MULTISCALE DYNAMICS IN HONEYBEE SOCIETIES

BY

VIKYATH DEVIPRASAD RAO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Sergei Maslov, Chair
Professor Nigel Goldenfeld, Director of Research
Associate Professor R. E. Lee DeVille
Assistant Professor Seppe Kuehn

Abstract

In this dissertation, I examine the social organization of a model organism, the honeybee, at multiple scales. I begin in Part I at the microbial scale, by studying the relationship between the social caste of individuals and the microbes they harbour in their gastrointestinal tracts. Using 16S rRNA sequence data, I reconstruct the gut microbiomes of honeybees of different castes. I find that the microbiomes of two previously-uncharacterized social castes—drones and queens—contain the same bacteria as those in the guts of worker bees. However, despite this similarity, I show that the compositions of these bacteria in drones and queens are sufficiently different that their microbiomes can be distinguished from those of workers.

In Part II, I study the honeybee society at the level of its individual constituents, in particular, the set of foragers. I characterize the distribution of foraging activity across these individuals in the society, and find that this is highly skewed, with some individuals contributing much more to the activity of the colony than others. I establish these results in the framework used to describe the wealth of individuals in human society, and also characterize the temporal variation and resilience of foraging activity.

In Part III, I describe a system to track individual honeybees and their interactions inside a two-dimensional observation hive with high spatiotemporal resolution. At the level of individual honeybees, I study the temporal statistics of trophallaxis, an important social interaction that occurs in honeybee societies, and find that the distribution of trophallaxis durations is similar to the distribution of face-to-face interactions among humans. I propose a scaling argument to explain the scaling exponent of these distributions, and test the argument in simple random-walk models of proximity interactions. I then study the honeybee society at the collective scale of the trophallaxis interaction network, and find that although bees exhibit bursty patterns of trophallaxis just as humans do in communication, the dynamics of simulated spreading on the trophallaxis networks is fast relative to randomized reference models, unlike in human temporal networks.

Acknowledgments

The person I have to thank first and foremost is my advisor, Nigel Goldenfeld. Nigel's incredible energy, truly wide-ranging interests, scientific leadership and personal kindness are exceptionally inspiring. This dissertation would not have been possible without his constant supply new and creative ideas, his ability to see connections that traverse disciplinary boundaries, and his vision in choosing which problems to work on. Our many long meetings trying out the latest crazy idea, sifting through the literature, focusing and rethinking our results, and skirting the inevitable hurdles of research, have all shaped my transition from a rather bookish onlooker of science to a researcher engaged in doing science. Nigel is also staunchly dedicated in caring for his students, both personally and professionally, and is exemplary in treating them fairly. I consider myself incredibly fortunate to have been able to grow up as a researcher under his mentorship, and to know him as a friend and a scientist.

Nigel also fostered a synergistic group environment in which his students freely share both progress and sticking points in their research activities, and in this context I've learnt much from my group-mates from years past and present: Michael Martini, Farshid Jafarpour, Maksim Sipos, Hong-Yan Shih, Tommaso Biancalani, Chi Xue, Minhui Zhu, Purba Chatterjee, Patricio Jeraldo and Luiza Angheluta. Thanks to Michael and Farshid in particular for their many years of friendship, and for sharing in numerous geeky discussions (though it is usually Nigel who one-ups all of us in being up-to-date on the tech world).

The research presented in this dissertation is multidisciplinary, and it has been possible thanks to our collaboration with Gene Robinson's group. I would like to thank Gene for trusting that someone with no training in biology could contribute to the projects that originated in his lab. Gene also took a lot of time out of his busy schedule to address our often-naïve questions with patient, thorough and insightful answers during meetings, via e-mail, or on our collaborative blog. I've been lucky in being able to work closely with several members of Gene's lab: Tim Gernat, Paul Tenczar, Claudia Lutz, Karen Kapheim and Adam Hamilton. When entering graduate school, I couldn't have foreseen that I would be studying honeybees, and I'm grateful to all them for carefully leading me into the world of honeybees with numerous fun discussions along the way. Tim and I have spent lot of time working together on analyzing the data that his marvellous

apparatus produces, and our different perspectives have always enriched our conversations; thanks also to Weibing Deng for joining us in some of these discussions. I worked with Paul and Claudia on my very first project in Nigel's group, and enjoyed getting to hear about Paul's outdoor adventures every few weeks. I was introduced to the exciting field of metagenomics when Karen contacted us about her bee microbiome data, and I learnt a lot from her about both the microbiome and ecology.

I would also like to thank Seppe Kuehn, Lee DeVille and Sergei Maslov for serving on my dissertation committee, and to Seppe for several interesting discussions on microbial ecology. I would like to thank Yoshi Oono for serving on my preliminary exam committee, and for many wide-ranging and thought-provoking discussions. Yoshi's longtime involvement as a physicist studying honeybees meant that I benefited greatly from his unique perspectives and his in-depth knowledge of biology.

Having served as a graduate teaching assistant for several semesters, I now appreciate all the more the dedication of my teachers at Amherst College during my undergraduate years. I would like to particularly thank Jonathan Friedman, in whose lab I first got a taste for scientific research, Will Loinaz for teaching many of my undergraduate physics courses with his characteristic affability, all the other members of the physics and math departments, and John Servos for a memorable course on the history of science.

Finally, as a social creature like the honeybee, I would not have been able to survive the long and sometimes perilous journey of graduate school without the social support I received on so many fronts. Foremost, I would like to thank my parents for having the courage ten years ago to send me out to study in a country they had never visited. I'm grateful to them and to my brother for their unconditional love and support throughout. I've also been lucky enough to have made so many good friends during the past six years in Champaign-Urbana that I cannot list them all here for fear of leaving someone out. I hope they will pardon me and instead give me a chance to tell them in person how important they have been.

I received support from a University Fellowship in fall 2014. Some of the material in this dissertation is based upon work supported by the National Science Foundation under Grant No. BCS-1246920. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Some of the research reported in this dissertation was also partially supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under award number R01GM117467. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Table of Contents

List of Figures

Chapter 1

Introduction and overview

In this dissertation, I take a multiscale approach to understanding a living system, namely a honeybee society. Life is characterized by non-linearity and collective effects. This means that one can study life at the level of genes, cells, biofilms, organisms, superorganisms or ecosystems, but to truly gain conceptual insights into its functioning, one has to also study the interactions between these disparate scales. In this chapter, I will argue for the relevance of a physics approach to addressing this issue, and I provide an overview of the research described in this dissertation.

1.1 Why should a physicist study honeybees?

In 1944, physicist Erwin Schrödinger published his book asking us *What is Life?* [104]. In this book, which was based on lectures for general audiences, Schrödinger speculated on the molecular basis upon which life operates, and his thinking influenced both Watson and Crick, who went on to co-discover the structure of the DNA molecule. However, to turn the question around somewhat, I wonder what Schrödinger would say today if we instead asked him, "what is physics?"

Physics has always been the broadest of the scientific disciplines in terms of the realms over which it asks questions. For this reason, I subscribe to the view that unlike the other sciences, physics is not defined by a domain on which it operates—for this does not seem such a useful definition—but rather by the choice of what it seeks to understand, and the methods it uses to address this goal. This is not to say that the methods of physics are always successful; indeed, as Nigel Goldenfeld likes to point out in his lectures, physics seems very successful only because it chooses to answer just those questions that are amenable to its methods. In the interest of turning away from philosophy and towards science, it suffices to state here that in some sense, physics always seeks to understand the essence of natural phenomena.[1]

Turning now to biology: as a discipline, biology (or the life sciences more broadly) has long been concerned with collecting observations about the subjects under its purview—living organisms—and trying to generalize

[1] In his stimulating book *The Nonlinear World* [89], Yoshi Oono calls this process "appreciating the world," and provides a much more thorough view of what constitutes the essence of natural phenomena.

from these facts. The sub-disciplines of biology are often created by a separation of length scales (ecology vs. cell biology) or time scales (evolution vs. population genetics). Because life is fundamentally a non-linear phenomenon, it has often been difficult to make biological predictions that go beyond the specific context from which the observations were derived.[2] This is now changing, as new technologies allow us to collect biological data that never existed before, and to do so at unprecedented scales. Furthermore, it is now becoming possible to bioengineer organisms and their interactions to test theories of how living systems work.

These advances are particularly relevant for the study of social behaviour, whether in the context of human societies or a society of eusocial insects like a honeybee colony. Today, biologists are no longer in the situation where they have to study a few individuals in a society over a some short observation period, and try to make general deductions about the organization of the society, despite the fact that social dynamics are in general far from equilibrium and are confounded by interference from different scales: genetic, microbial, environmental, and so on. Individual behaviour is not just determined by the social context, but also by genetics, and even by cross-species factors, such as interactions with microbes. For example, there is a body of work suggesting an association between the community of microbes living in human guts and autism [84], a condition in which social behaviour is overtly affected.

It is in this happy situation that the methods of physics now become useful. Given the wealth of data, we can take a physics approach to try to discern which features are specific to a particular model organism, and which are truly generic features of social behaviour. We can try to derive the simplest model that is able to capture generic features, much as a predator-prey model in ecology does not care whether the organisms are large mammals (such as a lynx and hare) or microbial (a bacteriophage and its bacterium). A honeybee society is particularly ideal as a model society because it provides a balance between simplicity and complexity: individual honeybee biology has been extensively studied and worker bees have well-defined social roles that are generally coupled to their age; yet a honeybee colony can behave as a single unit of biological organization even though no one individual is the central controller of the colony, and social roles are also modulated by the needs of the colony.

1.2 Overview of projects

In Part I of this dissertation, I study the effects of the microbial scale on the social structure of honeybees. The availability of metagenomics data is revolutionizing microbiology, and microbes are no longer simply

[2] I do not like Rutherford's famous phrase, "All science is either physics or stamp collecting," because it lends itself too easily to being used derisively. It is, however, possible to see it in a positive light if we agree that when we cannot derive wide-ranging laws like those in physics, the best we can do is to collect facts in the hope that one day we will do better.

seen as the cause of disease, but rather have much more profound effects on the larger organisms they inhabit, and thus on the societies that these organisms form. The field of metagenomics owes its existence to the pioneering use of the 16S rRNA gene in reconstructing the evolutionary history of microbes by Carl Woese, who was trained as a physicist. A honeybee colony has well-defined societal roles that are collectively regulated to allow the colony to function. I answer the question: is it possible to distinguish an individual bee's social caste not by looking at its activities but instead by simply looking at the microbes that inhabit its guts?

The remainder of the dissertation is concerned with studying the relationship between two scales of social organization: the individual and the collective, or societal level. In Part II, I describe our study of the division of foraging labour in a honeybee colony. The ability to track individual flights allows us to quantitatively measure how a honeybee colony has collectively chosen to distribute activity levels among its constituent workers. We establish these results using the framework of wealth distributions used in economics. We also study how the flight activity of individuals varies in times, and how there is a reserve workforce that increases their activity in response to a loss of foragers.

In Part III, I describe an apparatus that tracks every single individual inside a fully-functioning honeybee hive with high spatiotemporal resolution over extended periods of time. Using the data generated from these experiments, we measure the non-equilibrium dynamics of the colony and track interactions that occur inside the colony between individuals. We also measure the statistics of interactions in honeybee colonies, and find that there are similarities with human communication networks, some of which can be explained with simple models that capture features that are common to both kinds of interaction. Next, using network theory, we study both the aggregate statistical structure of these interactions and probe the temporal structure of the interactions in the framework of temporal networks. This allows us to study the relationship between the individual and the network levels of organization. We find that surprisingly, the temporal networks support fast spreading dynamics relative to randomized reference networks, unlike human communication networks.

1.3 My contributions

All the analyses and simulations that I performed for this dissertation were done under the guidance of Nigel Goldenfeld, who calculated and devised mathematical models, oversaw the development of the metagenomics pipeline, and suggested analyses to perform, graphs to plot and bodies of theory to use in analyses. As detailed below, the research in this dissertation was performed in collaboration with several members of Gene Robinson's honeybee laboratory in the Department of Entomology and the Woese Institute for Genomic

Biology (IGB), and with input and guidance on all aspects of honeybee biology from Gene Robinson.

The microbiome study described Part I was performed in close collaboration with Karen Kapheim, who was then a postdoctoral associate in the honeybee laboratory. Karen performed the experiments and the samples were sequenced twice: first on the 454 pyrosequencing platform, and then again on the Illumina MiSeq platform, when we decided (after studying the results of the downstream analyses) that the 454 reads were unusable due to the quality of the reads. I analyzed the original 454 reads with advice from Maksim Sipos (then a senior Ph.D. student in Nigel's group), and together we also developed a video tutorial for the TORNADO pipeline that had been developed for analyzing 454 reads. I analyzed the MiSeq reads with advice from Patricio Jeraldo (then a postdoctoral associate at IGB and Mayo Clinic); the IM-TORNADO pipeline for paired-end MiSeq reads was still under development at the time (but has since been published), so I ran each step of the development version pipeline independently, verifying the quality of the outputs at each stage. I also performed the downstream ecological clustering and ordination analyses under Karen's guidance.

The RFID tracking project described in Part II was performed in close collaboration with Claudia Lutz (then a neuroscience Ph.D. student) and Paul Tenczar (a citizen scientist at IGB). Paul and Claudia performed all the experiments. I designed and implemented all the analyses described in Part II in close consultation with Claudia and Paul.

The bCode tracking project described in Part III was performed in close collaboration with Tim Gernat, a research scholar in the honeybee laboratory and a Ph.D. candidate at the University of Leipzig. Tim designed and implemented the tracking system, performed all experiments and implemented the computer-vision algorithms for trophallaxis detection (subsection 7.3.1). I designed and implemented the methods for trip detection, forager identification and diurnal rhythms described in chapter 7, and all the work in chapter 8. Weibing Deng, then a postdoctoral associate in the Department of Mechanical Science and Engineering, implemented the dance detection method discussed in subsection 7.3.2, which I used as one of the inputs in my forager detection method (section 7.5). Tim and I jointly designed and performed the analyses in chapter 9; Tim created the randomized reference networks and I performed all the spreading simulations on the IGB Biocluster for high-performance computing.

All the figures and plots in this dissertation were generated by me, unless noted otherwise in the caption.

1.4 List of publications

The studies described in chapter 4 and chapter 6 have each already been published, and the work described in chapter 9 has been submitted for publication:

- Paul Tenczar*, Claudia C. Lutz*, Vikyath D. Rao, Nigel Goldenfeld and Gene E. Robinson (2014). "Automated monitoring reveals extreme interindividual variation and plasticity in honeybee foraging activity levels." *Animal Behaviour* 95:41–48.

- Karen M. Kapheim*, Vikyath D. Rao*, Carl J. Yeoman, Brenda A. Wilson, Bryan A. White, Nigel Goldenfeld and Gene E. Robinson (2015). "Caste-specific differences in hindgut microbial communities of honey bees *(Apis Mellifera)*." *PLoS One* 10(4): e0123911.

- Tim Gernat, Vikyath D. Rao, Martin Middendorf, Harry Dankowicz, Nigel Goldenfeld, Gene E. Robinson (2016). "Rapid spreading dynamics despite bursty interaction patterns in honeybee social networks." *(Submitted.)*

Further publications are planned for the work described in chapter 7 and chapter 8.

Chapter 2

A primer on honeybee biology

The goal of this short chapter is to collect together those biological facts on honeybees that are the bare essentials of background for understanding the projects discussed in this dissertation. The material in this chapter is mostly sourced from two books: Winston's *Biology of the honey bee* [133] and Seeley's *Wisdom of the hive* [105]; the latter is particularly informative on the topic of honeybee foraging. Because this chapter only provides a broad overview of honeybee biology, I will generally not include extensive references to the primary literature here; instead, those are provided as topics arise in the remainder of this dissertation.

2.1 Eusociality

Honeybees (*Apis mellifera*) live in societies consisting of thousands of individuals, and display a remarkable combination of variety in individual traits and social cooperation within the society. Honeybees are considered *eusocial* insects. Biologists have tried to isolate a small number of features that are common across many species, which define eusociality. There are arguments about how precisely to define eusociality [28], but for our purposes, the definition based on the following three features (first proposed by E. O. Wilson) is sufficient:

1. Individuals of the same species cooperate in caring for the young.
2. There is reproductive division of labour: individuals that are essentially sterile work alongside fertile individuals.
3. Within the society, there is an overlap of (at least two) generations, such that offspring work alongside the parents during some period of their lives.

The study of the evolution of eusociality is a fascinating subfield in itself, but is beyond the scope of this dissertation. We only mention here that in 2010, Nowak et al. [87] tried to model the emergence of eusociality using a population genetics model, and without using the prevailing framework of kin selection

theory. This model was very controversial and drew many dissenting responses, but it provides an interesting point of entry into the topic of the evolution of eusociality.

