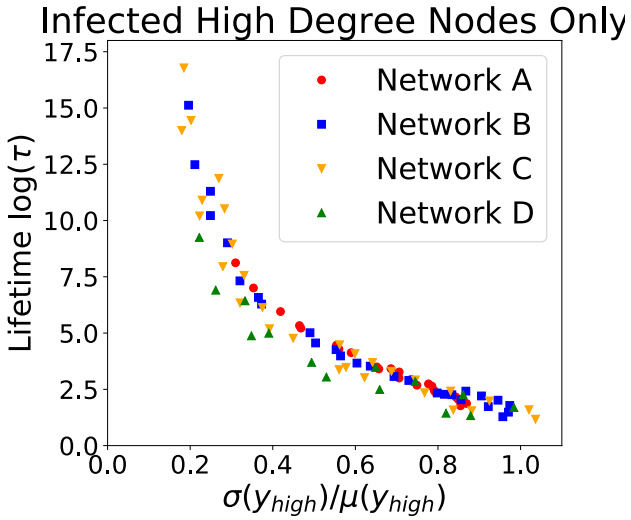

Figure 3.17: **Lifetimes vs. Relative Fluctuation Sizes**: The relationship between τ and the statistics for the high degree nodes only $\sigma(y_{\text{high}})/\mu(y_{\text{high}})$, plotted for four different networks. The data from each all four networks appear to collapse together, even more closely than the curves shown in Fig. 3.16.

The data points corresponding to each network fall onto distinct curves, with the most heterogeneous network (D) having the lowest τ and the least heterogeneous network (A) having the highest τ for a fixed value of $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$. The variation in how τ relates to $\sigma(y_{\text{total}})/\mu(y_{\text{total}})$ across networks suggests that the pathway to extinction, as a trajectory travels from the vicinity of the mean endemic level all the way to the absorbing state, changes depending on the network's heterogeneity. After all, for networks with two degree classes the QSD is a multivariate Gaussian distribution, where both y_{low} and y_{high} need to reach 0 separately in order for the disease to die out completely.

Fig. 3.17 plots τ against the ratio of moments of the high node degree distribution $\sigma(y_{\text{high}})/\mu(y_{\text{high}})$. Compared to the scatter plot in Fig. 3.16, the data points measured for all networks come much closer to collapsing onto a single curve. For the purposes of predicting the endemic state lifetime, focusing only on the behavior of the high degree nodes gives a better understanding of how the statistical properties of the

endemic state relate to the endemic state lifetime. (Not shown is τ plotted against $\sigma(y_{\text{low}})/\mu(y_{\text{low}})$, which instead causes the different scatter plots to spread apart from one another.) All of this suggests that the high degree nodes primarily drive the process through which the endemic state goes extinct.

Once again, it is worth noting how the four curves of data in the right hand side Fig. 3.16 collapse close to one another. The phenomenological relationship between the relative size of fluctuations in the quasi-static distribution and the mean time to extinction appears to be robust not only across different regions of parameter space, but also is robust to changes to the network's heterogeneity.

3.5.3 Paths to Extinction

To further investigate the role that the high degree nodes play in the extinction process, we plot the paths through configuration space $(y_{\text{low}}(t), y_{\text{high}}(t))$ that the ensemble of trajectories takes as it moves from the region near the mean endemic level to the absorbing state. For each network, a set of 10^4 trajectories simulated in order to carefully measure the exact path taken by each trajectory before going extinct. The data presented in Fig. 3.18 were generated using the following simulation parameters. For network A, $(\rho/\gamma = 1., \beta/\gamma = R_0 = 0.9)$; for network B, $(\rho/\gamma = 1., \beta/\gamma = R_0 = 0.7)$; for network C, $(\rho/\gamma = 1., \beta/\gamma = R_0 = 0.5)$; for network D, $(\rho/\gamma = 1., \beta/\gamma = R_0 = 0.5)$.

Figure 3.18 shows how the ensemble of trajectories proceeds from the mean endemic level down to the absorbing state. Each trajectory has a time when it goes extinct t_{ext}. The leftmost panel of each row of Fig. 3.18 shows the ensemble of trajectories during at a fixed time prior to t_{ext} for each trajectory. From left to right,

each panel shows the ensemble of trajectories at different times prior to t_{ext} as they proceed towards extinction. Each heat map represents a superposition of trajectories that outlines the characteristic path to extinction. The bright cross-shaped regions appearing in the leftmost panels of each row correspond to the mean endemic level, the starting points for each of the plotted trajectories.

Examining the characteristic paths to extinction for networks C and D highlights the important role that the high-degree nodes play in driving the extinction of the endemic state. For more heterogeneous networks, it appears that the path to extinction approaches a two-step process. Each row of Fig. 3.18 corresponds to a different network, and it is clear that the shape of the characteristic path to extinction depends on the network heterogeneity. For networks A and B, the paths to extinction appear to be symmetric, with both high-degree and low-degree nodes going to zero at the same rate. For the more heterogeneous networks C and D, however, appear to take an asymmetric extinction path. The paths to extinction in networks C and D begin with a rapid initial decrease in the number of infected high-degree nodes (y_{high}, along the y-axis). Only after y_{high} reaches 0 does the infection also die out in the low-degree nodes.

The effect that network heterogeneity has on the shape of the characteristic path to extinction shown here is consistent with previously reported results for the SIS model [56]. The SIS model represents one limit of the SIRS model, in which $\rho \to \infty$. Choosing a finite value of ρ for the general SIRS model does not appear to affect the characteristic path to extinction.

3.6 Discussion

We have explored and characterized the persistence behavior of the stochastic SIRS model on networks with varying topology. Analysis of the cumulant equations of the SIRS master equation yielded predictions for both the mean endemic level and the characteristic size of fluctuations in the quasi-static endemic state. These results were consistent with computer simulations of the same model for a set of annealed networks with varying amounts of heterogeneity.