In honeybees, reproductive division of labour consists of essentially-sterile female worker bees living alongside female queens and male drones. Honeybees exhibit the sex-determining system of *haplodiploidy*, in which males develop from unfertilized eggs while females develop from fertilized eggs; in other words, male bees only have a mother, while females have both a mother and father. The queen bee of a colony lays unfertilized eggs that become the drones of the colony, who go on to mate with virgin queens from other colonies; she also lays fertilized eggs that become the workers or virgin queens of the colony. Female workers do possess the ovaries that can produce viable, unfertilized eggs, but fewer than 1% have their ovaries sufficiently developed to do so. An even smaller percentage of males in colonies with queens come from worker eggs, as there is "worker policing," in which other workers destroy eggs laid by their worker colleagues and show aggression towards workers attempting to lay eggs.

2.2 The honeybee colony as an organizational unit

Honeybee colonies reproduce by the remarkable process of *swarming*. Prior to swarming, the queen lays eggs into specialized cells, and these are reared as virgin queens by the workers. The queen then leaves the original hive to seek a new site, together with about 60% of the workers, who also carry a large load of honey for use during transit and during the first few days in the new hive. In the hive left behind, workers continue to rear virgin queen larvae and worker brood. About a week following swarming, new virgin queens begin to emerge and go off on mating flights. Once a virgin queen has mated, the colony may fission again in an afterswarm, in which the mated queen departs with some workers, in the same fashion as a regular swarm. After one or two afterswarms, there are too few workers left for the process of fissioning to continue, and so the virgin queens that are still in the original hive fight to the death until a single one reigns over the colony. Occasionally a colony will also reproduce by the process of *supersedure* rather swarming: here, the queen is simply eliminated and replaced by a younger queen.

The other reproductive members of the colony are the drones, who are responsible for mating with virgin queens from other colonies. In fact, most drones die before mating, either due to age or because they are ejected from the hive by the workers. In the spring and summer months, drones take mating flights to try to mate with virgin queens; if they succeed, they die promptly after mating. By the time fall arrives, any drones still remaining in the colony are ejected. A young queen will mate with ten or more drones, and so her daughters in the colony trace multiple patrilines.

Swarming occurs primarily during springtime. During the winter, when temperatures drop, bees stay inside the hive and form a *winter cluster* to maintain the nest temperature and protect the queen. Generally, clusters form when temperatures drop below 18°C, and the porosity of the cluster changes to accomplish thermoregulation [88]. In the winter, bees consume honey, which is made during the rest of the year by processing nectar from flowers. During the rest of the year, bees forage for both nectar and pollen, which are the two basic foodstuffs that they consume.

2.3 Inside the beehive

What would we see if we could transport ourselves into a honeybee hive in the wild? The size of the colony varies depending of the time of year and its lifecycle. In late spring, before swarming, a full-size colony may have around 30,000 individuals. The vast majority of these individuals are (female) worker bees; there is a single queen, and during spring, on the order of 1,000 drones. These different castes have different lifespans: queens live for 1–3 years, drones for about 4 weeks, and the lifespan of workers varies depending on the season and subspecies of honeybee; in the summer months, workers live on the order of five weeks.

Workers bees perform different tasks, depending on their age, a phenomenon known as *temporal polyethism*. Younger worker bees perform in-hive tasks, such as cleaning cells, rearing brood, building comb, tending to the queen, and offloading and handling food from foragers. There is a progression in which in-hive tasks a worker bee will perform with age, and there are physiological changes accompanying ageing that are related to task performance.

As a worker bee gets older, she moves on to performing tasks close to the hive entrance, such as ventilating and guarding the hive. Eventually, a worker bee will take her first flights, known as *orientation flights*, and typically on the next day, the bee will begin foraging, which is the task associated with the oldest age. Mortality is much higher for bees performing forager than those remaining in the hive, so it appears to make sense that the oldest bees are assigned this most risky of tasks. Although task performance is generally determined by age, it can also be modulated by the needs of the colony, as we will see later in this dissertation. Thus individuals are able to perceive the collective need of the colony, and respond accordingly. The transition from hive bee to forager in particular has been modelled using the idea of social inhibition [11].

Bees communicate inside the hive using a variety of methods. Chemicals feature prominently: pheromones diffuse through the hive and bees secrete chemicals on their bodies that can be sensed by other bees using their antennae. The famous dance language is one method of movement-based communication, and is discussed in more detail in section 5.1. Another important movement-based signal is DVAV (dorsoventral

abdominal vibration), also used in foraging communication. Antennation is used to sense movement and chemicals. Finally, the act of liquid-food–exchange is believed to also be a form of communication, and one that will be particularly important for us. We discuss the mechanics of trophallaxis in more detail in subsection 7.3.1.

Part I

The honeybee gut microbiome and social organization

Chapter 3

Introduction to metagenomics

Most of life on Earth is microbial: it is estimated that there are a staggering 10^{29} single-celled organisms living on the planet [66]. We humans go about our daily lives mostly unaware of these tiny organisms, even though they make up roughly 30% of the total biomass on Earth. For much of Earth's history, microbes have played an important part in shaping the planet itself: for example, the great oxygenation event that occurred 2.3 billion years ago was driven by oceanic cyanobacteria [45]. And microbes continue to play an important role today in regulating the Earth's climate via the global carbon cycle [9].

Microbes were first discovered under the lenses of Leeuwenhoek around 350 years ago, and ever since, we have been trying to understand the relationship between microbes and other living organisms. Through the work of microbiologists such as Louis Pasteur and Robert Koch in the nineteenth century, it came to be understood that microbes can cause diseases in humans and other animals, and also that microbes are responsible for processes such as fermentation and food spoilage. Microbiologists soon began to culture different microbes in the lab to study their biochemical interactions, leading to landmark advances in immunology.

The discovery of the molecular structure of DNA in the middle of the last century opened a window into the genetic basis of life [130]. Biologists were now able to study the relationship between genes—the molecular blueprint for life—and biological function. Genes also provided a molecular basis for reconstructing the history of life: Carl Woese used one particular gene, coding for the 16S ribosomal RNA subunit of ribosomes, to construct a phylogenetic tree of life, and in this process, discovered the Archaea, a set of microbial organisms that constitute one of the three main branches—or domains—of the tree of life [134].

The emergence of new technologies in the second half of the twentieth century led to the growth of new, quantitative subfields of biology, such as genomics and molecular biophysics. Today, research in quantitative microbiology can be roughly classified by scale [110]:

- single-molecule experiments and simulations, studying the physical properties of biological molecules *in vivo, in vitro,* or *in silico*;

- single-cell experiments in culture, such as measurements of gene expression, transcriptome analysis, and cell-level processes such as division and motility;
- experiments using bulk cultures of cells, such as measurements of bacterial colony growth, biofilm formation, quorum sensing;
- experiments using uncultured cells in bulk, such as metagenomics and metatranscriptomics.

This last method of studying microbes represents a particularly significant conceptual step forward, as it links the microbial world with the field of ecology. It opens the possibility of studying *communities* of microbes, in which the individual organisms are interacting in various ways, determined both by their genetics and their environment. It is believed that the vast majority of bacteria cannot be cultured in the lab [115]—it is difficult to even give a precise estimate of the extent of this so-called Great Plate Count Anomaly because the very notion of species of bacteria is ill-defined. Any definition based on lab cultures is useless in describing unculturable bacteria, while genetic definitions are confounded by the propensity of bacteria to exchange genes horizontally [34].

The study of bulk communities of bacteria was enabled by advances in sequencing technology, and the term *metagenomics* was coined in 1998 for the study of genetic material recovered in bulk from an environmental sample [54]. In section 3.1 and section 3.2 of this chapter, I will describe the sequencing technologies used in metagenomics, and how the methods of metagenomics allow us to probe communities of bacteria.

The field of metagenomics also opens the possibility of studying the interaction between communities of microbes and higher organisms. The ecological dynamics of a microbial community are dependent on its environment, and if that happens to be located within another organism—say the gut of a mammal—then there will inevitably be feedback (in both directions) between the microbial community and its host organism. Thus, certain phenotypes of a host organism may be influenced by a particular microbial community that it hosts, and conversely, the microbial community may be structured by the host environment.

In the context of humans, this discovery exposes a completely new perspective on health and disease, as these can now be seen to be influenced not just by the genetic makeup of an individual but also by the set of microbial organisms that she hosts. It is estimated that there are 10 times as many microbial cells residing in a human body than human cells, and 100 times as many genes in the human microbiome compared to the human genome [6]. Such comparisons even raise the question of what it means to be human—clearly, an answer that relies on genetics alone is too simplistic.

These comparisons also shed new light on the question of nature versus nurture: while we understand

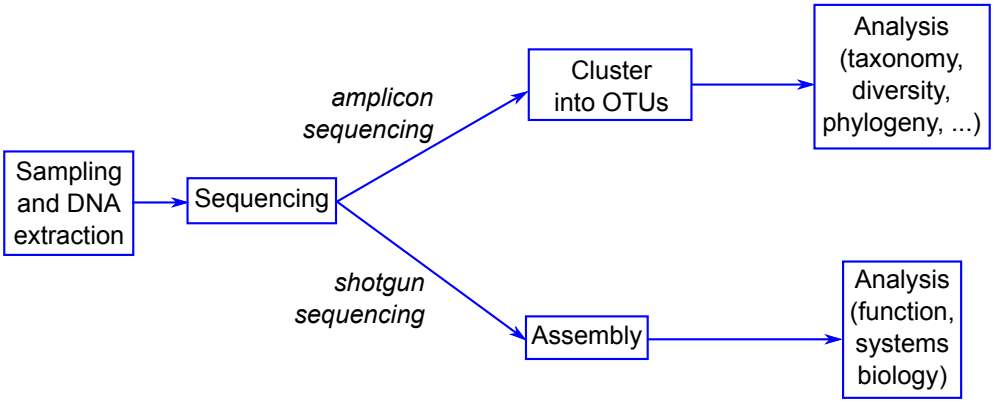

Figure 3.1: **Metagenomics flow diagram.** Overview of the different steps in analyzing metagenomics data. Based on the flowcharts in Refs. [110, 119]

well the genetic basis of heredity, human offspring also inherit some of their mother's microbes while passing through the birth canal [32, 33]. However, the microbiome is not static in time, and during a human life cycle, the structure of the microbiome changes [19, 138]. The composition of the microbiome is known to be influenced by diet, cohabiting individuals and pets [112], geography, and other features that continue to be discovered.

While our understanding of the human microbiome and its dynamics continues to improve, it is sufficient to note here that the microbiome—which can affect human health—is not simply inherited in a linear manner from parent to offspring, but can be influenced by the social context in which an individual finds herself, since this influences what she eats, whether she has a pet, etc. In chapter 4, I will describe how we studied the connection between the microbiome and social organization in the model system of a honeybee colony, where the societal roles of individuals can be more simply defined than in a human society. The remainder of this chapter will provide background to the methods used in that study.

3.1 Sampling a microbial environment and sequencing

Figure 3.1 gives a broad overview of a typical metagenomics project. Once DNA samples are extracted from the environment of interest, there are two different approaches that are commonly used. In *amplicon sequencing*, only one or more specific marker genes are chosen to be amplified in the DNA extracted, and these are then sequenced. To study bacteria in the sample, the 16S rRNA gene is usually used, as it contains highly variable regions that can be used to distinguish different kinds of bacteria on a phylogenetic basis [134]. In *shotgun sequencing*, the sample DNA is randomly split into small fragments, which are then amplified and sequenced. The sequenced fragments can be computationally matched to complete reference genomes or to

specific reference marker genes to create a so-called co-assembled library. Alternatively, it is also possible to create a *de novo* assembly of the genomes present in the sample, although this is computationally more challenging, and has only been accomplished in low-diversity environments [119, 63].

Amplicon sequencing aims to provide a community profile for the sample of interest. As described in subsection 3.2.1, the sequenced reads are clustered into *operational taxonomic units* (OTUs) based on sequence similarity, and these serve as a microbial analogue of the species concept in macroscopic ecology. Based on the community profile, one can address questions of microbial ecology, such as diversity, community structure, niche formation, etc. On the other hand, shotgun sequencing provides information about the entire ecosystem of genes, and in principle provides information about the metabolic functioning of the community.

Metagenomics became possible with the advent of rapid and low-cost sequencing. In the following two sections, we briefly review Sanger sequencing, which was developed almost 40 years ago and became the first widely-used sequencing method, as well as the set of so-called next-generation sequencing technologies, which allow for vast quantities of DNA to be sequenced.

3.1.1 Sanger sequencing

In 1977, Frederick Sanger and colleagues developed a method to sequence single-stranded DNA that was later commercialized by Applied Biosystems and came to be widely used for sequencing DNA. The method is based on incorporating chain-terminating dideoxynucleotides (for each of the four bases in turn) in place of the standard deoxynucleotides during synthesis from the template strand. So if, for example, the dideoxynucleotide ddATP is included together with the standard dATP, then the synthesis reaction results in fragments that all terminate at base A. The lengths of these fragments can be found using gel electrophoresis, and from this, all locations of the base A in the template sequence may be deduced. This synthesis process is repeated for the other three bases, allowing us to reconstruct the complete template strand.

Although Sanger's original protocols have been modified in more recent times—incorporating, for example, the use of microfluidic devices [68]—the basic principle is the same. For the sequencing of long strands of DNA (such as the complete genome of an organism), the strand is first fragmented into smaller pieces that are then amplified by polymerase chain reaction (PCR). Sanger sequencing is highly accurate, and the lengths of the "reads" that are reconstructed from these fragments can be up to 800–1000 bases long. The complete genomic DNA can then be computationally assembled from these shorter reads.

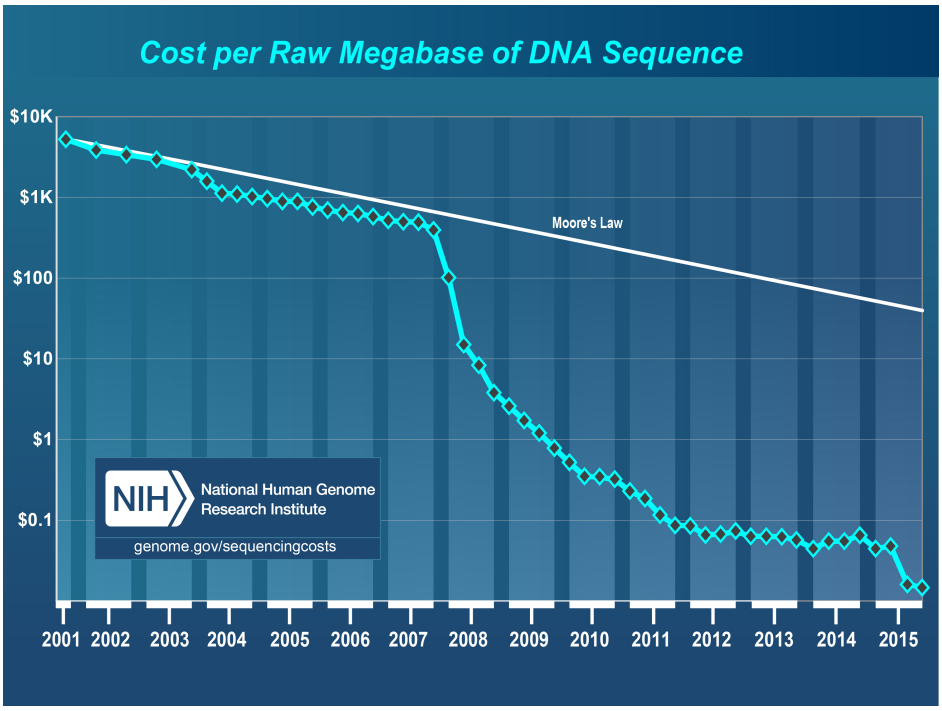

Figure 3.2: **Decline in sequencing cost over time.** The National Human Genome Research Institute (NHGRI) periodically publishes estimates of the cost per-base of DNA sequencing at `https://www.genome.gov/sequencingcosts`.

3.1.2 Next-generation sequencing methods

A landmark achievement in DNA sequencing was the completion of the Human Genome Project in 2003. Since that time, the cost of sequencing DNA has been falling at a faster-than-exponential rate (see Figure 3.2), thanks to the emergence of so-called next-generation sequencing (NGS) technologies. For a complete review of these technologies, the reader is referred to the recent review article by Goodwin et al. [51]. Here, we review two NGS technologies—454 pyrosequencing and Illumina MiSeq—that we used in studying the honeybee gut microbiome (see chapter 4), to give a flavour of how these modern methods work, and to compare them with first-generation methods based on Sanger sequencing. Both these methods are based on *sequencing by synthesis*: a polymerase is used to synthesize the complementary sequence of a template fragment, and the nucleotides that are used in this synthesis process are read out in some way.

In 454 pyrosequencing, the read out takes place using a light signal as each nucleotide is incorporated into the elongating strand that is being synthesized. A chemoluminescent enzymatic reaction is used to detect the release of a pyrophosphate upon the addition of a single nucleotide. During each cycle, solutions containing each of the four fluorescently labelled bases A, T, G, and C are sequentially added and removed from the reaction, and after each addition, the reaction surface is imaged with a CCD camera to see if the

base has been incorporated into the elongating strand; once the nucleotide has been incorporated, the next cycle begins. In this way, each base of the template strand is sequentially read out, one at a time. The typical read lengths are 400–600 bases for 454 platforms. Insertion-deletion errors are the most common type of error in 454 platforms, but the overall error rate is comparable to other NGS technologies [51]. The method is also especially susceptible to errors in homopolymer regions (i.e. a subsequence of identical bases, such as AAAA), as the intensity of light has to be used as a proxy for the length of such regions. Due to these issues, and the relatively high cost per base compared to other NGS technologies, 454 platforms were discontinued in 2013 by parent company Roche.

The suite of platforms developed by Illumina are based on adding a distinct fluorescent dye to each of the four deoxynucleotides, and then detecting this as the elongating strand is synthesized, one nucleotide at a time. During each cycle, a single nucleotide is incorporated into the elongating strand. The reaction surface is imaged to see which of the four bases was added. The fluorescent dye and chain-terminating group that were originally on the newly-added deoxynucleotide are washed away, and the next cycle can begin. The Illumina MiSeq platform has a read length of 300 bases for single strands. The platform also allows for paired-end sequencing, in which the template strand is sequencing from each end independently; this allows for a maximum read length of 2×300 bases. This technology is now highly mature, and as a result, Illumina presently dominates the short-read sequencing industry [51].

While NGS sequencing methods tend to be much cheaper and have a higher throughput than Sanger-based technologies, their primary limitation is the relatively short read length (typically about 500 bases). This means that computational tools that can leverage short reads are particularly important.

3.2 Bioinformatics pipeline

In this section, we discuss in more detail amplicon sequencing of the 16S rRNA gene. This gene is found in the DNA of all bacteria and archaea, and codes for the small subunit of the ribosome. Figure 3.3 shows the secondary structure of the 16S subunit in *Escherichia coli*. The gene consists of some regions that are highly conserved over evolutionary timescales, particularly those regions that are bonded among themselves. However, as Figure 3.3 shows, there are also unbonded loops, and other sections of the gene, that are more susceptible to mutation. These are divided into variable and hypervariable regions; there are 9 so-called hypervariable regions, labelled V1–V9, that can be used to recover phylogenetic information about the gene (and thus of the organism). Since the read length in NGS technologies is smaller than the length of the complete 16S gene (1542 bases for *E. coli*), a fragment of the gene containing some subset of these

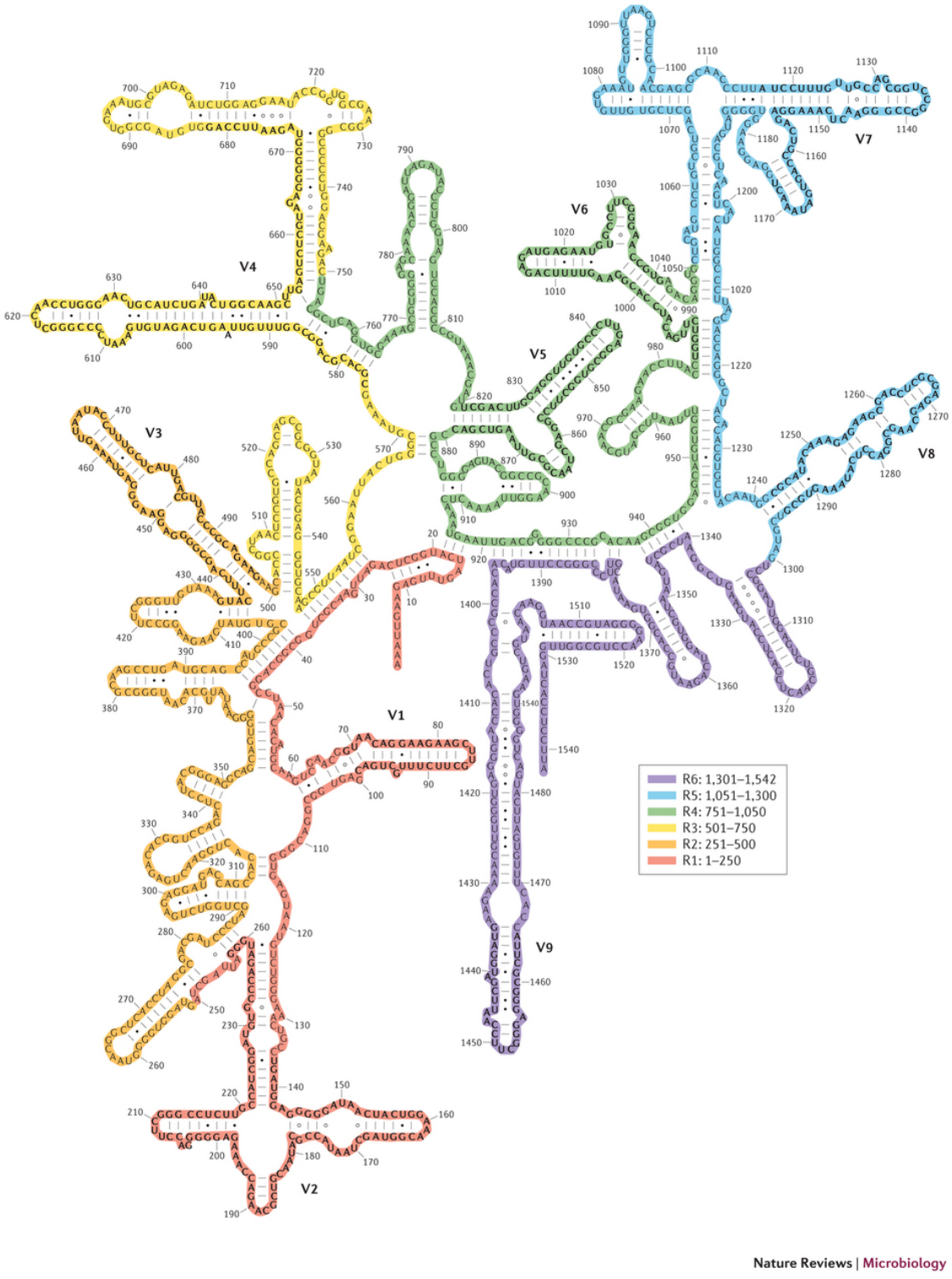

Figure 3.3: **Secondary structure of the 16S rRNA gene of *E. coli*.** In this figure, the gene is divided into 6 colour-coded regions of about 250 bp each. The hypervariable regions are labelled as V1–V8. Reproduced with permission from Ref. [137].

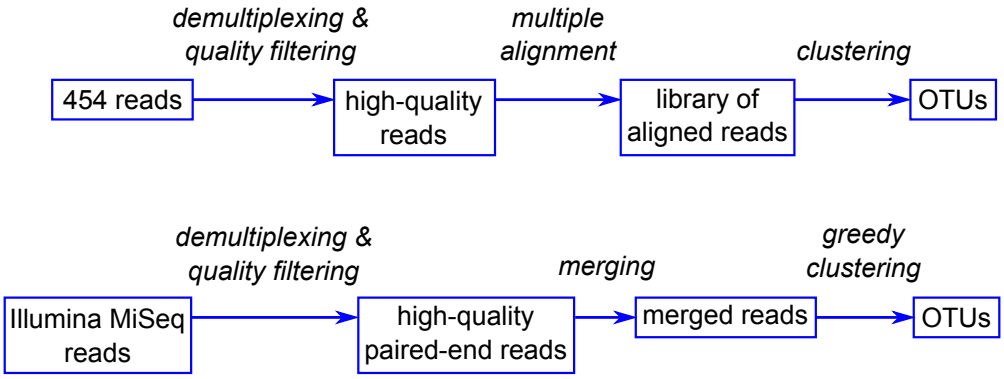

Figure 3.4: **Bioinformatics pipeline.** The main difference between the pipelines for the two technologies is that multiple alignment is used in processing 454 sequences, but is not required for MiSeq reads.

hypervariable regions is sequenced for metagenomic community reconstruction. It has been shown that the V1–V3 region is particularly suitable for this purpose [62], and this is the region we used in our study of the honeybee gut microbiome, described in chapter 4.

The protocols used to reconstruct a microbial community from a set of reads vary depending on the sequencing technology used to obtain the reads. In the following two sections, we describe the procedure for 454 pyrosequencing and for Illumina MiSeq. At the time of writing, the former has essentially become a legacy technology, but the issues encountered provide insight how these protocols are devised. The protocols described were developed in IGB (and subsequently the Mayo Clinic) and came to be called TORNADO (Taxon Organization from RNA Dataset Operations) [111, 64].

3.2.1 454 pyrosequencing

Figure 3.4 outlines the general procedure we used for processing 454 reads. The reads are typically in the form of a plain text FASTA file, which gives both the read ID and the sequence of bases for that read. In addition, each read may have additional quality data, such as a Phred quality score for each base; in this case, the FASTQ file format may be used. Each read typically begins with an primer sequence (which is used to amplify the sequence during PCR) followed by a barcode (typically made up of 10 bases) marking the sample that the read comes from, and finally the sequence from the sample. In the first, quality-filtering step, the primer is trimmed and the reads are demultiplexed (meaning separated by barcode), and then the barcodes are removed). Additional steps may be taken at this stage, such as only retaining reads above a certain length (typically 100 bases) and trimming bases that could not be distinguished (these appear as the letter N in the FASTA file, rather than one of the four bases). The sequences are also *dereplicated*: that is, identical reads that appear more than once are consolidated into a single read, and information about the

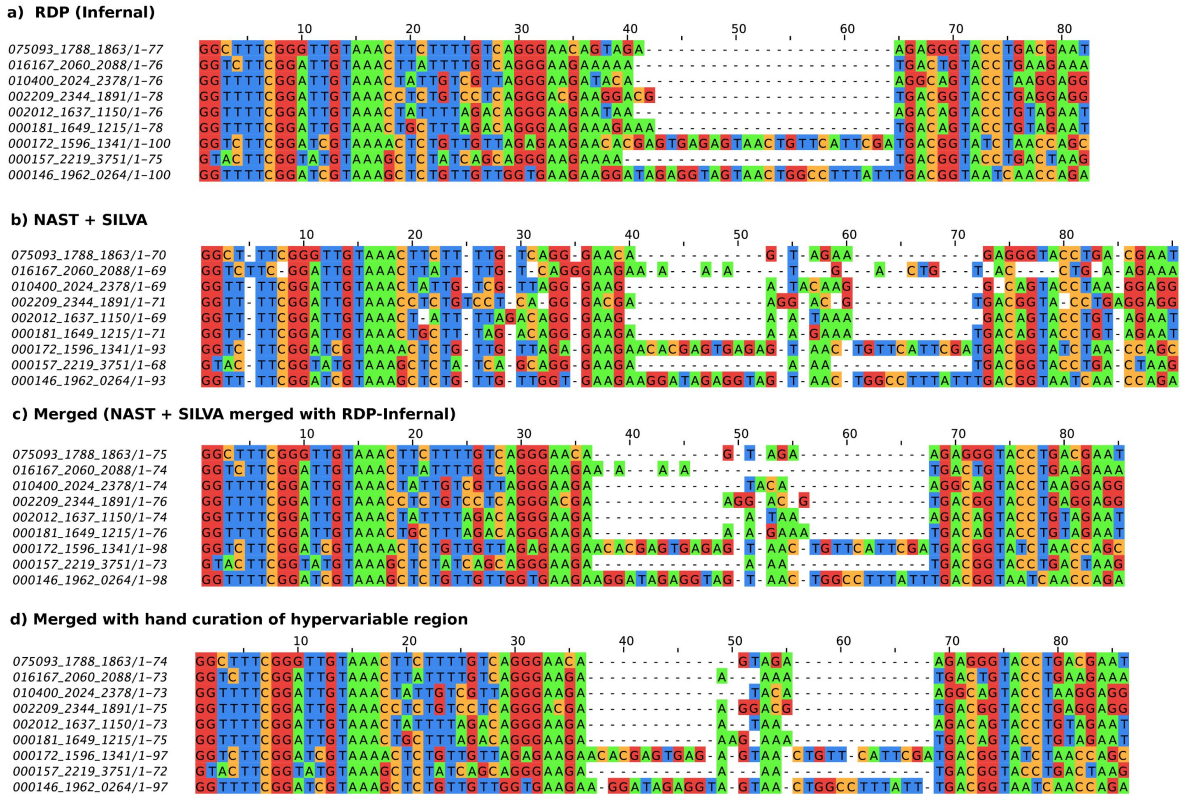

Figure 3.5: **Multiple alignment using different methods.** a) Infernal aligns based on secondary structure, and if some samples contain hypervariable regions that do not fit into the structural template, these are left unaligned. b) NAST with alignment to the SILVA database results in some alignment errors in conserved regions when there is no perfect match to the template library. c) The merged alignment uses Infernal to align conserved regions and NAST+SILVA to align hypervariable regions. d) Further hand curation of the hypervariable regions may be necessary. Figure reproduced from Ref. [111]

number of times this read appears is noted as metadata (typically appended to the read ID). 454 sequencing is also susceptible to chimera errors, in which a single read represents two sequences fused together; the UCHIME algorithm is used to detect and remove chimeric reads [40].

After quality filtering is complete, the complete library of reads is aligned. The goal of alignment is to insert gaps so as to align sequences according to their phylogenetic relationship, such that mismatches and indels (insertions and deletions) at particular positions represent a reconstruction of the evolutionary relationship between a pair of reads. This is an important step, because the ultimate goal is to cluster the reads by similarity, and alignment allows reads to be compared base-by-base.

There are several algorithms commonly used to perform multiple alignment. The TORNADO pipeline for 454 sequences uses a combination of two alignment methods. The NAST algorithm [30] implemented in the mothur software package [103] aligns sequences by comparing them to a handcurated library of reads

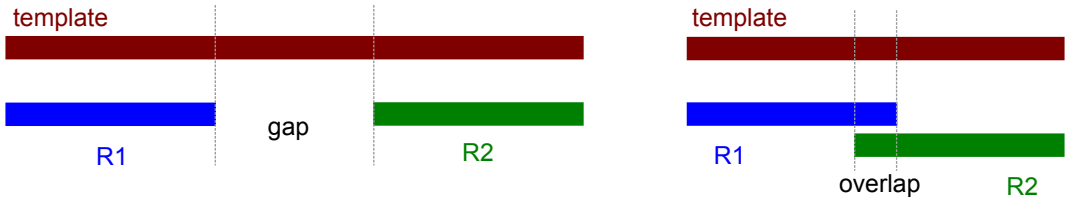

Figure 3.6: **Paired-end reads.** The template strand is read from both ends, and depending on its length, there could either be a gap or an overlap between the pair of reads R1 and R2.

(the SILVA database is used in mothur). Another aligner is Infernal, part of the Ribosomal Data Project (RDP) [22], which also uses secondary structure information to align 16S sequences. Figure 3.5 shows a small set of reads aligned with each of these algorithms independently. The NAST alignment tends to have alignments that in places do not respect secondary structure, while Infernal incorporates secondary structure but is susceptible to errors when the occasional taxon appears with high variation in between the conserved regions. Thus it is best to use both methods, with additional hand curation of misaligned hypervariable regions as necessary.

Once the set of reads has been aligned with respect to each other, they are clustered by similarity. The dissimilarity between a pair of reads is typically measured by a Hamming distance, which is the fraction of bases that are mismatched between the pair. This can be done using complete-linkage clustering (implemented in mothur), which is a hierarchical clustering method based on a distance matrix created by calculating pairwise Hamming distances. (We discuss the details of clustering in more detail in subsection 3.3.2; although the context discussed there is different, the idea behind the method is exactly the same.) By convention, reads that differ by less than 3% are clustered together, and each cluster of reads is called an operational taxonomic unit (OTU).

The end result is that for each sample, we have a list of OTUs present in that sample, together with the *abundance* of that OTU in the sample, which is simply the number of reads that cluster into the OTU. Note that the abundance does not provide information about the *absolute* number of microbes of a given OTU present in a given sample: rather, the abundance is simply proportional to be number of microbes, and in comparing across different OTUs, we can compare relative (but not absolute) abundances.

3.2.2 Paired-end Illumina MiSeq reads

As Figure 3.4 shows, the procedure for reads from the Illumina MiSeq platform is generally similar to that for 454 pyrosequencing. There are two main difference:

- The MiSeq platform produces paired-end reads, that is, the template strand is read from both ends,

and so each strand produces a pair of read R1 and R2, as shown in Figure 3.6.

- Because the reads from MiSeq are of much higher quality than the 454 platform, there is no need to perform the computationally expensive operation of creating a multiple alignment. Instead, the reads can be clustered heuristically into OTUs.

The first issue is dealt with differently depending on where there is a gap between the pairs or a region of overlap (see Figure 3.6). If there is a gap, the pair of reads R1 and R2 may be simply concatenated. If there is a region of overlap, they have to merged to form a consensus read; this may be done using USEARCH [42]. The second, heuristic clustering step may also be performed using USEARCH.

3.3 Analyzing microbial compositions

The end result of the bioinformatics pipeline is essentially a table showing the abundance (i.e. number of reads) for each OTU in each sample. This is similar to traditional ecological surveys, in which different samples may correspond to different habitats, niches, spatial locations, etc. and the abundance is simply the number of individuals of a particular species observed in that sample. The main difference is that microbiome data are *compositional*, i.e. the absolute number of reads depends on the specific experimental protocols used, so we have to normalize the number of reads by the sample total. Another typical feature of microbiome data is that they are high-dimensional and sparse: the OTU vector of relative abundances for a sample has a lot of entries but many of them are zero [73]. These two features can in some cases introduce statistical complications, such as in accurately measuring correlations between samples [49].