We argued that the mean time to extinction τ was governed by the properties of the quasi-static distribution, μ and σ. For both the simulations as well as the numerical estimates produced using the cumulant equations, there appeared to be a simple, low-dimensional mapping from the properties of the QSD and the mean time to extinction. This led to a straightforward phenomenological relationship between QSD and τ that remains robust across wide variations in the different parameter inputs.

When introducing network heterogeneity, we found that for each different network there was the same straightforward relationship between τ and σ/μ, with only a slight variation across networks with different topologies. By focusing on the behavior of the high degree nodes, we were able to collapse the data together, effectively integrating out the variation due to network topology. This suggested that the fluctuations in the endemic infection level for high degree nodes were most important for driving the extinction process. This intuition was verified by plotting the characteristic paths to extinction and seeing that for heterogeneous graphs those paths are most strongly defined by the rapid extinction of the high degree nodes. Overall, the phenomenological relationship between the measurable properties of the QSD and the mean time to extinction does not change depending on network topology.

Our analysis was conducted using annealed networks, which were mathematically convenient for direct comparison with the computer simulations. It remains to be seen whether there is a similar straightforward relationship between τ and the properties of the QSD for networks with quenched disorder, where the edges are frozen and each node interacts only with its neighbors rather than interacting with a heterogeneous mean field. It has been shown that in quenched networks high-degree nodes drive the long-term dynamics and extinction properties of the endemic state [30, 70], but it is not yet known whether there is a similar relationship between the endemic state and the mean time to extinction as we have shown for annealed networks.

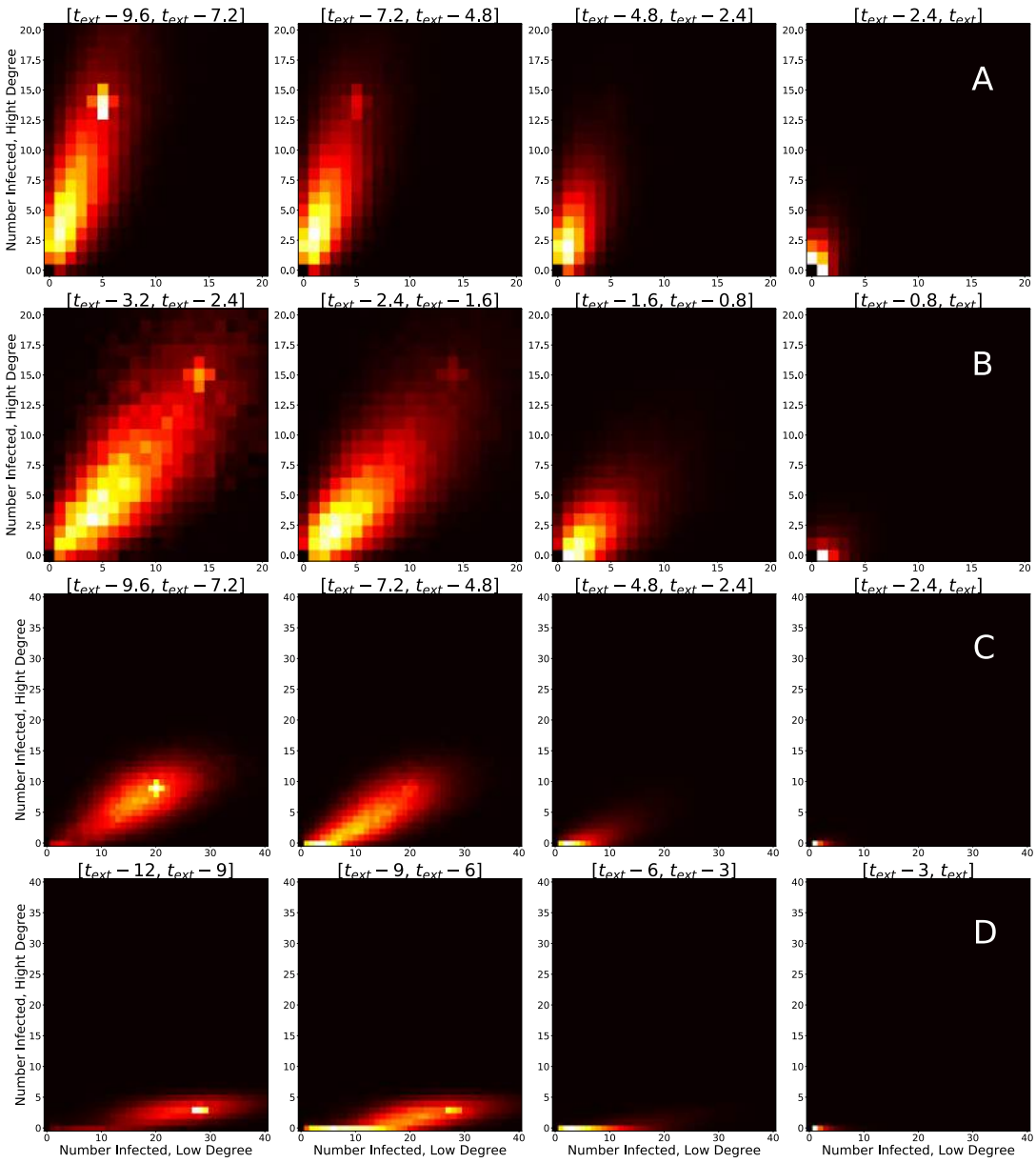

Figure 3.18: **Paths to Extinction for Networks with Varying Heterogeneity**: Each row corresponds to a different network. Within each row, each panel shows a heat map representing the ensemble of trajectories for a particular time interval prior to extinction (t_{ext}), proceeding from early times towards the time of extinction from left to right. Each heat map represents a superposition of trajectories that outlines the characteristic path to extinction.

APPENDIX A

INTERPRETING TOPIC MODEL OUTPUT

This table lists the properties of each of the $N = 50$ topics identified using LDA trained on the condensed matter physics (cond-mat) articles on arXiv. **Keywords** represent a few of the words most strongly associated with a topic. The **Sample Article** is the reference number of an article with a high probability assigned to a topic. We interpret each topic as representing a research field, and the **Interpretation** is the name that we use to refer to that research field.