Nevertheless, many of the standard statistical measures used in ecology can also be used in microbial ecology to study the relationship between different bacterial OTUs and their environment. In this section, we briefly review some ecological concepts used in our study of the honeybee gut microbiome.

3.3.1 Measuring diversity

Ecologists traditionally think of the total diversity in a landscape as being composed of two parts: the mean diversity at the local scale of habitats (known as α-diversity) and differentiation among those different habitats (β-diversity) [132]. In the context of microbial communities, α-diversity is equivalent to within-sample diversity, and one measure of this is the *Simpson index*. If a particular sample contains OTU i at relative abundance p_i, the Simpson index for that sample is simply

$$\text{Simpson index} \qquad D \equiv \sum_i p_i^2. \tag{3.1}$$

The mean of this value across samples is one measure of the α-diversity of the entire sample set. Another index that is used is Shannon's diversity index, which is just the information-theoretic entropy of the OTU vector:

$$\text{Shannon index} \qquad H \equiv -\sum_i p_i \ln p_i. \tag{3.2}$$

This has the advantage that it is not dominated by the most abundant OTU, but instead provides an index that is sensitive to both the typical abundance and the evenness of the within-sample distribution. One way to understand this difference between these indices is to think of both as weighted means over the OTU vector $\{p_i\}$, with D weighted by the relative abundance and H weighted by the negative logarithm of the relative abundance.

We characterize diversity in the bee microbiome using Shannon's index, as described in subsection 4.4.2. We also note here that for microbial communities, it is possible to create more sophisticated measures of diversity using phylogenetic information about the OTUs. For example, each OTU can be placed on a phylogenetic tree (derived from comparing the OTU consensus sequences), and diversity can be measured by adding the lengths of branches on such a tree [73].

As for β-diversity, one way of quantifying the distance between two microbiome samples A and B (with abundances n_{Ai} and n_{Bi} for OTU i) is using the Bray-Curtis dissimilarity,

$$d(A,B) = \frac{\sum_i |n_{Ai} - n_{Bi}|}{n_{A+} + n_{B+}} \qquad \text{(Bray-Curtis)}, \tag{3.3}$$

where n_{A+} (n_{B+}) is the total number of reads in sample A (B).

This is a standard measure used in ecology, but again, such measures can be modified to incorporate phylogenetic information. UniFrac is one such distance measure, devised in 2005 [74], that has come to be widely used in metagenomics.

3.3.2 Clustering and ordination

Two important and related methods used in ecology are clustering and ordination. The goal of clustering in this context is to explore possible similarities between, say, two habitats, or microbial samples; more generally, in the fields of statistics and machine learning, clustering is an important method in *unsupervised learning,* which addresses the task of discovering hidden structure in unlabelled data. Ordination is a related method of exploring data by finding an embedding of the points of interest—habitats—in some mathematical space, such that objects that are similar (in some mathematical sense) are close together, while those that are very different are far apart. This is typically of interest when the space is low-dimensional: thus, the

task often ends up becoming a problem of *dimension reduction*—taking the description of data in some high-dimensional space and mapping that description to a low-dimensional space (ideally two or three dimensions, so that one can directly visualize the data).

In this section we review some methods for clustering and ordination commonly used in ecology. We will provide an explanation for how these methods work to understand how they well they are suited for different analyses, but will not write down the detailed mathematical formulation of each method, as these are well known and can be easily looked up. For concreteness, let us assume in this section that we are dealing with OTU relative abundance data, represented by a matrix $\boldsymbol{X}$ such that the element $X_{\alpha i}$ represents the relative abundance of OTU i in sample α, for $i = 1, \ldots, p$ and $\alpha = 1, \ldots, N$; we will also represent the αth row of $\boldsymbol{X}$ as the OTU vector $\vec{x}_\alpha$. In the more general terminology of statistical learning, $\vec{x}_\alpha$ would be called the *feature vector* for the αth object or data point.

When there is some notion of distance or dissimilarity $d_{\alpha\beta}$ between a pair of samples α and β, a common method of clustering that is used is *hierarchical clustering*. The goal is to create clusters of points such that within each cluster, points are separated by "small" distances, in some sense, while a pair of points that come from two different clusters should be separated by a "large" distance. The quantitative outcome is determined by the specific method of hierarchical clustering used: for example, in complete-linkage clustering, two points within a cluster must, by construction, be separated by no more than a maximum threshold distance.

Qualitatively, the algorithmic task of hierarchical clustering may either be performed agglomerative, meaning that each point is initially its own cluster and clusters agglomerate as a distance threshold is increased; or it could be performed divisively, in which case we begin with all points in a single cluster, and break up that cluster as the distance threshold is decreased. In both cases, one can build up a dendrogram showing the clusters present as the distance threshold is varied. Note that although these clustering algorithms will always produce clusters, the clusters may not always be meaningful, i.e., it may not be possible to achieve the goal stated above simply because the points do not truly form clusters in the p-dimensional feature space. In metagenomics, hierarchical clustering is often used in two distinct ways: first, in the actual construction of OTUs, where the notion of dissimilarity is the Hamming distance between OTU consensus sequences, as described in section 3.2; and second, in trying to cluster samples by the similarity of their microbiomes. These two uses have different goals: the first gives an operational analogue of the species concept in bacteria; the second tries to address the ecological question of how microbial community differentiates depending on its habitat.

Turning to ordination, perhaps the simplest and most common method used is *Principal Components Analysis (PCA)*. PCA does not rely on a pairwise distance between samples, but instead uses correlation

between pairs of samples. Typically in PCA, one first calculates the $N \times N$ correlation (or covariance) matrix $\mathrm{Cov}(\vec{x}_\alpha, \vec{x}_\beta)$, and then performs eigenvalue decomposition of this matrix to find the axes along which highest variation occurs. Thus the nth principal axis is the eigendirection corresponding to the nth largest eigenvalue, and looking simply at the first few principal axes should reveal patterns in correlation. PCA has the advantage that there is a natural interpretation of the axes, and the degree of variation along an axis is quantifiable simply by the relative magnitude of the corresponding eigenvalue. The disadvantage of PCA is that it is entirely a linear method, and it implicitly relies on a notion of distance that is just the Euclidean distance between feature vectors—and this may have no biological basis.

One method that overcomes this restriction on the notion of distance is *Principal Coordinates Analysis (PCoA),* also known as classical multidimensional scaling. In PCoA, one specifies the dimension of the embedding (typically, low-dimensional) space to begin with, and the algorithm tries to produce a representation of the points in this space, while minimizing the differences between the pairwise distances $\tilde{d}_{\alpha\beta}$ in the embedding space and the input dissimilarity $d_{\alpha\beta}$. Specifically, a quantity called the *strain* captures this difference, and the classical scaling algorithm moves points around in the embedding space in a way that minimizes the strain. A related algorithm is *non-metric multidimensional scaling (nMDS)*, which instead tries to preserve the ordering of the entries in the two matrices of distances $d_{\alpha\beta}$ and $d_{\alpha\beta}$. Both these methods have the advantage over PCA that they are non-linear; the tradeoff is that the axes of the embedding space generally have no simple interpretation.

In ecology, Bray-Curtis dissimilarity (Equation 3.3) is often used, and so hierarchical clustering, PCoA and nMDS are ideal methods for exploratory data analysis.

Chapter 4

Caste-specific differences in the honeybee gut microbiome

The tools of metagenomics developed over the past twenty years have enabled studying the microbial compositions of various environments, such as the soil, the ocean, the gastrointestinal tracts of various organisms, and more. Amplicon sequencing of the 16S rRNA gene is usually used to produce a static snapshot of the community composition. This composition does, of course, change in time, and it is becoming more common for studies to measure longitudinal compositional data [47]. However, a basic question that one can ask is: to what extent is the microbiome specific to that environment? Environmental conditions determine which microbes can survive and grow, and interactions among the microbes determine the relative abundances of different organisms. Microbes can in turn shape the environment that they inhabit. In some cases, such as the community in a methanogenic bioreacter, it appears that the community is in a state far from equilibrium, with the community composition constantly changing in time [48].

In the case of animals that form societies—whether humans or insects—one can study the relationship between the microbiome and the social role of the individual. To what extent does the host's social phenotype influence the microbiome composition? And can the microbiome in turn affect the host's social behaviour? While such questions may be most interesting in the context of human societies, particularly given their relevance to human health and disease, the complexity of human society and interactions means that we would require a vast amount of data to meaningfully address such questions. In this chapter, we study the relationship between the microbiome and host social phenotypes in the model system of a honeybee society, where the social roles are more simply defined, and where many of the relevant biological interactions among individuals have been well studied.

4.1 Introduction: host-symbiont dynamics

Microbes residing in the gastrointestinal tract of animals are considered to be symbionts, as they perform important physiological tasks, such as immune system development, digestion and detoxification, in the case of humans [37]. Variation in the gut microbiome across individuals is emerging as a source of phenotypic

diversity [82]. Honeybees and other social insect species are ideal systems in which one can explore the relationship between the microbial community composition and phenotypic variation, because individuals within the same society exhibit extreme differences in behaviour; in chapter 6, we quantify this for the foraging behaviour of honeybees. This variation in behaviour is accompanied by variations in diet, physiology and lifespan, which in turn can affect the composition of the microbial community.

In honeybee societies, there are clear physiological differences between the female queen, workers (also female) and male bees. Queens mate early in life and then remain in the hive, specializing in the production of eggs, which are then reared by the workers. Compared to the other female caste, queens are larger in size, have more highly developed reproductive organs, and typically live for 2–3 years—up to two orders of magnitude longer than female worker bees, which live for 5–7 weeks [133]. The typical worker bee's life cycle consists of specializing in tasks within the colony (such as brood care, cleaning, and comb building) early in life, and then transitioning to outdoor tasks such as foraging. Male bees do not work inside the hive early in their life, but instead spend their time resting and begging for food from nurse bees. Their behaviour and physiology is centred on mating, and they die soon after the mating flight. Male bees that survive the mating flight are ejected from the colony [133].

Diet is an important factor in determining phenotypic differences among different social groups. During the larval stage, future queen larvae receive more royal jelly from nurses than do future worker larvae, and this difference in diet has been linked to differences in caste-associated gene expression [67, 75]. As adults, queens are fed a more protein-rich diet than worker bees, while the diet of workers varies over their lifespan [25]. Nurses eat a pollen-rich diet, while foragers do not eat pollen [26]. However, all bees are fed by the nurses, and in this way, foragers do receive some protein and lipids. Meanwhile, adult male bees have low levels of pollen consumption and low intestinal proteolytic activity relative to workers [116].

Given these differences in diet and physiology among the different social groups, we expect that there should also be a link between the gut microbiome and social group. On the other hand, within any social animal species, close physical interactions create the potential for colony-wide transmission of gut microbiomes, and this mixing might lead all individuals within a colony to share the same microbiome. Indeed, in the case of humans, it has been shown that co-habiting individuals share microbiota among themselves and their pets [112]. Furthermore, among bees, similar gut microbiomes have been observed for workers from different colonies, continents, and races of honeybees [78, 83, 101, 76, 80].

In this chapter, we describe the first study in which the microbiomes of queen and male honeybees is explicitly characterized. Furthermore, although the gut microbiomes of 9- to 30-day-old workers has been shown to be highly similar [77], in our study, we examine explicity the variation in the microbiome of workers

bees engaged in nursing and foraging behaviour. We also compare microbiome composition among different social groups, within and across different honeybee colonies.

4.2 Methods

4.2.1 Sample collection

Queens, males, nurses and foragers were collected from honeybee colonies at the University of Illinois at Urbana-Champaign. No special permissions were required to collect from these colonies because they are owned by the university and *Apis mellifera* is not an endangered or protected species. A single queen, six males, six foragers, and six nurses were collected from each of three typical colonies headed by naturally mated queens in September 2011. We collected seven additional foragers from a pollen feeder in an indoor flight chamber that housed two colonies in February 2011. We pooled hindguts of four individuals from this collection into one sample and analyzed three individuals separately. We also included samples of seven-day-old virgin queens that were collected in summer 2003 after being reared in a "queen bank," namely, a colony that is used to store queens. These non-reproductive queens were held in cages within a colony so they could be fed by nurses, but were not yet functioning as queens. We pooled hindguts of four individuals from this collection into one sample and analyzed four individuals separately. Individual bees were collected live and stored at $-80°$C. The hindgut and its contents were removed from each bee using sterile dissection, and they were stored in PowerSoil-htp Bead Solution (Mo Bio, Carlsbad, CA) at $-20°$C.

4.2.2 DNA extraction and sequencing of hindgut contents

DNA was extracted from individual hindguts using a PowerSoil DNA isolation kit (Mo Bio, Carlsbad, CA), following the manufacturers instructions with an additional bead-beating step (120 s at 6.5 m/s). The hypervariable V1–V3 region of the 16S rRNA gene was amplified by PCR using tagged 27f and 534r primers designed with some degenerate nucleotides. This region of the 16S rRNA gene has been demonstrated to most accurately reflect the full length 16S rRNA gene [62], and is highly amenable to taxonomic reconstruction [123, 18]. Amplicons were sequenced using Illumina MiSeq technology, giving paired-end reads (2×300 bases in length). We performed a titration run with 65 samples pooled according to their DNA concentrations. We then renormalized the 60 samples that yielded more than 1,100 high-quality pairs of reads for a full bulk run. The five samples that yielded fewer than 1,100 paired-end reads were failed PCR amplifications (as indicated by a lack of a band on the gel). These samples yielded few (24–325 paired-end) reads, and therefore served as negative controls: samples that failed to produce a band on a gel also failed

to generate a robust microbiome profile. These five samples were excluded from further analysis.

4.2.3 Bioinformatics processing

The total number of sequence reads obtained was 23,229,704 (11,614,852 pairs). The reads were clustered into OTUs by implementing the steps used in the Tornado 2 metagenomics pipeline [62], described in section 3.2. Since the read-pairs were overlapping, they were first merged using USEARCH v7.0 [41]. Of the 11,614,852 raw read-pairs, 9,242,697 (79.6%) were successfully merged. The OTU centres (consensus sequences) were first identified by dereplicating the merged reads and clustering with USEARCH at the default radius of 3.0, corresponding to 97% sequence identity. This resulted in 112 OTU centers; these were then taxonomically classified using mothur [103] (with `classify.seqs`, using the Wang method and the Greengenes database). Those OTU centers that were unclassified at the species level were discarded, leaving a total of 83 OTUs. Finally, we used USEARCH (`usearch_global` command with the requirement of 97% sequence identity) to try to assign all the merged reads to one of the OTU centers. In this way, 6,252,345 merged reads were successfully assigned to one of the 83 OTUs.

4.2.4 Classification of OTUs into phylotypes

We classified each sequence at the family level, and used BLAST [2] to identify those OTUs that had been previously found in bees (blastn, e-value < 0.0001). We identified four OTUs as belonging to eukaryotes, and removed these from further analyses. OTUs classified as bacteria previously found in bees are denoted by the phylotype names first used by Babendreier et al. [5]. All OTUs belonging to each bee phylotype were from the same families: "alpha-1" – Bartonellaceae, "alpha-2.1" – Acetobacteraceae, "alpha-2.2" – Acetobacteraceae, "bifido" – Bifidobacteriaceae, "firm-4" – Lactobacillaceae, "firm-5" – Lactobacillaceae, "gamma-1" – Orbaceae, "gamma-2" – Orbaceae, "beta" – Neisseriaceae. Descriptions of some of these have since been published, in which case the genus name is presented following the phylotype name. The "beta" phylotype is *Snodgrassella alvi*. The "gamma" phylotypes and some unclassified sequences may include *Gilliamella apicola* (formerly "gamma-1"), *Frischella perrera* (formerly "gamma-2"), "gamma-3," and "gamma-4". The "alpha-2.2" phylotype is *Parasaccharibacter apium*.