Topic #	Sample Keywords	Sample Article	Interpretation
1	critical scaling ising transition temperature spin glass dimension random order phase correlation lattice	cond-mat/9709165	Spin Glasses; Magnetic Frustration
2	network node distribution degree random graph complex dynamic population scalefree market pattern	cond-mat/9709165	Complex Networks; Population Dynamics
3	condensate boseeinstein atom trap gas bose interaction potential condensation trapped atomic bec	0812.0499	Bose-Einstein Condensates

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
4	pressure phase alloy compound gpa temperature transition structural crystal lattice diffraction xray superconductivity	0712.2955	Superconducting Phases; High Pressure Phases
5	quantum state qubit entanglement spin dot decoherence coupling single control gate coupled information	0903.2030	Quantum Computing; Quantum Information
6	dynamic noise quantum state oscillator dynamical regime frequency nonequilibrium driven coupled evolution fluctuation	1411.2637	Quantum Oscillators
7	spin magnetic ferromagnetic magnetization effect current anisotropy polarization exchange layer interaction coupling	1205.2835	Spins in Materials; Spintronics

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
8	experimental recent theoretical physic experiment phenomenon present review work physical discus understanding	1306.1774	Review Articles
9	coupling interaction spinorbit phonon electronphonon effect phonons electron strong mode rashba polaron	cond-mat/9911404	Polarons
10	phase transition diagram order critical temperature point state quantum region firstorder behavior	cond-mat/0602237	Phase Transitions; Quantum Phase Transitions
11	quantum optical dot exciton semiconductor emission electron energy excitons hole laser excitation	0906.3260	Quantum Dots; Mesoscale Physics

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
12	temperature conductivity thermal transport dependence low effect resistivity heat coefficient scattering thermoelectric	cond-mat/0210047	Transport Measurements
13	wave soliton nonlinear periodic lattice potential instability velocity oscillation mode dynamic propagation	0904.4417	Solitons; Stationary States
14	vortex magnetic pinning lattice superfluid flux core superconductors current critical superconducting defect	cond-mat/9908317	Superconductor Vortices
15	scanning microscopy measurement tunneling image force local tip surface imaging probe atomic resolution	1009.2393	Microscopy

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
16	device material application design control cell efficiency performance memory potential power circuit technology	0804.1389	Electronic Devices
17	approximation density potential energy calculation solution effective functional exact expression expansion order	cond-mat/0007282	Mathematical Physics
18	spin lattice chain magnetic quantum heisenberg state interaction antiferromagnetic phase order exchange	1404.0194	Magnetic Frustration; Spin Chains & Lattices
19	magnetic temperature heat measurement magnetization susceptibility transition specific crystal single compound ferromagnetic	1411.2135	Magnetic Material Properties

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
20	gas lattice atom interaction fermi superfluid optical boson fermion ultracold state quantum	0806.4310	Ultracold Atoms Dynamics
21	film thin layer substrate temperature sample surface growth thickness grown deposition nanoparticles	1502.07223	Oxide Thin Films
22	spin relaxation magnetic nuclear electron temperature rate resonance nmr dynamic frequency hyperfine	1501.02897	Nuclear Magnetic Resonance
23	quantum hall electron magnetic state effect landau level fractional twodimensional edge filling	1109.6219	Quantum Hall Effect

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
24	nanotube carbon nanowires transistor device gate channel effect voltage nanowire tube contact transport	1112.4397	Electronic Devices; Nanoscale Devices
25	polymer chain protein interaction dna solution simulation length charge molecule force charged concentration	cond-mat/0504108	Soft Condensed Matter; Polymer Physics
26	impurity disorder interaction kondo liquid localization effect disordered electron quantum fermi anderson	1209.1606	Disordered Systems
27	frequency optical cavity mode light microwave wave resonance resonator dielectric radiation photonic	1212.0237	Optics; Metamaterials

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
28	dynamic glass liquid temperature simulation relaxation transition molecular water density fluid correlation glassy	1209.3401	Glasses
29	graphene layer edge bilayer electronic dirac gap band monolayer graphite sheet nanoribbons	1309.5398	Graphene
30	topological symmetry state insulator phase quantum fermion gauge dirac chiral majorana breaking edge	cond-mat/0506581	Topological Phases
31	simulation monte carlo algorithm problem numerical quantum present efficient technique scheme calculation	0705.4173	Simulation Methods; Monte Carlo

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
32	quantum dot transport conductance tunneling electron current effect voltage charge lead contact junction	0706.2950	Mesoscale Transport
33	scattering mode spectrum excitation peak frequency energy optical raman neutron inelastic phonon	cond-mat/0308170	Inelastic Scattering Experiments
34	distribution random correlation matrix statistic fluctuation probability gaussian ensemble large statistical density	cond-mat/9704191	Condensed Matter Theory; Random Matrices

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
35	flow particle granular fluid velocity shear force dynamic simulation friction hydrodynamic viscosity	cond-mat/9511105	Soft Condensed Matter; Granular Physics
36	current junction josephson superconducting ring magnetic flux critical effect array wire temperature tunnel	cond-mat/9811017	Superconducting Devices; Josephson Junctions
37	entropy equilibrium energy nonequilibrium fluctuation statistical heat thermodynamic distribution relation thermodynamics theorem temperature	1111.7014	Thermodynamics

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
38	hubbard interaction electron correlation charge mott state lattice insulator correlated phase coulomb band hopping	cond-mat/0508385	Mott-Hubbard Model
39	band surface fermi state electronic gap electron energy photoemission calculation level spectroscopy	1101.5615	Electronic Spectra; ARPES
40	scaling exponent percolation size cluster dimension critical alpha lattice law fractal distribution	cond-mat/0608223	Critical Phenomena
41	stress elastic strain material deformation dislocation shear modulus mechanical crack solid response fracture	cond-mat/0410642	Mechanical Properties of Materials

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
42	state energy ground bound number spectrum density excited level particle potential excitation	cond-mat/9712133	Quantum States
43	superconducting superconductivity doping superconductors cuprates temperature order state pseudogap magnetic charge	1504.06972	Cuprate Superconductors
44	magnetic ferroelectric phase transition orbital ordering polarization temperature order manganite state charge	1309.0291	Ferroelectrics
45	matrix quantum entanglement operator boundary lattice chain entropy exact group solution spin representation	cond-mat/0211081	Condensed Matter Theory

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
46	superconducting state superconductors superconductivity superconductor gap pairing symmetry dwave temperature order pair	cond-mat/0307345	Superconductivity
47	surface interface domain wall growth boundary force nucleation droplet bulk substrate layer	0809.1779	Surface Physics; Surface Chemistry
48	calculation atom energy density molecule electronic surface functional molecular cluster defect hydrogen	1312.4272	Density Functional Theory

Continued on next page

Table A.1 – continued from previous page

#	Keywords	Article	Interpretation
49	particle diffusion process motion dynamic brownian rate reaction random stochastic transport probability	1207.6190	Nonequlibrium Stat Mech; Stochastic Processes
50	crystal membrane nematic liquid surface curvature defect order rod orientation elastic phase	1304.0575	Soft Condensed Matter; Structured Fluids

APPENDIX B

NETWORK ASSEMBLY RESULTS FOR ALL TOPICS

This table summarizes the behavior of each topic's corresponding co-authorship network. For each topic (denoted by # and **Interpretation**), we show the number of articles for both the arXiv and Web of Science data sets (**# Articles arXiv** and **# Articles WoS**, respectively). Also shown is the assembly behavior of the co-authorship network for each topic (**GC Transition**). Referring back to Figures 2.6 and 2.7, "No GC" refers to no giant component formation, where cliques of authors remain disjointed. "Treelike GC" refers to cases where cliques of authors join together to form an extended, treelike giant component that has a large diameter. "Dense GC" refers to cases where cliques join together to form a densely connected giant component with many overlapping cliques and a small diameter.

#	Interpretation	# Articles arXiv	GC Transition arXiv	# Articles WoS	GC Transition WoS
1	Spin Glasses; Magnetic Frustration	1558	Dense GC	1765	Dense GC
2	Complex Networks; Population Dynamics	2677	Dense GC	731	Dense GC
3	Bose-Einstein Condensates	1020	Dense GC	105	No GC
4	Superconducting Phases; High Pressure Phases	695	Dense GC	3780	Dense GC
5	Quantum Computing; Quantum Information	1135	Dense GC	677	Dense GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
6	Quantum Oscillators	523	No GC	238	No GC
7	Spins in Materials; Spintronics	840	Dense GC	1461	Dense GC
8	Review Articles	369	No GC	429	No GC
9	Polarons	60	No GC	85	No GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
10	Phase Transitions; Quantum Phase Transitions	60	No GC	52	No GC
11	Quantum Dots; Mesoscale Physics	821	Dense GC	6489	Dense GC
12	Transport Measurements	366	No GC	1189	Dense GC
13	Solitons; Stationary States	280	Treelike GC	233	Treelike GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
14	Superconductor Vortices	450	Treelike GC	756	Dense GC
15	Microscopy	122	No GC	439	Treelike GC
16	Electronic Devices	324	Treelike GC	5375	Dense GC
17	Mathematical Physics	752	Treelike GC	786	Treelike GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
18	Spin Chains & Lattices	1539	Dense GC	1623	Dense GC
19	Magnetic Material Properties	1087	Dense GC	5714	Dense GC
20	Ultracold Atoms Dynamics	1300	Dense GC	104	No GC
21	Oxide Thin Films	1116	Treelike GC	54823	Dense GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
22	Nuclear Magnetic Resonance	196	Treelike GC	620	Treelike GC
23	Quantum Hall Effect	742	Dense GC	799	Dense GC
24	Electronic Devices; Nanoscale Devices	422	Treelike GC	3663	Dense GC
25	Soft Condensed Matter; Polymer Physics	1276	Dense GC	993	Dense GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
26	Disordered Systems	444	No GC	285	No GC
27	Optics; Metamaterials	676	Treelike GC	2125	Dense GC
28	Glasses	1187	Dense GC	844	Dense GC
29	Graphene	402	Dense GC	392	Dense GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
30	Topological Phases	913	Dense GC	433	Dense GC
31	Simulation Methods; Monte Carlo	754	No GC	356	Treelike GC
32	Mesoscale Transport	1416	Dense GC	1774	Dense GC
33	Inelastic Scattering Experiments	220	Treelike GC	756	Dense GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
34	Condensed Matter Theory	641	No GC	100	No GC
35	Soft Condensed Matter; Granular Physics	1375	Dense GC	421	No GC
36	Superconducting Devices	596	Dense GC	1191	Dense GC
37	Thermodynamics	1574	Treelike GC	140	No GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
38	Mott-Hubbard Model	798	Dense GC	901	Dense GC
39	Electronic Spectra; ARPES	457	Dense GC	1451	Dense GC
40	Critical Phenomena	786	No GC	109	No GC
41	Mechanical Material Properties	525	No GC	2345	Dense GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
42	Quantum States	62	No GC	29	No GC
43	Cuprate Superconductors	1030	Dense GC	858	Dense GC
44	Ferroelectrics	1043	Dense GC	2591	Dense GC
45	Condensed Matter Theory	1595	Treelike GC	467	Treelike GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
46	Superconductivity	576	Dense GC	528	Dense GC
47	Surface Physics; Surface Chemistry	467	No GC	938	Dense GC
48	Density Functional Theory	1002	Dense GC	8439	Dense GC
49	Nonequlibrium Stat Mech	993	Treelike GC	130	No GC

Continued on next page

Table B.1 – continued from previous page

#	Interpretation	# Articles arXiv	GC ArXiv	# Articles WoS	GC WoS
50	Soft Condensed Matter; Structured Fluids	432	Treelike GC	139	Treelike GC

APPENDIX C

SECOND DERIVATION OF CUMULANT EQUATIONS

An alternative way to derive the cumulant equations (Eq. 3.19) is to directly calculate the time-dependent behavior of the moments of the QSD by averaging over all possible changes to the moments as defined by the master equation. This is known as the diffusion approximation [59, 60, 73] and is less mathematically detailed than the use of generating functions but is algebraically simpler and yields the same results.