4.2.5 Summary of sequences and classification

After filtering, we retained 6,251,931 of the total sequences, distributed over 79 OTU clusters. These sequences were distributed across 60 samples, as shown in Figure 4.1. This set of 60 samples was composed of 2 reproductive queens, 2 virgin (7-day-old) queens, 16 males, 22 foragers, 18 nurses; 4 queen and 2 male

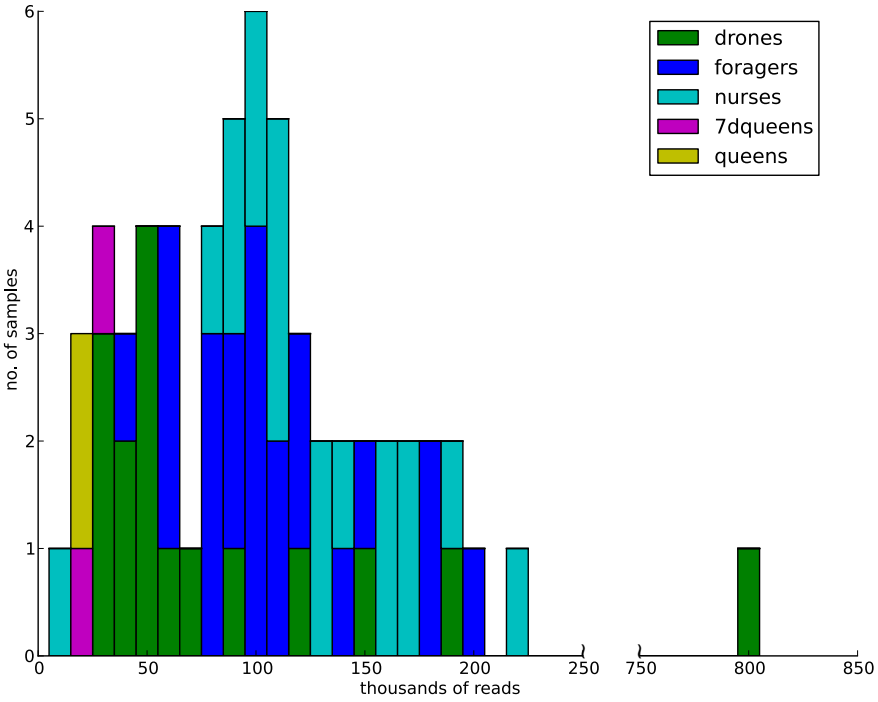

Figure 4.1: **Histogram showing the distribution of the number of reads across different castes.** Bins are of width 10,000 reads. In general, the nurse samples have the largest number of reads, while the queen samples have the least.

samples were removed from analysis due to their low PCR yields. The number of sequences was significantly different between castes (i.e., nurses, foragers, males, queens) (Kruskal-Wallis $\chi^2 = 17.78$, $p = 0.0005$, $n = 60$), with nurses having the most reads and queens having the least. The variation in read count did not affect the number of OTUs detected in each group. Neither DNA concentration nor number of reads were significant predictors of OTU number in a negative binomial regression after removing the one outlier sample with greater than 3.8-fold more reads than all other samples (LR $\chi^2 = 3.02$, $n = 59$, $p = 0.22$). We detected an average of 26.4 ± 6.93 OTUs per individual. The number of OTUs detected was not significantly different between castes (Kruskal-Wallis $\chi^2 = 4.85$, $p = 0.18$, $n = 60$). The OTUs were classified into 44 taxonomic groups at the family level. Over 99% of the total sequences belonged to OTUs that were previously classified as honeybee phylotypes [5].

4.3 OTUs of the honeybee gut microbiome

Figure 4.2 shows the relative abundances of phylotypes across samples. The microbiomes of nurses and foragers were highly similar to those observed in previous studies of worker honeybee microbiomes. Of the 4,402,282 sequences from nurses and foragers, 4,401,036 (99.97%) belonged to OTUs classified as phylotypes

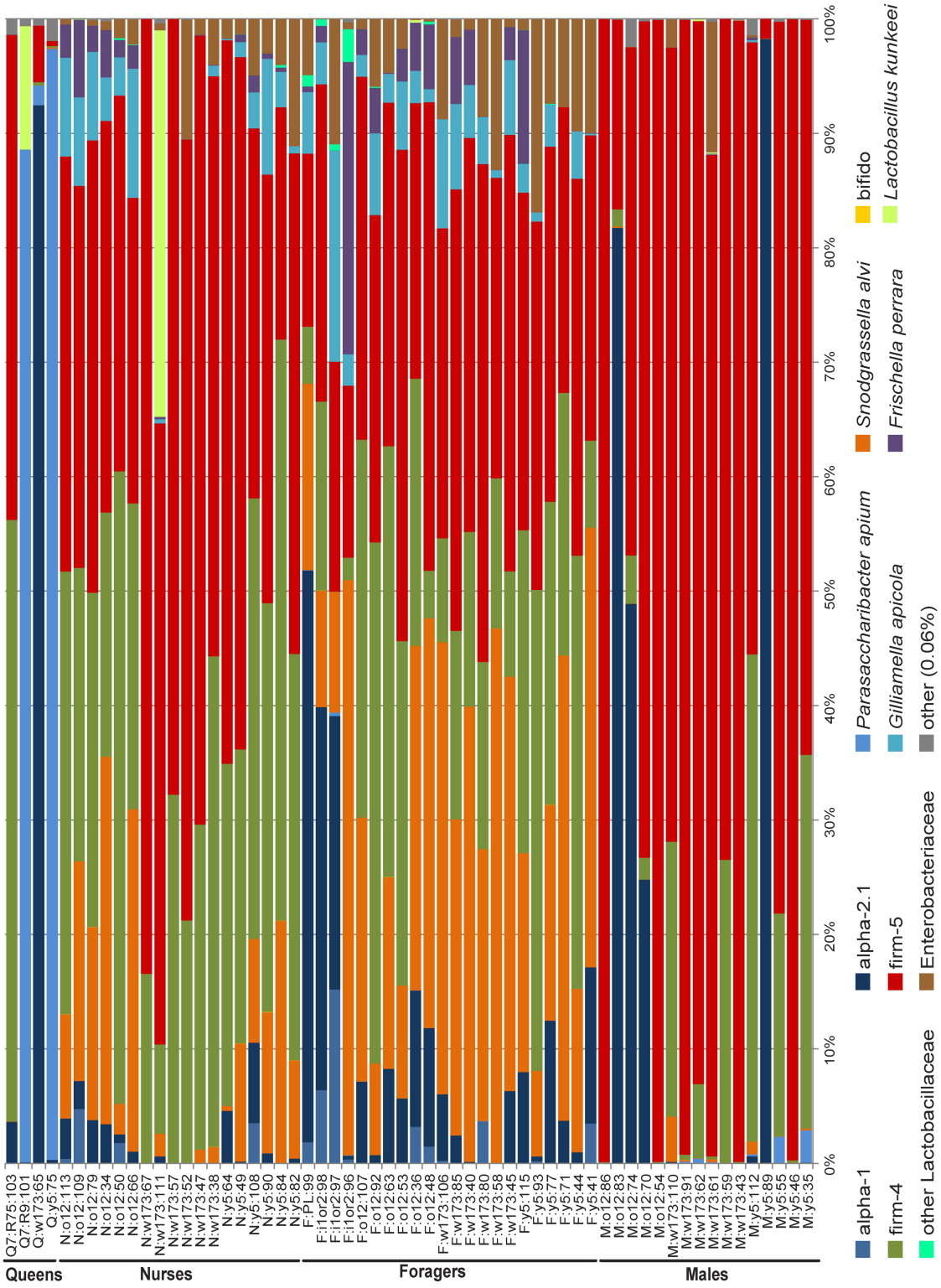

Figure 4.2: **Family-level taxonomic composition of individual gut microbiomes in honeybees.** Proportion of each taxa in total microbiome composition is represented by the proportion of the coloured bar.

previously characterized in honeybees or were from bacterial families previously detected in honeybees (e.g., other *Lactobacillaceae*). Our results show that males and queens—whose gut microbiomes had not been previously characterized—have most of the same bacteria in their hindguts as workers. Over 99% and 98.87% of the total reads from males and queens, respectively, were classified as bacteria previously detected in honeybees. However, despite this similarity in phylotypes present, the relative proportion of certain bacteria did vary significantly among castes. For example, queens had a higher relative abundance of the alpha-proteobacteria *Parasaccharibacter apium* than the other castes. Caste membership was a significant predictor of *P. apium* abundance (LR $\chi^2 = 10.51$, $n = 60$, $p = 0.001$). Only three (7.5%) workers had more than ten sequences of *P. apium*; all three workers were from the indoor colonies. The large abundance of *P. apium* was not characteristic of all queens, however. One of the young 7-day-old queens had no detectable *P. apium*, but the two queens that did have high *P. apium* (one old, one young) had more than 13 times higher abundance of the phylotype than any other individual.

The *Lactobacillus* species firm-5 was the most abundant phylotype in males, foragers, and nurses, representing 75.7%, 30.4%, and 49.4% of total sequences from these castes, respectively. Firm-5 was the fourth most abundant bacteria found in queens, however, with *P. apium*, alpha-2.1, and firm-4 ranking ahead of firm-5 (Figure 2). Caste membership was a significant predictor of firm-5 abundance (LR $\chi^2 = 6.97$, $n = 60$, $p = 0.0083$). Males had significantly higher relative abundance of firm-5 than any other group (ANOVA $F = 18.20$, $n = 60$, $p < 0.0001$, post-tests $p < 0.05$), while nurses had a significantly higher proportion of firm-5 than foragers.

4.4 Statistical analyses

To study patterns in the distribution of OTUs among different samples, a Bray-Curtis dissimilarity matrix was constructed for pairwise dissimilarities between samples. All downstream analyses were based on this matrix, and were performed using the software package Primer 6 with the PERMANOVA+ add-on [20].

4.4.1 Unsupervised clustering of samples

We used two methods to cluster the samples based on Bray-Curtis dissimilarity: average-linkage clustering to create a dendrogram for hierarchical clustering, and non-metric multidimensional scaling (nMDS).

Figure 4.3 shows the dendrogram that results from average-linkage clustering. At 35% similarity, all nurses and foragers cluster together with the exception of one forager and one nurse. The nurse that did not cluster with the other workers was an outlier with a small number of reads (see Figure 4.1). There is

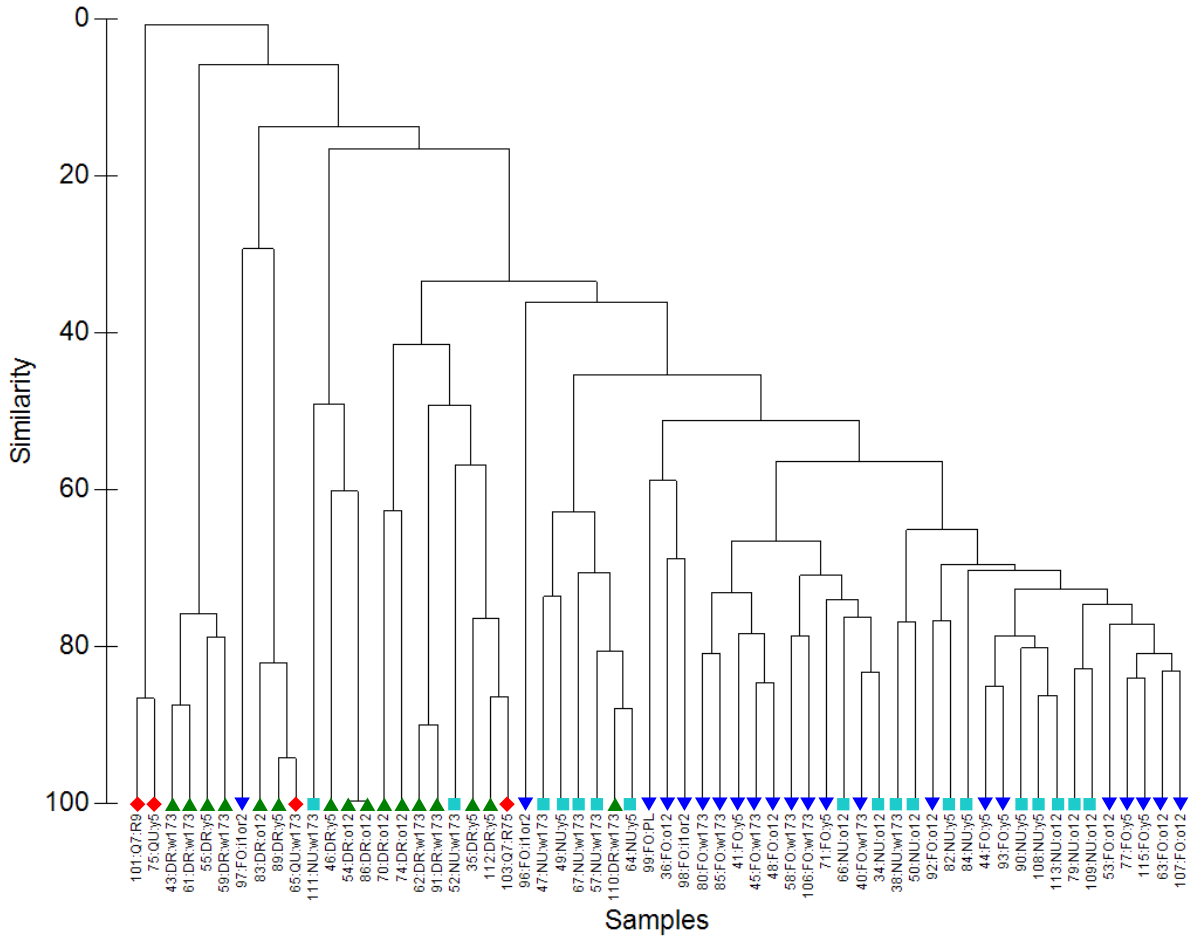

Figure 4.3: **Dendrogram from average-linkage clustering of samples based on Bray-Curtis dissimilarity.** The four castes are represented by different symbols. Leaf labels follow the format caste:colonyID:individualID.

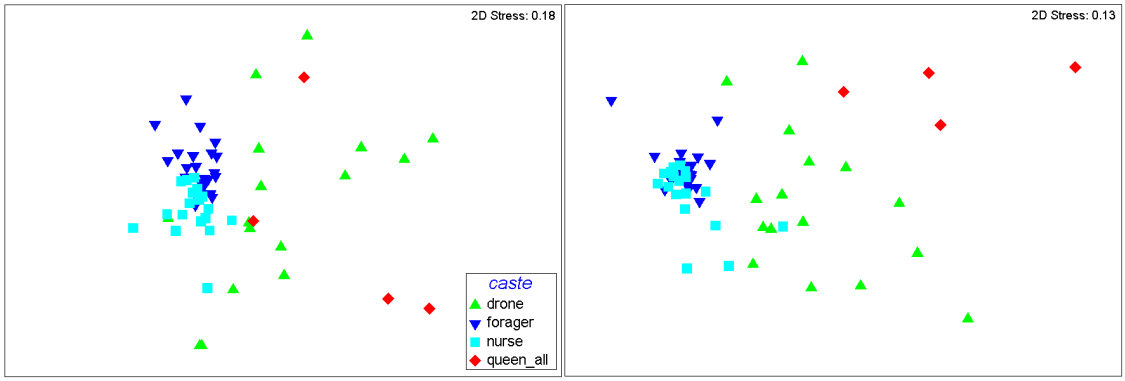

Figure 4.4: **nMDS in two dimensions.** The plots are shown both for relative OTU abundance data (left) and for log-transformed abundance data (right).

also one male sample that falls within this cluster of mostly worker bees; this sample was an outlier with the largest number of reads. Although there was notable diversity among the male samples (i.e., they formed multiple clades with variable similarity), there were several clusters of males with high similarity that did not include queens, nurses or foragers.

These clustering patterns are supported in the non-metric multidimensional scaling plot based on Bray-Curtis dissimilarities (Figure 4.4), which shows overlap among the nurse and forager samples, and separation of these worker samples from most of the males and queens. Foragers from indoor and outdoor colonies did not form distinct clusters, and neither did young and old queens, so these were combined in the downstream analyses. The nMDS algorithm was run for Bray-Curtis dissimilarities derived from both the original relative-abundance data, as well as for log-transformed relative-abundance data. The purpose of the log transformation was to reduce the effect of highly abundant OTUs; this yields an nMDS ordination with a slightly better (lower) stress value.

4.4.2 Shannon diversity

The Shannon index is commonly used as a measure of diversity in ecology, and is simply the information-theoretic entropy in the relative abundances of OTUs. Given a sample that has N OTUs with relative abundances $\{p_i\}_{i=1}^{N}$, the Shannon diversity for the sample is defined to be

$$H = -\sum_{i=1}^{N} p_i \log p_i \tag{4.1}$$

Figure 4.5 shows the Shannon index for the four different groups. Shannon diversity was significantly different between groups (ANOVA: $F = 49.15$, $p < 0.0001$, $n = 60$). Nurses and foragers had the highest

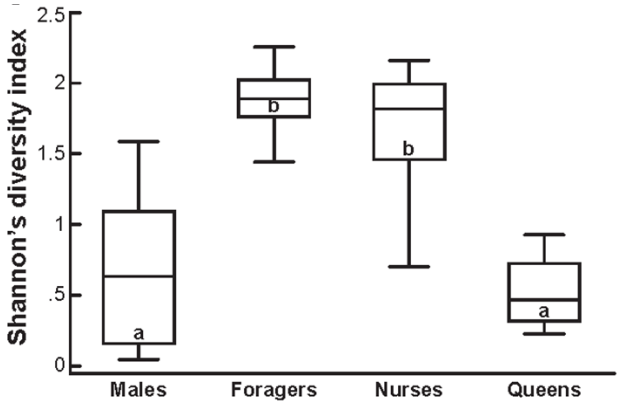

Figure 4.5: **Shannon diversity index for different social groups.** Lines represent the median index value, boxes demarcate the interquartile range, whiskers are the adjacent values, and markers are the outer adjacent values. Figure provided by Karen Kapheim.

diversity index, and both had significantly higher diversity than males or queens (post-test: $p < 0.05$), but were not significantly different from each other.

4.4.3 Canonical analysis of principal coordinates

We tested the strength of the association between the multivariate OTU data and caste using the method known as canonical analysis of principal coordinates (CAP) [4], implemented in Primer 6. In this ecological ordination method, the Bray-Curtis dissimilarity matrix is first used in principal coordinates analysis (also known as metric multidimensional scaling) to obtain a representation of the data in the basis of an orthonormal set of axes. Canonical discriminant analysis is then used on this representation to find a set of axes that maximizes the delimitation among the four different castes that each sample belongs to. We found that there was a clear separation of castes by canonical axes 1 and 2, which together explain 39% of the variance (Figure 4.6).

We then used leave-one-out cross-validation to classify samples based on multivariate OTU abundance data, and found that OTU frequencies correctly predicted group membership for 47 of 60 (78.3%) individuals. Microbiome data was most predictive of male (93.8% correct), nurse (72.2% correct), or forager (77.3% correct) caste membership, but the strength of this association was lower for queens (50% correct). The five foragers that were misidentified were predicted to be nurses, and four of the five misidentified nurses were predicted to be foragers, again emphasizing the similarity between the microbiota of the two types of worker. One male was classified as a nurse, and two queens (one young, one old) were classified as males.

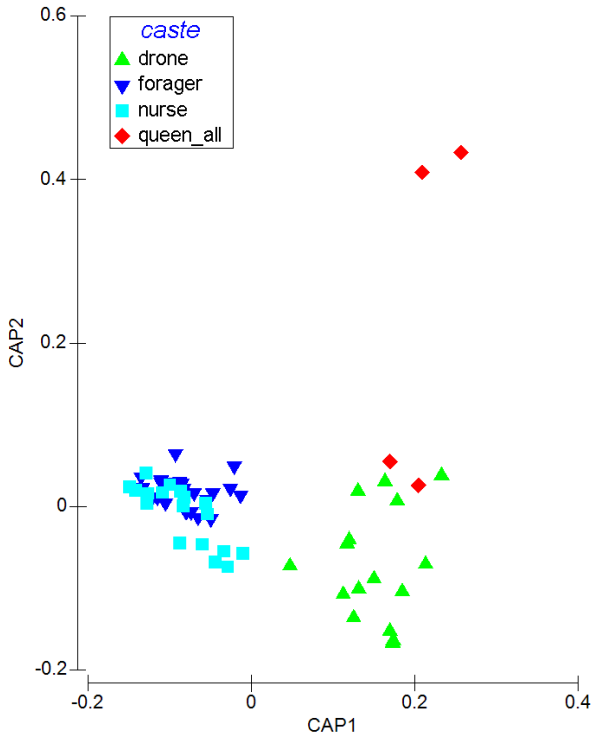

Figure 4.6: **Ordination by canonical axes of prinicipal coordinates.** The first axis delineates workers from non-workers, while the second separates two of the queen samples from the rest.

4.4.4 Partitioning variance between caste and colony

Since the bees came from different colonies, this is another factor that might influence the composition of individual microbiomes, in addition to caste. PERMANOVA is a semi-parametric method to perform multivariate analysis of variance with multiple groups [3], and is implemented as an add-on for the Primer 6 software package. We used PERMANOVA for a mixed-effects model with caste as a fixed effect, colony as a random effect, and including the interaction between caste and colony. The PERMANOVA+ add-on computes pseudo-F statistics to test the null hypothesis that the centroid of the abundance vectors for the different factors are the same, and computes p-values using Monte Carlo sampling.

We found that caste, colony and the interaction between caste and colony were all significant factors in PERMANOVA. Post-hoc pairwise comparisons revealed that male bacterial communities were significantly different than that of workers, but not queens (see Table 4.1). The marginal differences between the two worker castes (nurses and foragers) disappeared when this analysis was repeated for the log-transformed data (Table 4.2), suggesting that these differences are driven by differences in the proportions of highly abundant OTUs. In addition, log-transformation reveals marginally significant differences between queens and workers, perhaps indicating that queens have a unique rare microbiome.

	males	foragers	nurses	queens
males		$t = 2.69,\ p = 0.0005$	$t = 2.23,\ p = 0.0054$	$t = 1.08,\ p = 0.4288$
foragers			$t = 1.83,\ p = 0.0301$	$t = 1.45,\ p = 0.1911$
nurses				$t = 1.42,\ p = 0.2172$
queens				

Table 4.1: **Results from pairwise comparisons of OTU variance among castes.** PERMANOVA based on Bray-Curtis dissimilarities.

	males	foragers	nurses	queens
males		$t = 3.16,\ p = 0.0008$	$t = 2.60,\ p = 0.0024$	$t = 1.48,\ p = 0.1606$
foragers			$t = 1.36,\ p = 0.1496$	$t = 2.34,\ p = 0.0527$
nurses				$t = 2.25,\ p = 0.0512$
queens				

Table 4.2: **Results from pairwise comparisons of OTU variance among castes.** PERMANOVA based on Bray-Curtis dissimilarities computed for log-transformed OTU count data.

4.5 Discussion

Our results support the hypothesis that sociality facilitates the development of a shared microbiome. We identified a set of core microbial taxa detected among at least one bee of each of the four social groups, consisting of phylotypes found previously in honeybee workers. The microbiomes of workers in our study was similar to those found in previous studies that also analyzed worker bees [83, 101].

The similarity between nurse and forager microbiomes is surprising, because variation in the microbiome is generally associated with host diet, behaviour, and physiology [31, 129], all of which vary between these two groups of specialized bees. Changes in diet and gut physiology accompany the age-related transition from hive work as nurses to foraging outside of the hive, and it is possible that there are corresponding shifts in the abundance of certain bacteria in the microbiome. In our study, nurses had significantly more *Lactobacillus* species (firm-5) than foragers. This group of bacteria function in carbohydrate metabolism and transport [44], and so differences in their relative abundance may reflect variation in the hindgut environment due to differences in the diets of nurses and foragers.

The stability of the worker microbiome may be maintained by frequent re-inoculation among interacting nestmates via food exchange. However, nurses also feed male and queen bees within the hive, but despite this, the microbiomes for those two castes have lower diversity than the worker microbiomes (Figure 4.5). Perhaps the increased strain diversity among workers is the result of the regular exchange of environmentally derived sources of microbes between returning foragers and receiving nurses. Foragers are likely responsible for introducing new strains of bacteria from the environment to the colony, which may then be filtered and distributed among colony members by nurses. Nurses feed all colony members with food made in specialized

glands in the head (hypopharyngeal glands), and thus the entire colony's nutrients, and possibly microbes, pass through nurses, but do not necessarily end up in their gut. For example, we did not detect *P. apium* in the hindguts of nurses, but it was highly abundant in queens. Previous research has found this bacterium in the crops of foragers [23] and the crops and hypopharyngeal glands of nurses, as well as in the royal jelly they feed to the queens [24]. All individuals in a colony likely receive a diverse suite of bacteria from nurses, but most bacteria either do not persist or are extremely rare in all colony members.

Some caste-related differences in microbiome composition may be related to differences in reproductive activity. For example, *P. apium* was significantly more abundant in queens than all other castes, and was nearly absent from most workers, who do not reproduce. Genes expressed in strains of these bacteria found in workers are linked to ferrodoxins and oxidoreductases associated with metabolism as well as membrane/envelope biogenesis [44], which are important components of oogenesis (the formation of eggs). We therefore speculate that they may be useful for converting large amounts of dietary protein to egg production.

Finally, chemical communication within the colony may be regulated in part by the distribution of microbes. It has been reported that bacteria are responsible for group-specific social odours among hyenas [118], and bacteria have been shown to influence mating preferences in *Drosophila melanogaster* [108]. Pheromones play an important role in regulating honeybee social behaviour, and it is possible that having a colony-specific microbiome facilitates nestmate recognition.

Part II

Interindividual variation and plasticity in foraging activity

Chapter 5

Introduction to foraging, division of labour and wealth distributions

In this chapter we provide some background to our study of bee flight activity, described in chapter 6. First, we describe how honeybees forage for food, recruiting other foragers using the dance language. One question that this raises is: what is the optimal method of searching a space—a field, say—for food? We explore this question with a discussion of Lévy flights, which have been proposed as an optimal foraging strategy. We end with a discussion of wealth distributions in economic data; in chapter 6, we use the framework discussed here in characterizing the distribution of foraging activity among honeybees.

5.1 Foraging strategies in honeybee colonies

A honeybee colony's nutritional needs are met by two raw food sources obtained from flowers: pollen and nectar. Nectar is the sugary substance provided by flowers to attract pollinators such as honeybees, and is processed by them into honey, for the purpose of long-term storage. Apart from being another food source, pollen eventually produces the male gamete cells in plants; when a bee lands on a flower to attempt to collect either of these substances, some of the pollen that it previously collected (or was stuck to its body) is transferred to the flower, allowing pollination to happen between the two plants that the flowers belong to.

Foragers are the set of worker bees responsible for finding these substances and bringing them back to the hive. Additionally, foragers collect water, which is used both for nutritional purposes and for thermoregulation of the hive. Honeybee colonies in the wild require 15–30 kg of pollen, and 60–80 kg of honey, to survive. Based on these requirements, and given that the average forager on a single foraging trip carries 15 mg of pollen or 32 mg of nectar, Winston [133] estimates that the foragers in a colony make a total of 1×10^6 trips annually to collect pollen, and about 4×10^6 trips annually for nectar.

In order to perform this impressive feat, many different aspects have to be coordinated among the workers in the society. First, the age at which a particular bee forages is modulated based on the colony's needs and the availability of food. The colony has to discover particular flowers to forage from, and allocated

an appropriate number of workers to each flower; furthermore, bees expend different amounts of energy depending on the species of flower, as some are easier to forage from than others. The colony has to also regulate how the foragers are divided among the different food sources—pollen, nectar or water—depending on the colony's needs, as well as (in the case of water) the hive temperature.

Foragers may be divided by their two subroles: scouts and recruits. And at any given time, a forager can be either employed (i.e. "has been assigned" to a particular food source and is making trips back and forth) or unemployed (either because she is a novice forager or because she has recently abandoned a depleted patch of flowers). Scouts are those foragers responsible for finding new food sources. Upon the discovery of a new source, scouts return to the hive and recruit other bees to that source using the famous waggle-dance language discovered by Karl von Frisch [128]. The dance takes place in one section of the hive, close to the entrance, known as the dance floor. Here, among a throng of so-called follower bees—potential recruits—she performs movements whose general form in shown in Figure 5.1. The dance takes place in two repeated phases: a waggle phase, consisting of waggle runs and high-frequency (~ 260 Hz) vibrations of the wings; followed by a return phase. The direction of the waggle run relative to the Earth's gravitation field communicates information about the direction of the food source (see Figure 5.2), while the duration of the run correlates with distance to the food source. This method of communication is impressively robust: for example, bees are able to correct the direction of the waggle run as the sun changes its position along the celestial equator, as shown in Figure 5.2; and they can do it despite cloud cover, by relying on ultraviolet frequencies of sunlight. At the same time, the scout bee also communicates the odour of the flower at the food source, and recruits appear to use this to help pinpoint the location.

Once a forager makes a successful trip and brings back food to the hive, the food has to then be stored. Foragers that bring back pollen do so by carrying a little ball of pollen on each hind leg. When they return to the hive, they deposit the pollen into storage cells, and leave the rest of the processing to younger worker bees. Foragers that bring back nectar or water, meanwhile, offload the goods to a receiver bee near the hive entrance via trophallaxis. The receiver bees either distribute the nectar among other bees or process it into honey for storage. This processing happens in their bodies via both physical and biochemical manipulations, and takes about 25 minutes until it is ready to be placed into a honeycomb, which is then sealed to prevent fermentation.

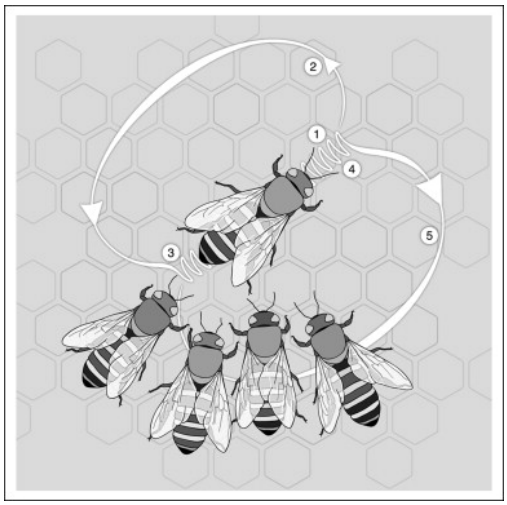

Figure 5.1: **The honeybee waggle dance.** The dancer performs a waggle run (1) followed by a turn to the left (2) to circle back to the starting point (3) of the waggle run. It then starts another waggle run (1), followed by another turn—usually to the other side (4)—and the cycle repeats. Reproduced with permission from Ref. [52].

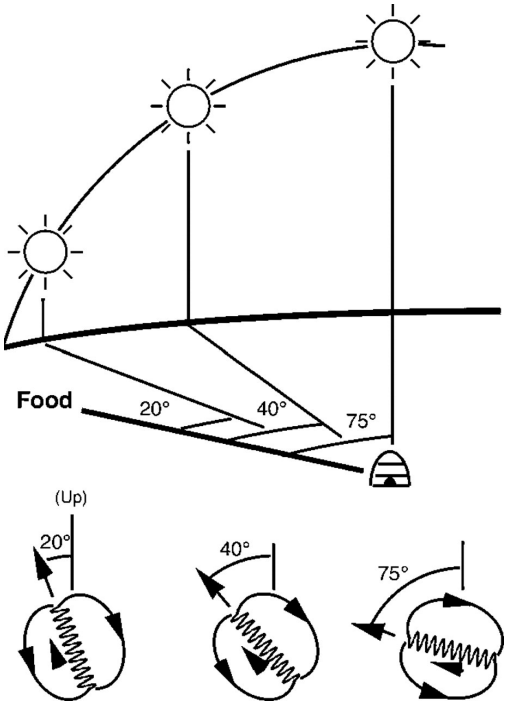

Figure 5.2: **Communicating direction using the waggle dance.** The direction of the waggle run relative to the gravitational field maps to the direction of the food source relative to the sun. Figure reproduced from Ref. [36]

5.2 Lévy flights and foraging

How does a scout bee go about finding new food sources? That is, how does she set her flight path to sample the environment around the hive, maximizing her chances of encountering a food source, while minimising the energy she expends on the process? The study of how animals move and why they move in a particular way is an important part of animal behaviour, and physicists have made some important contributions to quantitatively understanding animal movement. Although this subject is slightly out of the scope of our study described in chapter 6, we review some of the basic ideas from forager theory here, as they motivated our initial interest in automated tracking of honeybees flights. For a much more comprehensive overview of foraging theory, we recommend the monograph by Viswanathan et al. entitled *The Physics of Foraging* [126].

Let us consider an agent moving in a 2-dimensional landscape, perhaps searching for food. The simplest null model for the motion of the agent would be that of a free random walk, i.e. the agent takes a step of length l at each time step τ, reorienting itself in a random direction after each step. The Fokker-Planck equation describing the probability distribution in space $\vec{x}$ and time t of this random walker is simply the linear diffusion equation,

$$\frac{\partial}{\partial t}P(\vec{x},t) = D\nabla^2 P(\vec{x},t), \tag{5.1}$$

with $D = l^2/6\tau$ as the diffusion coefficient.

The second spatial moment of this distribution—the mean-squared displacement—grows linearly in time:

$$\langle (x - x_0)^2 \rangle = 2Dt. \tag{5.2}$$

A simple test of whether this null model actually describes our agent's movement is to see whether its mean-square displacement (MSD) does in fact scale linearly in time. We can write the scaling of the MSD in the generalized form $\langle x^2 \rangle \sim t^{2H}$, where H is called the Hurst exponent, with $H = 1/2$ corresponding to regular diffusion.

In real biological systems, it is not necessarily the case that $H = 1/2$. Our focus is on the *superdiffusive* case $H > 1/2$, as this is relevant to foraging. The simplest modification to incorporate in order to make the model more biologically realistic is to include directional persistence: if an animal makes a step in a certain direction, it is more likely to stick to that direction in the next time step, rather that choosing a direction isotropically at random. However, this is not sufficient to obtain superdiffusive scaling, as the memory effect induced by this only has a finite range [126].

Instead, one can make the step size a stochastic quantity drawn from some distribution $\Lambda(l)$, rather being

than a fixed length l. Lévy flights arise when this step-size distribution has a power-law tail:

$$\Lambda(l) \sim l^{-(1+\alpha)}, \tag{5.3}$$

where the exponent α is known as the Lévy index. For $\alpha \geq 2$, the central limit theorem applies and leads to regular diffusion. However, for $0 < \alpha < 2$, the variance diverges and the random walker's positional distribution $P(\vec{r}, t)$ converges instead to a Lévy-stable distribution, and this is what defines Lévy flights. The second moment for this distribution diverges, so one cannot study the MSD, but instead we can defined an analogous width variable (e.g. half width at half maximum), and this pseudo-MSD grows in time with a subdiffusive Hurst exponent $H = 1/\alpha$.