Let the notation $\langle\cdot\rangle$ define the moment of $(\cdot)$ in the QSD. For example, $\langle x\rangle$ is the first moment of S, or the mean number susceptible. $\langle y\rangle$ is the mean number infected. For higher-order moments,

$$\langle xy\rangle = \sigma(xy) + \langle x\rangle\langle y\rangle = \sigma(xy) + \mu(x)\mu(y)$$
$$\langle x^2\rangle = \sigma^2(x) + \langle x\rangle^2 = \sigma^2(x) + \mu(x)^2$$
$$\langle y^2\rangle = \sigma^2(y) + \langle y\rangle^2 = \sigma^2(y) + \mu(y)^2$$

To find the time derivative of a quantity $\langle\cdot\rangle$, one calculates the ensemble average over all possible changes to the quantity $(\cdot)$:

$$\frac{d\langle\cdot\rangle}{dt} = \langle \sum_{\text{events}} (\text{change to quantity } (\cdot)) \times (\text{rate of change to quantity } (\cdot))\rangle$$

where the rates of change the ones defined for the SIRS model in Table 3.3.2.

To derive the time derivative of the mean number susceptible $\mu(x) = \langle x\rangle$:

$$\begin{aligned}\frac{d\langle x\rangle}{dt} &= (+1)\rho\langle N - x - y\rangle + (-1)\beta\langle xy\rangle/N \\ \frac{d\mu(x)}{dt} &= \rho\left(N - \mu(x) - \mu(y)\right) - \beta\left(\sigma(xy) + \mu(x)\mu(y)\right)/N\end{aligned} \tag{C.1}$$

To derive the time derivative of the mean number susceptible $\mu(y) = \langle y\rangle$:

$$\begin{aligned}\frac{d\langle y\rangle}{dt} &= (-1)\gamma\langle y\rangle + (+1)\beta\langle xy\rangle/N \\ \frac{d\mu(y)}{dt} &= -\gamma\mu(y) + \beta\left(\sigma(xy) + \mu(x)\mu(y)\right)/N\end{aligned} \tag{C.2}$$

The second-order moments are a little trickier: just as the equations for the first-order moments depend on the equations for the second-order moments, the exact second-order moment equations will depend on third-order moments. In order to avoid generating an infinite number of interdependent moment equations, one can apply a Gaussian moment closure approximation. A Gaussian distribution only has nonzero first-order and second-order cumulants, with all higher-order cumulants equal to zero. Assuming that the QSD of the endemic state is approximately Gaussian (as in the left hand side of Fig. 3.8), it is possible to truncate the infinite sequence of moment equations to only require the first- and second- order equations.

In practice, when deriving the second-order moment equations, a third-order moment with the form $\langle abc \rangle$ will appear. Gaussian moment closure makes it possible to re-write this third-order term as an algebraic combination of lower-order moments by assuming that the third-order cumulant $C(abc)$ is zero and expanding it in terms of its moments:

$$C(abc) = \langle abc \rangle - \langle ab \rangle \langle c \rangle - \langle ac \rangle \langle b \rangle - \langle bc \rangle \langle a \rangle + 2\langle a \rangle \langle b \rangle \langle c \rangle$$

$$C(abc) = 0$$

$$\Rightarrow \langle abc \rangle = \langle ab \rangle \langle c \rangle + \langle ac \rangle \langle b \rangle + \langle bc \rangle \langle a \rangle - 2\langle a \rangle \langle b \rangle \langle c \rangle$$

To derive the time derivative of the covariance $\sigma(xy) = \langle xy \rangle - \langle x \rangle \langle y \rangle$ requires calculation of the change to the quantity (xy) for each possible type of event. For an infection event, the change to the quantity (xy) is $((x-1)(y+1)) - xy = x - y - 1$. For a recovery event, the change to the quantity (xy) is $(x(y-1) - xy = -x$. For a

loss of immunity event, the change to the quantity (xy) is $(x+1)y - xy = y$.

$$\begin{aligned}
\frac{d\sigma(xy)}{dt} &= \frac{d\langle xy\rangle}{dt} - \langle x\rangle\frac{d\langle y\rangle}{dt} - \langle y\rangle\frac{d\langle x\rangle}{dt} \\
\frac{d\langle xy\rangle}{dt} &= \langle (x-y-1)\,\beta xy/N\rangle + \langle (-x)\gamma y\rangle + \langle (+y)\,(N-x-y)\rangle \\
&= -\gamma\langle xy\rangle + \rho\left(N\langle y\rangle - \langle xy\rangle - \langle y^2\rangle\right) - \beta\langle xy\rangle/N \\
&\quad + \beta\left(2\langle xy\rangle\langle x\rangle + \langle x^2\rangle\langle y\rangle - 2\langle x\rangle^2\langle y\rangle\right)/N \\
&\quad - \beta\left(2\langle xy\rangle\langle y\rangle - \langle x\rangle\langle y^2\rangle - 2\langle y\rangle^2\langle x\rangle\right)/N \\
\langle x\rangle\frac{d\langle y\rangle}{dt} &= -\gamma\langle x\rangle\langle y\rangle + \beta\langle x\rangle\langle xy\rangle/N \\
\langle y\rangle\frac{d\langle x\rangle}{dt} &= \rho\left(N\langle y\rangle - \langle x\rangle\langle y\rangle - \langle y\rangle^2\right) - \beta\langle y\rangle\langle xy\rangle/N
\end{aligned}$$

Simplifying,

$$\begin{aligned}
\frac{d\sigma(xy)}{dt} &= -\beta\left(\mu(x)\mu(y) + \sigma(xy)\right)/N - \gamma\sigma(xy) - \rho\left(\sigma(xy) + \sigma^2(y)\right) \\
&\quad + \beta\left(\mu(y)\sigma^2(x) + \mu(x)\sigma(xy) - \mu(x)\sigma^2(y) - \mu(y)\sigma(xy)\right)/N
\end{aligned} \tag{C.3}$$

Similarly, one may derive the time derivatives for each of the variances $\sigma^2(x)$ and $\sigma^2(y)$ using the same procedure, obtaining:

$$\begin{aligned}
\frac{d\sigma^2(x)}{dt} &= \beta\left(\mu(x)\mu(y) + \sigma(xy)\right)/N + \rho\left(N - \mu(x) - \mu(y)\right) \\
&\quad - 2\beta\left(\mu(x)\sigma(xy) + \mu(y)\sigma^2(x)\right)/N - 2\rho\left(\sigma(xy) + \sigma^2(x)\right)
\end{aligned} \tag{C.4}$$

and

$$\begin{aligned}
\frac{d\sigma^2(y)}{dt} &= \beta\left(\mu(x)\mu(y) + \sigma(xy)\right)/N + \gamma\mu(y) \\
&\quad + 2\beta\left(\mu(y)\sigma(xy) + \mu(x)\sigma^2(y)\right)/N - 2\rho\sigma^2(y)
\end{aligned} \tag{C.5}$$

Note that each of the above equations (Eqs. C.1, C.2, C.3, C.4, C.5) is identical to the analogous equation derived directly from the master equation shown above in Eq. 3.19.

This procedure may also be repeated for the heterogeneous SIRS model (Eq. 3.23), where this time the interaction term depends on a sum over K different classes. The result is the same as Eq. 3.24.

APPENDIX D

ADDITIONAL PLOTS

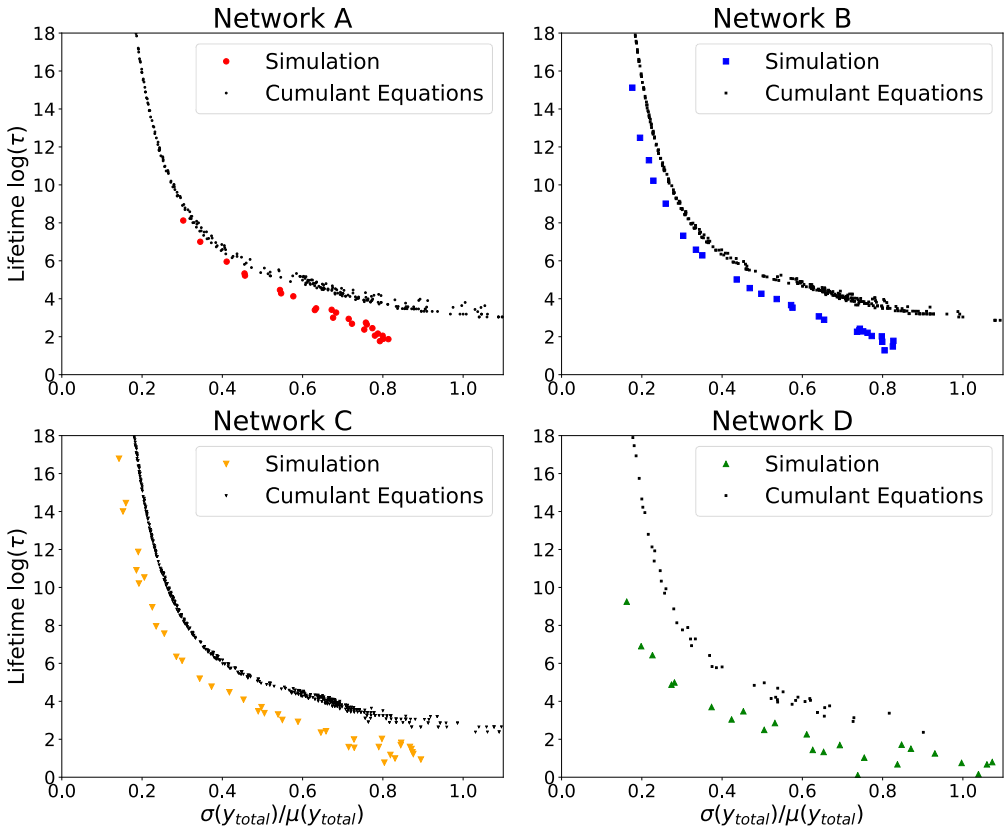

Figure D.1: **Comparison of Cumulant Equations Against Each Network, Shown Separately**: Referring back to Fig 3.16, which plots τ vs $\sigma(y_{total})/\mu(y_{total})$ for each network (A, B, C, D). The data for all four networks are superposed together. For the sake of clarity, this figure shows the same four data sets plotted separately, with each one juxtaposed with the cumulant equations' predictions. Once again, the cumulant equations are qualitatively consistent with the simulation data, but systematically overestimate τ.

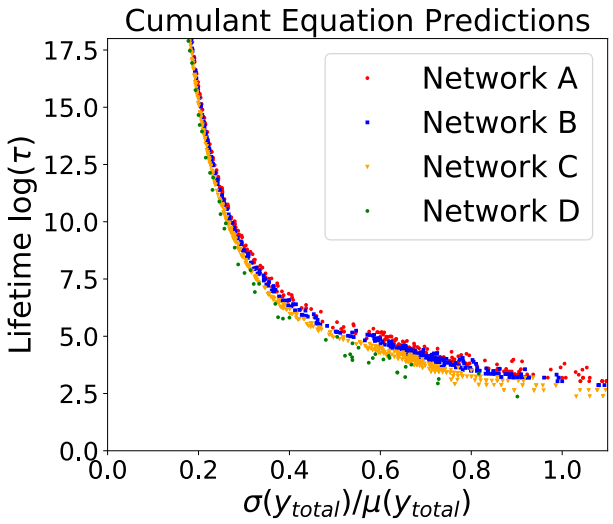

Figure D.2: **Comparison of Cumulant Equations Against Each Network, Shown Separately**: The cumulant equations' predictions, seen in Fig. D.1, juxtaposed together. Once again, the data collapse together to outline a relatively simple, low-dimensional relationship between τ vs $\sigma(y_{total})/\mu(y_{total})$.