Why might Lévy flights be used in foraging? To get an intuition for this, let us first clearly define the foraging problem: the forager moves in some space in which there are some target sites (such as a food source). We will assume that the forager's search process is destructive: that is, once it reaches the target site, it depletes the resource, so that it is no longer a target.

Suppose the forager uses a ballistic strategy, that is, it chooses a direction in space and simply moves in that fixed direction. This has the advantage that if the target lies along the ballistic path, then the forager reaches it in the smallest time possible. The tradeoff is that if the target does not happen to lie along the path, then the forager never reaches the target. Now suppose instead that the forager uses a regular random walk. This has the advantage that the space is explored much more thoroughly than in the ballistic strategy. The downside is that, due to the nature of the random walk, the forager will often end up returning to places it has already explored.

A strategy based on Lévy flights balances these two extremes. The forager spends many time steps in one region of space, because most of the step sizes sampled from the distribution in Eq. 5.3 will be relatively small. Occasionally, however, it will take a relatively large step, allowing it to ballistically move away from the local region that it has been exploring, and thus preventing it from repeatedly searching that region (which is the drawback of foraging with a regular random walk). This qualitative idea was formulated in a mathematical model by Viswanathan et al. [127], in which they showed that the Lévy index $\alpha = 1$ was optimal in maximizing the ratio of the number of visited target sites to total distance traversed.

So have living creatures actually evolved to discover this optimal strategy? The first empirical study—which motivated the theory of Lévy flight–based foraging—looked at the foraging of albatrosses in the south Atlantic [125]. This study found a power-law tail with exponent -1 in the histogram of flight durations. However, subsequent work revealed problems with parts of the data, and cast the discovery of Lévy flights

in doubt [43]. Nevertheless, with the advent of better statistical methods to detect power-law tails in data, newer studies are establishing more firmly that Lévy flights do occur in nature [95].

The flight patterns of honeybees have been studied using harmonic radar technology [97]. Bees are equipped with passive tags that radiate a harmonic of an incoming radio wave, allowing their location in flight to be detected over distances of up to 250 m. Using this method, it has been established that honeybee do appear to perform scale-free Lévy flights with $\alpha \approx 1$, both when trying to find a food source [94] and when trying to located their hive after being displaced [93].

In the project described in chapter 6, we do not measure the spatial trajectories of flights, so we cannot measure the statistics of flight distances. What we measure instead is the distribution of flight activity, and we characterize this using the framework of wealth distributions, as discussed in the next section.

5.3 Wealth distributions, Lorenz curves and Gini coefficients

We turn now from a power law that appears in studying the mechanistic movements of animals to one that occurs for the distribution of wealth or labour among individuals in a society. In a honeybee society, workers are genetically related, and the set of workers specializing in a particular task at a moment in time are, from the point of view of the colony, essentially indistinguishable. What does this mean for the amount of labour that each performs? Is the distribution of labour across bees essentially a delta function, with bees performing the same amount of work? Or is the distribution skewed in some way? In chapter 6, we address these questions using automated monitoring of a large fraction of the foraging workforce of several honeybee colonies. Here we provide some context for our methods of analysis.

The distribution of labour among individuals in a human society is difficult to measure, but economists have long studied the related distribution of wealth in human societies. This distribution often has a power law tail, and in this case, is called the *Pareto distribution*, named after the Italian engineer and economist Vilfredo Pareto (1848–1923). The probability density function in wealth w for the Pareto distribution may be written as follows:

$$P(W = w) = \frac{\alpha w_0^\alpha}{w^{\alpha+1}} \Theta(w - w_0), \tag{5.4}$$

where Θ is the Heaviside step function, which truncates the distribution below the cutoff w_0. The exponent is defined so that α is simply the exponent for the complementary cumulative distribution,

$$P(W > w) = 1 - P(W \leq w) = \left(\frac{w_0}{w}\right)^\alpha. \tag{5.5}$$

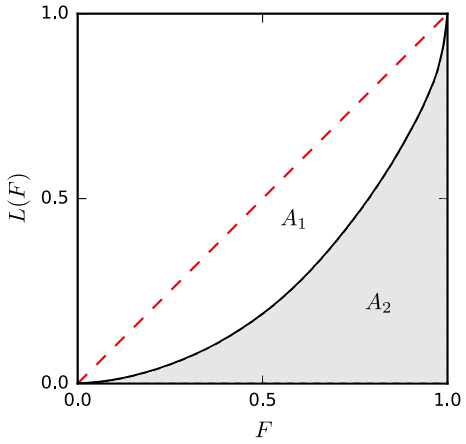

Figure 5.3: **The Lorenz curve.** The dashed line corresponds to an equal distribution of wealth; deviations show up as a concave skew away from the dashed line. This skew is measured by the Gini coefficient, which is the ratio of the area A_1 in between the dashed line and the curve to the total area $A_1 + A_2$.

This cumulative distribution $F(w) \equiv P(W \leq w)$ is often plotted in the form of a *Lorenz curve*,

$$L(F) \equiv \frac{\int_0^F w(F')\ dF'}{\int_0^1 w(F')\ dF'}. \tag{5.6}$$

This may be thought of as a cumulative share vs. rank plot: F is essentially a continuous rank variable, and $L(F)$ gives the share of total wealth in the subpopulation of individuals of rank at least F.

If all individuals have the same wealth, then the cumulative share just grows linearly with rank: $L(F) = F$. If there is an unequal distribution, the cumulative share increases faster as the rank increases (see Figure 5.3. The degree of inequality is captured by the *Gini coefficient* G, which is proportional to the area in between the Lorenz curve and the line $L(F) = F$. Specifically, the Gini coefficient is defined in terms of the areas under these curves (see Figure 5.3) as

$$G \equiv \frac{A_1}{A_1 + A_2} = 2A_1. \tag{5.7}$$

In the special case where the wealth distribution is Pareto, the Lorenz curve and Gini coefficient are set by the power-law exponent α:

$$L(F) = 1 - (1 - F)^{1-1/\alpha} \quad \text{and} \quad G = \frac{1}{2\alpha - 1}. \tag{5.8}$$

So what does the wealth distribution really look like for a country? Figure 5.4 shows the (cumulative) distribution of income in the US for each year from 1983–2001. These distributions show two regimes:

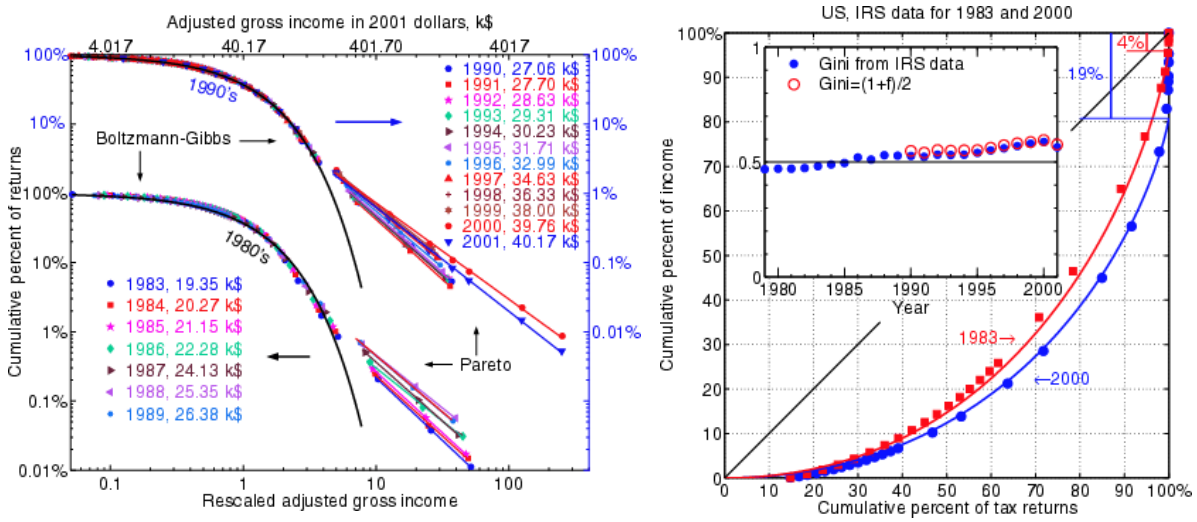

Figure 5.4: **US income data.** Left: Cumulative probability distributions of US tax returns from 1983–2001. The distributions are for tax return values normalized by the mean value. Right: Lorenz curve for the first and last year in this period. The inset shows the Gini coefficient as a function of time. Figures reproduced from Ref. [109].

an exponential distribution with a power-law tail. These have been studied in the field of econophysics and attempts have been made to model money exchange and debt in order to try to reproduce these distributions [136].

Chapter 6

Extreme interindividual variation and plasticity in foraging activity

6.1 Introduction

The idiom "busy as a bee" traces back to 14th-century England, appearing in Chaucer's *Canterbury Tales*,[1] and raises a number of interesting questions about the organization of a society of bees. First, a natural question to ask is whether all bees are equally busy—are there not perhaps some bees that work harder than others, just as in a human society? And from the perspective of optimizing the efficiency of a honeybee society, what is the ideal distribution of busyness among the worker bees? Since the reproductive unit of honeybees is the colony rather than individual workers, one might expect that the dynamics of the individuals bees should be determined not by their own individual needs, but by the collective needs of the society.

In reality, entomologists have observed that in many eusocial species and for a range of behaviours, individual workers show a great deal of variation in their levels of activity [12, 35, 60, 79]. The term elitism is often used to describe the phenomenon in which a subset of the individuals performing a given task display a much higher activity level than others. Several explanations have been proposed for the adaptive value of variation in activity level within the set of individuals performing a task. Individuals with lower activity may represent a reserve workforce, enabling the colony to rapidly respond to sudden and unpredictable demands or opportunities [92]. Alternatively, individuals within a workforce may have behavioural specializations that make them more or less able to perform a task efficiently [90]. Moreover, factors promoting variation in activity level may be different across different species, or for different tasks within a single species.

Despite a vast body of literature devoted to foraging behaviour of honeybees, reports of systematic studies of elitism in honeybee foragers are absent, and only a few studies in the last half century have alluded to elite forager honeybees. Sekiguchi and Sakagami [106] described the presence of a few individual honeybee foragers with unusually high flight activity relative to other foragers from the same colony. Ribbands [96] also reported a maximum of 150 trips in a single day made by individual nectar foragers to an artificial feeder. These findings indicate the potential for great variation in activity level within a colony's foraging

[1] Thanks to Paul Tenczar for pointing out the provenance of this quote and its relevance for our study.

population, but to our knowledge, no prior study has investigated whether elite honeybee foragers represent a distinct sub-group, how the tendency to extreme foraging activity might change over the individual lifespan, or how foraging elitism relates to colony-level regulation of foraging activity.

The likely reason why these questions have not been previously explored in honeybees is the difficulty of monitoring flight activity. Thus far, biologists have relied on manual observation and recording of individual behaviour in honeybees, and the kind of continuous observations that would be required to quantitatively address these questions are challenging to perform manually. The recent appearance of technologies that enable automated detection of individual insects and other small animals has offered the possibility to pursue research that previously faced these obstacles. In this chapter, we describe how we addressed these questions by tracking the lifetime flight activity of several hundred individuals in several different colonies of honeybees using existing radio-frequency identification (RFID) tagging technology.

Our use of RFID technology allowed us to analyze the flight activity of a large sample of individually identified bees over an extended time period. This dataset allowed us to measure the distribution of activity levels among honeybee foragers, and to investigate potential differences between high-activity foragers and other foragers within the same colony We also examined the effect of the removal of high-activity foragers on the activity levels of the remaining tagged individuals.

6.2 Experimental setup

6.2.1 Monitoring technology

Bees were tagged with laser light–activated "p-Chip" microtransponders (PharmaSeq, New Jersey). These are passive tags that use RFID (radio-frequency identification) technology: that is, the tag surface contains photocells that are activated by light from a laser detector device, and these then emit a radio-frequency signal that is read by an antenna attached to the reader. This signal contains an identification number that is unique to the tag, and can be read by the detector at a distance of up to 10 mm away from the tag. The detector is connected to a computer via USB, and the output is processed by p-Chip detector software provided by PharmaSeq.

Due to the small (1.5 mm) diameter of the activator beam, two tags were attached to each bee to increase the likelihood of detection (Figure 6.1). Each tag had dimensions 500 μm $\times$ 500 μm $\times$ 100 μm and weight 90 μg. Two tags fit easily on the thorax of the bee, and their combined weight was very small (0.6%) relative to the average load carried by a nectar forager [133], so it is unlikely that the presence of the tags impaired natural behaviour.

Figure 6.1: **Two p-Chip tags are placed on each bee.** Each tag is associated with a unique ID to enable identification of individual bees. Photograph courtesy of Paul Tenczar.

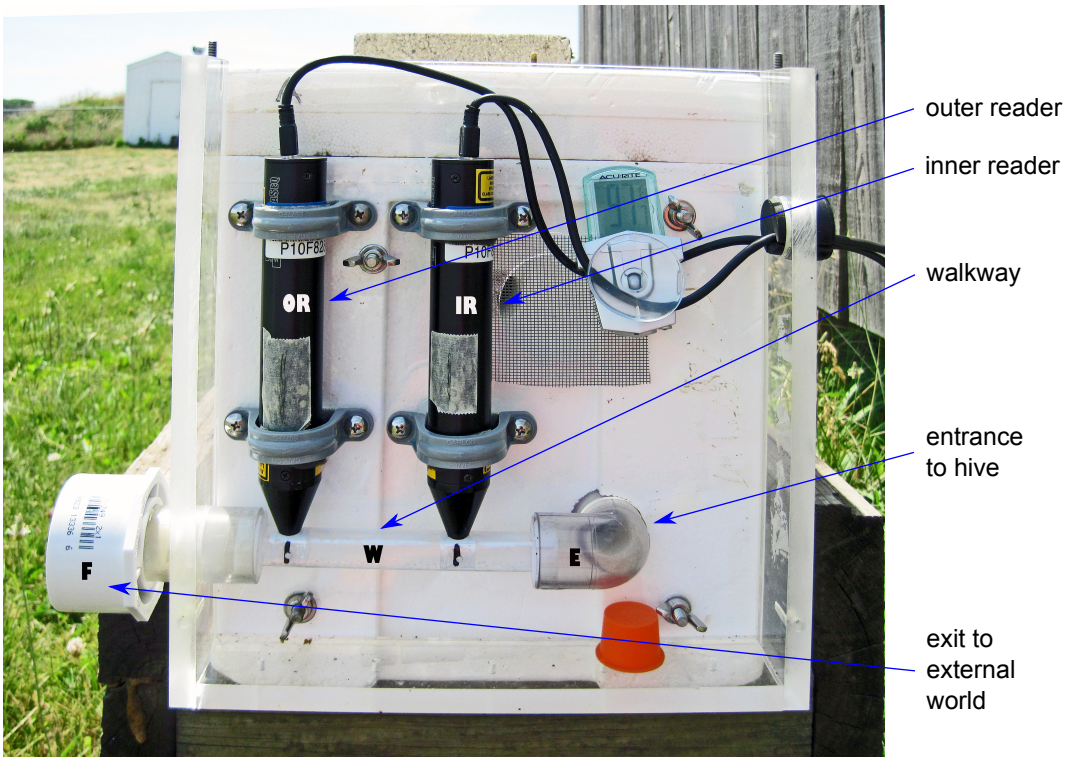

Figure 6.2: **Setup of detectors in the walkway leading to the hive.** Bees live inside a Styrofoam hive box with a single entrance to the outside, via a walkway consisting of a plastic tube. Two detectors are placed above the tube to monitor the entrance. Photograph courtesy of Paul Tenczar.

To read tagged bees, two detectors were placed above a plastic tube of cross section 10×10 mm^2 that served as a walkway to the entrance of the hive (see Figure 6.2). Bees passed sequentially under each detector as they entered and exited the hive via this walkway. The order of detection by each detector could be then be used to infer the direction of travel.

The top and sides of the walkway were coated with Fluon (Bioquip Products, Rancho Dominguez, California) while the floor of the walkway had numerous small drilled holes to provide grip; this encouraged bees to walk on the floor of the walkway, maximizing the likelihood that their tags would be detected. Upon detecting a tag on a passing bee, the detector passed the RFID signal to the attached PC, and we recorded the tag ID, detector number, and a time stamp for the detection.

6.2.2 Colony setup

Frames of honeycomb containing pupae were collected from 5–10 typical colonies maintained according to standard procedures at the University of Illinois Bee Research Facility, Urbana, Illinois, with naturally mated queens, and stored in a dark, humid incubator at 34°C. Adults were removed from the comb within 24 h of emergence, marked on the thorax with paint or provided with tags as described below, and introduced to an experimental colony.