BIBLIOGRAPHY

[1] ArXiv Submission Rate Statistics, 2012. Accessed online at http://arxiv.org/help/stats/2012_by_area/index.

[2] ArXiv Submission Rate Statistics, 2016. Accessed online at http://arxiv.org/help/stats/2016_by_area/index.

[3] Certain data included herein are derived from Clarivate Analytics Web of Science TM. ©Copyright Clarivate Analytics 2017. All rights reserved.

[4] Cornell University - College Of Arts and Sciences - Plagiarism. Accessed online at http://plagiarism.arts.cornell.edu/tutorial/index.cfm.

[5] Federal Office of Research Integrity, Malfeasance and Misconduct - Definitions, Accessed online at https://ori.hhs.gov/education/products/ucla/chapter8/default.htm.

[6] Physical Review D Editorial Policies and Practices, Accessed online at https://journals.aps.org/prd/authors/editorial-policies-practices.

[7] Alberto Aleta, Andreia NS Hisi, Sandro Meloni, Chiara Poletto, Vittoria Colizza, and Yamir Moreno. Human mobility networks and persistence of rapidly mutating pathogens. *Royal Society Open Science*, 4(3):160914, 2017.

[8] Roy M Anderson, Robert M May, et al. Population biology of infectious diseases: Part i. *Nature*, 280(5721):361–367, 1979.

[9] Jesús R Artalejo. On the time to extinction from quasi-stationarity: A unified approach. *Physica A: Statistical Mechanics and its Applications*, 391(19):4483–4486, 2012.

[10] Normal T. J. Bailey. *The Elements of Stochastic Processes*. John Wiley & Sons, 1964.

[11] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[12] Maurice S Bartlett. Measles periodicity and community size. *Journal of the Royal Statistical Society. Series A (General)*, 120(1):48–70, 1957.

[13] MS Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 4, page 109, 1956.

[14] MS Bartlett. The critical community size for measles in the united states. *Journal of the Royal Statistical Society. Series A (General)*, pages 37–44, 1960.

[15] Luis M A Bettencourt and David I Kaiser. Formation of Scientific Fields as a Universal Topological Transition. *arXiv.org*, April 2015.

[16] Luis M A Bettencourt, David I Kaiser, and Jasleen Kaur. Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3):210–221, July 2009.

[17] Mario Biagioli. Recycling Texts or Stealing Time?: Plagiarism, Authorship, and Credit in Science. *International Journal of Cultural Property*, 19(03):453–476, December 2012.

[18] Francis L Black. Measles endemicity in insular populations: critical community size and its evolutionary implication. *Journal of Theoretical Biology*, 11(2):207–211, 1966.

[19] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[20] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[21] Johan Bollen, Herbert Van de Sompel, Aric Hagberg, Luis Bettencourt, Ryan Chute, Marko A Rodriguez, and Lyudmila Balakireva. Clickstream Data Yields High-Resolution Maps of Science. *PLoS One*, 4(3):e4803, March 2009.

[22] Katy Börner, Noshir Contractor, Holly J. Falk-Krzesinski, Stephen M. Fiore, Kara L. Hall, Joann Keyton, Bonnie Spring, Daniel Stokols, William Trochim, and Brian Uzzi. A multi-level systems perspective for the science of team science. *Science Translational Medicine*, 2(49):49cm24–49cm24, 2010.

[23] Katy Börner and Richard M Shiffrin. Mapping knowledge domains. *Proceedings of the National Academy of Sciences*, 101:5183–5185, January 2004.

[24] Kevin W Boyack, Richard Klavans, and Katy Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005.

[25] Samuel A Bozzette, Rob Boer, Vibha Bhatnagar, Jennifer L Brower, Emmett B Keeler, Sally C Morton, and Michael A Stoto. A model for a smallpox-vaccination policy. *New England Journal of Medicine*, 348(5):416–425, 2003.

[26] Geoff Brumfiel. Turkish physicists face accusations of plagiarism, 2007.

[27] Charles D Brummitt, Raissa M D?Souza, and Elizabeth A Leicht. Suppressing cascades of load in interdependent networks. *Proceedings of the National Academy of Sciences*, 109(12):E680–E689, 2012.

[28] Sergey V Buldyrev, Roni Parshani, Gerald Paul, H Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1029, 2010.

[29] Ronald S Burt. Structural Holes and Good Ideas. *American journal of sociology*, 110(2):349–399, September 2004.

[30] Claudio Castellano and Romualdo Pastor-Satorras. Thresholds for epidemic spreading in networks. *Physical review letters*, 105(21):218701, 2010.

[31] Daniel T Citron and Paul Ginsparg. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences*, 112(1):25–30, 2015.

[32] Daniel T Citron and Samuel F Way. Network assembly of scientific communities of varying size and specificity. *Journal of Informetrics*, in review.

[33] Damian Clancy and Sang Taphou Mendy. The effect of waning immunity on long-term behaviour of stochastic models for the spread of infection. *Journal of Mathematical Biology*, 61(4):527–544, November 2009.

[34] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020, 2006.

[35] Vittoria Colizza, Marc Barthélemy, Alain Barrat, and Alessandro Vespignani. Epidemic modeling in complex realities. *Comptes rendus biologies*, 330(4):364–374, 2007.

[36] J N Darroch and E Seneta. On quasi-stationary distributions in absorbing continuous-time finite markov chains. *Journal of Applied Probability*, 4(1):192, April 1967.

[37] Derek J de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[38] Derek J de Solla Price. *Little Science, Big Science... and Beyond*. Columbia University Press, 1986.

[39] David R de Souza and Tânia Tomé. Stochastic lattice gas model describing the dynamics of the SIRS epidemic process. *Physica A: Statistical Mechanics and its Applications*, 389(5):1142–1150, March 2010.

[40] Ian Dobson, Benjamin A Carreras, Vickie E Lynch, and David E Newman. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2):026103, 2007.

[41] Andrew J Dolgert. Discrete stochastic models in continuous time for ecology. *arXiv preprint arXiv:1506.08483*, 2015.

[42] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[43] Toni Feder. Experimenting with plagiarism detection on the arxiv. *Physics Today*, 60(3):30, 2007.

[44] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[45] E. Garfield and I. H. Sher. New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14(3):195–201, 1963.

[46] Jim Giles. Preprint server seeks way to halt plagiarists, 2003.

[47] P Ginsparg, P Houle, T Joachims, and J H Sul. Mapping subsets of scholarly information. *Proceedings of the National Academy of Sciences*, 101:5236–5240, April 2004.

[48] Paul Ginsparg. Arxiv at 20. *Nature*, 476(7359):145–147, 2011.

[49] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[50] Michelle Girvan, Duncan S. Callaway, M.E.J. Newman, and Steven H. Strogatz. Simple model of epidemics with pathogen mutation. *Physical Review E*, 65(031915), March 2002.

[51] Nicholas C Grassly, Christophe Fraser, and Geoffrey P Garnett. Host immunity and synchronized epidemics of syphilis across the united states. *Nature*, 433(7024):417, 2005.

[52] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[53] Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[54] Zhen Guo, Shenghuo Zhu, Yun Chi, Zhongfei Zhang, and Yihong Gong. A latent topic model for linked documents. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 720–721. ACM, 2009.

[55] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[56] Jason Hindes and Ira B Schwartz. Epidemic extinction and control in heterogeneous networks. *Physical review letters*, 117(2):028302, 2016.

[57] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, 2002.

[58] D. Kaiser. *Drawing Theories Apart: The Dispersion of Feynman Diagrams in Postwar Physics*. University of Chicago Press, 2005.

[59] M J Keeling. Metapopulation moments: coupling, stochasticity and persistence. *Journal of Animal Ecology*, 2000.

[60] Matt Keeling and Pej Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

[61] Jeffrey O Kephart, Steve R White, and David M Chess. Computers and epidemiology. *IEEE Spectrum*, 30(5):20–26, 1993.

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

[63] Vincent Larivière, Cassidy R. Sugimoto, Benoit Macaluso, Staša Milojević, Blaise Cronin, and Mike Thelwall. arxiv e-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, 65(6):1157–1169, 2014.

[64] Deokjae Lee, K-I Goh, B Kahng, and D Kim. Complete trails of coauthorship network evolution. *Physical Review E*, 82(2):026112, 2010.

[65] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 551–556. SIAM, 2007.

[66] Travis Martin, Brian Ball, Brian Karrer, and M E J Newman. Coauthorship and citation patterns in the Physical Review. *Physical Review E*, 88(1):012814, July 2013.

[67] Robert M May, Roy M Anderson, et al. Population biology of infectious diseases: Part ii. *Nature*, 280(5722):455–461, 1979.

[68] Andrew K McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[69] Lauren Ancel Meyers, Babak Pourbohloul, Mark EJ Newman, Danuta M Skowronski, and Robert C Brunham. Network theory and sars: predicting outbreak diversity. *Journal of theoretical biology*, 232(1):71–81, 2005.

[70] Miguel A Muñoz, Róbert Juhász, Claudio Castellano, and Géza Ódor. Griffiths phases on complex networks. *Physical review letters*, 105(12):128701, 2010.

[71] Christopher R Myers, David J Schneider, and Sarabjeet Singh. The structure of infectious disease outbreaks across the animal-human interface. *Bulletin of the American Physical Society*, 60, 2015.

[72] Noboru Nakanishi and Izumi Ojima. Notes on Unfair Papers by Mebarki et al. on "Quantum Nonsymmetric Gravity". *arXiv.org*, December 1999.

[73] Ingemar Nåsell. On the time to extinction in recurrent epidemics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):309–330, 1999.

[74] Ingemar Nåsell. Stochastic models of some endemic infections. *Mathematical biosciences*, 179(1):1–19, 2002.

[75] M. E. J. Newman. Scientific collaboration networks. I. network construction and fundamental results. *Phys. Rev. E*, 64:016131, Jun 2001.

[76] M. E. J. Newman. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, Jun 2001.

[77] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[78] M E J Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5200–5205, April 2004.

[79] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[80] G Nodelijk, MCM De Jong, A Van Nes, JCM Vernooy, LAMG Van Leengoed, JMA Pol, and JHM Verheijden. Introduction, persistence and fade-out of porcine reproductive and respiratory syndrome virus in a dutch breeding herd: a mathematical analysis. *Epidemiology & Infection*, 124(1):173–182, 2000.

[81] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63(066117), May 2001.

[82] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[83] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical Review E*, 65(3):036104, 2002.

[84] John J Potterat, L Phillips-Plummer, Stephen Q Muth, RB Rothenberg, DE Woodhouse, TS Maldonado-Long, HP Zimmerman, and JB Muth. Risk network structure in the early epidemic phase of hiv transmission in colorado springs. *Sexually transmitted infections*, 78(suppl 1):i159–i163, 2002.

[85] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press, 2004.

[86] Saul Schleimer, Daniel S Wilkerson, and Alex Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85. ACM, 2003.

[87] Vishal Sood and Sidney Redner. Voter model on heterogeneous graphs. *Physical review letters*, 94(17):178701, 2005.

[88] Daria Sorokina, Johannes Gehrke, Simeon Warner, and Paul Ginsparg. Plagiarism detection in arxiv. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 1070–1075. IEEE, 2006.

[89] Steven H. Strogatz. *Nonlinear Dynamics and Chaos*. Westview Press, 1st edition, 2001.

[90] Xiaoling Sun, Jasleen Kaur, Staša Milojević, Alessandro Flammini, and Filippo Menczer. Social dynamics of science. *Scientific Reports*, 3, 2013.

[91] A N Tabah. Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual Review of Information Science and Technology*, 34:249–86, 1999.

[92] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433. ACM, 2006.

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

[94] H Wilf. Generatingfunctionology,(1990). *ISBN: 0-12-751956-4*.

[95] Fang Wu, Bernardo A Huberman, Lada A Adamic, and Joshua R Tyler. Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(1):327–335, 2004.

[96] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.