Each bee that had to be tagged was first anaesthetized on ice and kept on a container of shaved ice covered with aluminium foil. A vacuum pickup tool (Hakko Model 394, Osaka, Japan) was used to select and manipulate upright two tags into a small dish. The tags' serial numbers were read using a hand-held detector and recorded. A small dollop of Loctite Super Glue Gel Control (Henkel, Düsseldorf, Germany) was applied to the bee's thorax. Tags were picked up with the vacuum tool, a damp toothpick or a pair of microdissection forceps, and positioned on the bee's thorax. The bee was then placed in a cradle to stay upright for the minute or so required for glue drying and recovery. The entire tagging process for a single bee lasted 3–5 min.

Once the requisite number of bees was tagged, single-cohort colonies, composed of 1000–1500 bees were established according to standard procedures [98]. The small populations allowed us to tag and thus monitor a proportionally high number of the colony's foragers. Each colony was composed of a laying queen, about 1000 paint-marked workers bees, and 100–500 tagged worker bees. These were placed in a small Styrofoam hive box and given three honeycomb frames containing nectar and pollen, but no eggs, larvae or pupae.

colony	setup	perturbation	no. of bees (tagged/total)	mean detection rate (%)
O1	15 Jun 2011	none	102 / 1060	49 ± 16
O2	26 Aug 2011	none	119 / 1051	39 ± 14
O3	4 Jun 2012	none	369 / 1348	49 ± 13
E1	10 Sep 2011	forager removal	163 / 1084	37 ± 13
E2	22 Aug 2012	forager removal	391 / 1442	54 ± 13

Table 6.1: **Summary of the five different experiments performed** The detection rate was calculated for each bee as described in subsection 6.3.2; the values here are mean ± one standard deviation, calculated across bees (see Figure 6.4 for the complete distributions).

6.2.3 Summary of experiments

All activity-tracking experiments were conducted in Urbana, Illinois, in the summers (May–October) of 2011 and 2012. Specifics of each experimental colony are detailed in Table 6.1. Each experimental colony was maintained for 5–7 weeks, during which time the flight activity of tagged bees was continuously recorded.

Three colonies (O1, O2, O3) were located outside, with unrestricted access to the surrounding landscape. No experimental manipulations were performed on these colonies; their activity provided a baseline of foraging activity on naturally occurring resources.

Two additional colonies (E1, E2) were located inside an outdoor mesh enclosure (roughly 6 m wide, 20 m long, 3 m high) and provided *ad libitum* pollen, water, and 50% sucrose solution, placed in feeders positioned near the colony and replenished daily. In the fourth week after colony setup, hand-held laser detectors were used to manually detect and record the presence of tagged bees at both the sucrose and pollen feeder over a 5-hour period (09:00–14:00) for three consecutive days. On the second day of this recording period, all bees observed foraging at the feeders between 11:00 and 12:00 were killed. These removal perturbations were performed to study the response of the colony to the loss of a portion of the foraging workforce.

6.3 Data analysis

The raw data obtained from the two detectors consists of a series of tuples $(t,\ i,\ r)$, where t is a time stamp, i is a tag ID number, and $r \in \{\textsc{i},\ \textsc{o}\}$ represents either the inner detector I or the outer detector O. During the tagging process, a list is created to allow us to map $i \mapsto b$, where b is a bee ID number; note that there are two tags i for each bee b. After this mapping is performed, our data consists of a series of tuples $(t,\ b,\ r)$, which is used in all downstream analyses.

$r_{j-1},\ r_j$	s_{j-1}		
	0	1	2
I,I	(1, 0)	(0, 0)	(0, 1)
I,O	(1, 0)	(1, 1)	(0, 2)
O,I	(1, 1)	(1, 1)	(1, 0)
O,O	(2, 2)	(2, 0)	(1, 0)

$r_{j-1},\ r_j$	s_{j-1}		
	0	1	2
I,I	(2, 2)	(0, 0)	(0, 1)
I,O	(2, 1)	(0, 1)	(0, 2)
O,I	(2, 2)	(2, 2)	(0, 1)
O,O	(2, 2)	(2, 0)	(0, 2)

Table 6.2: **Lookup tables to infer a bee's state s_j.** Using the current detection event r_j, the previous detection r_{j-1} and previous state s_j, these tables are used to infer s_j. The table entries are tuples of form (s_j, flag). The table on the left is used for short time steps $\Delta t_j \leq 15$ s, while the one on the right is used for long time steps.

6.3.1 Inference of round-trip excursions

If we look at the data for a single bee, we see a time series of detection events, such as IOOIIIIOIO... (together with the corresponding times for each detection event). The reason we used two detectors I and O was that we hoped to infer when a round trip occurs: that is, a round trip starting inside the hive should consist of the subsequence IOOI, with a relatively long time span separating the two O events (while the bee is outdoors), and a relatively short time spent in the walkway between I and O, both while leaving the hive and while returning.

However, in reality, each detection event happens with some probability p_{det}, which in principle depends on the individual bee being detected, the particular detector in consideration, and the time at which it is detected: thus, $p_{det} = p_{det}(b, r, t)$. The dependence on the individual and on time may arise from such factors as the placement of the tags on that particular individual, the individual's propensity to move quickly past the detectors, how much dirt has accumulated on the detector photocell surface, etc. In the next section, we describe how we estimated p_{det}, ignoring the possible dependence on r and t; however, it should be clear that if p_{det} is anything other than close to unity, it becomes more challenging to translate a time series of detection events—such as IOOIIIIOIO—into a time series for round trips. Furthermore, it is possible for an irresolute bee to enter the walkway (setting off I), but then immediately turn back (setting off I again)! Here we describe the heuristics we used to make the inference of excursions, given these challenges.

We think of a bee as being in one of three states at any instant of time: inside the hive (state $s = 0$), in the walkway—i.e. in between the two detectors I and O—($s = 1$), or outside ($s = 2$). In the first step of the inference algorithm, for each bee b_0, a state $s_j \in \{0, 1, 2\}$ is assigned to each data point j in the time series $\{(t_j,\ b_0,\ r_j)\}_{j=1}^{N(b_0)}$ of detections for that bee. This assignment is made based on three values: the detector r_j that has been activated at the time under consideration; the previous detector r_{j-1} activated by this particular bee; and the prior state of the bee s_{j-1}. The assignment also depends on the time step $\Delta t_j \equiv t_j - t_{j-1}$.

Table 6.2 shows the two lookup tables used to make the inference. These tables provide the new state information s_j for the state of the bee in the time interval $[t_{j-1},\ t_j]$, as well as a 'flag' variable, which is a heuristic measure of the quality of that inference, as described below. The table on the left is used for short time intervals, i.e. $\Delta t_j \leq 15$ s, while the table on the right is used for $\Delta t_j > 15$ s. This threshold value of 15 s was chosen as it is roughly the expected upper bound on the time spent in the walkway (in between the two detectors).

To understand these tables, consider the following example: detector O is activated at the time under consideration; the previous activation recorded was detector I and the state recorded at that time was 0. Thus we need to look up row I,O and column $s_{j-1} = 0$ in order to figure out the state of the bee s_j. If Δt_j is short, then the first table tells us $s_j = 1$; the interpretation of this is that the bee was initially inside the hive ($s_{j-1} = 0$), it entered the walkway and activated detector $r_{j-1} =$ I, and then crossed the walkway and has now activated detector $r_{j-1} =$ O. Thus, it spent the time interval in the walkway, meaning that we assign $s_j = 1$, as the table tells us. Since this time interval is short, this seems a very reasonable inference, so the flag variable is set to 0.

Suppose instead that the time interval was longer than our chosen threshold value of 15 s. In that case, it is likely that the bee actually made a trip outdoors, but the outer detector failed to record the transition from the walkway to the outdoors. Thus the second table tells us that in this case, the bee must have spent most of its time outdoors—$s_j = 2$—but the quality of the inference is lower due to the missed detection (thus the flag is 1).

The entries of the table are created by thinking of the most likely scenario in this way, and recording the quality of each inference with a flag (0, 1 or 2—roughly, the number of missed detections). Once each detection was annotated in this way with a state s_j for the preceding time interval, we merged the walkway states $s_j = 1$ with the state $s_j = 0$. In this way, we go from a binary sequence of detections $r_j \in \{$I, O$\}$ to a binary sequences of states, with $s_j \in \{0, 2\}$.

Figure 6.3 shows the proportion of flags for each dataset. Because the proportion of high-quality inferences (flag 0) was lower than 20% in all cases, we deduced that a bee's probability of being detected by a single detector is likely not close to unity, and this confounds our ability to infer round-trip excursions with good accuracy. To deal with this, we first estimated the detection rate for each bee (see subsection 6.3.2) and then used a metric of activity that did not rely on the inference of round trips (subsection 6.3.3).

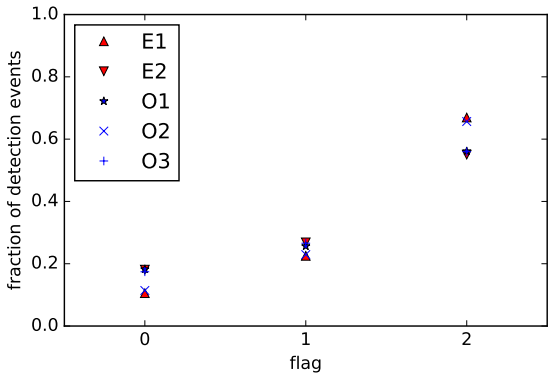

Figure 6.3: **Distribution of quality flags.** A flag of 0 represents a high-quality inference, while a flag of 2 is a poor-quality inference.

6.3.2 Detector success rate

As we discussed in the previous section, each detector r successfully detects a bee b with probability $p_{det} = p_{det}(b, r, t)$. Our goal in this section is to estimate this probability, making the simplifying assumptions that it is that same for the two detectors and that it does not depend on time: that is, $p_{det} = p_{det}(b)$.

Suppose that a particular bee passes the pair of detectors n times while making foraging trips, and that in n_2 of these events, both detectors are activated, in n_1 events a single detector is activated, and in n_0 events, the bee passes undetected. Then the expectation value for the fraction of events in which both detectors are activated is simply p_{det}^2, while the probability that only a single detector is activated is $2p_{det}(1 - p_{det})$.

In our data, we can count the number of times n_2 the two detectors are activated in quick succession (IO or OI, with a time separation of no more than 15 s). If we have N data points in total for that bee, then the number of single detector activations must be $n_1 \equiv N - 2n_2$. In addition, we have an unknown number of times n_0 in which both detectors failed to activate.

We can then estimate

$$2\hat{p}_{det}(1 - \hat{p}_{det}) = \frac{n_1}{n_0 + n_1 + n_2}, \qquad \hat{p}_{det}^2 = \frac{n_2}{n_0 + n_1 + n_2}, \tag{6.1}$$

and eliminating the unmeasurable n_0, we obtain

$$\hat{p}_{det} = \frac{2n_2}{n_1 + 2n_2} = \frac{2n_2}{N}. \tag{6.2}$$

In this way, we estimate $p_{det}(b)$ for each bee b. Figure 6.4 shows the distribution of $\hat{p}_{det}(b)$ across different bees and Table 6.1 includes the mean detection rate, for each dataset. We see that the mean detection rate

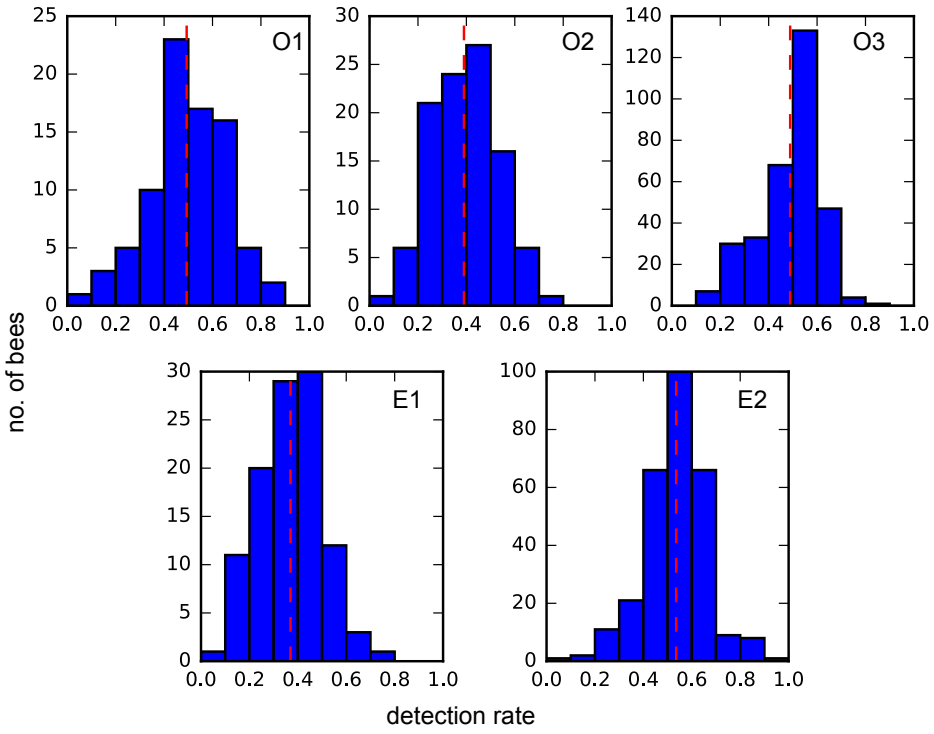

Figure 6.4: **Distribution of detection rate across bees.** Histograms using ten bins spaced evenly on $[0, 1]$; the dashed red line indicates the mean value.

is around 35–55%, and there is quite a lot of variance about this value (standard deviation around 15%).

Figure 6.5 shows the distribution across bees of 'single' and 'double' reads (represented by n_1 and n_2 above). We can see from this that the distribution of reads is generally highly skewed: there is a small number of bees with a relatively high number of reads. In the next two sections, we describe some further steps needed before we can translate this statement into one about the foraging activity of different bees.

6.3.3 Measuring daily activity

The relatively low mean value of the probability of detection $p_{det}(b)$ makes it difficult to infer round-trip excursions of bees, so we instead used the number of detection events, or reads, as a metric for foraging activity. To account for the wide variation in $p_{det}(b)$ across bees, we corrected each bee's daily number of reads by the estimated value $\hat{p}_{det}(b)$ for that bee. Bees with $\hat{p}_{det}(b) < 0.2$ were discarded from downstream analyses.

Figure 6.6 shows the total colony activity, as measured by the total number of reads (corrected for detection rate) on a particular day. We see that aside from the general trend that activity is lower during the first few days (when bees are orienting), the day-to-day fluctuations are very large. Days with close to

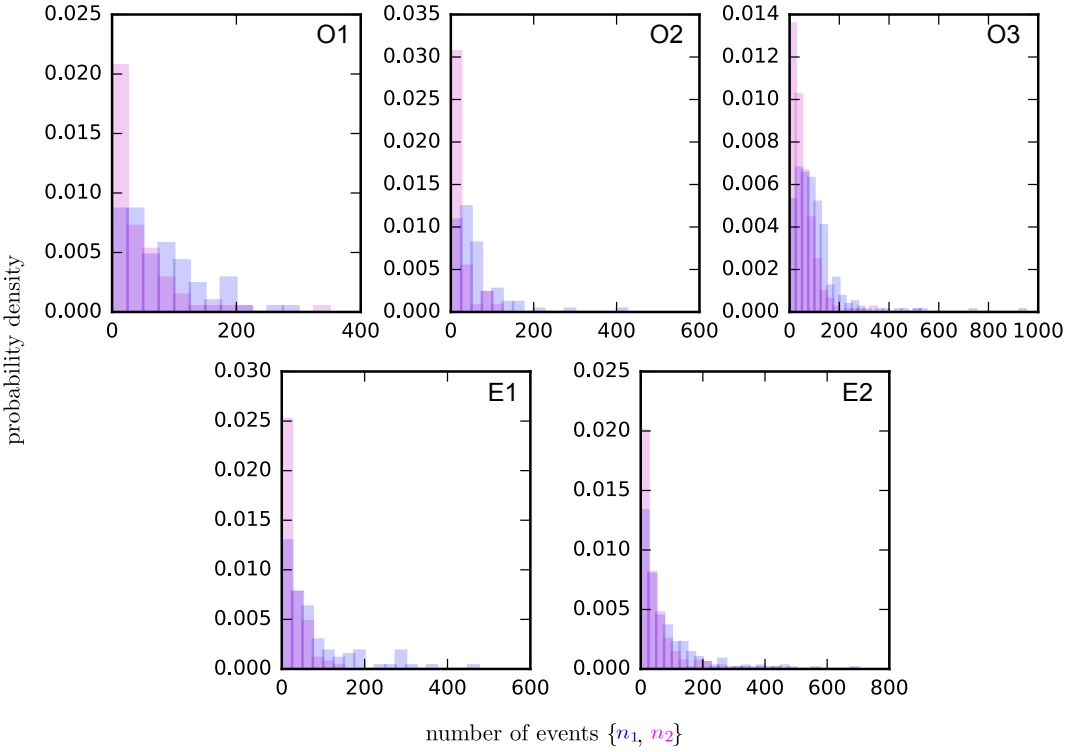

Figure 6.5: **Distribution of event counts across bees.** These plots estimate the probability distribution for the number of single reads n_1 (blue) and the number of double reads n_2 (magenta), for each dataset. The bin width is 25 reads.

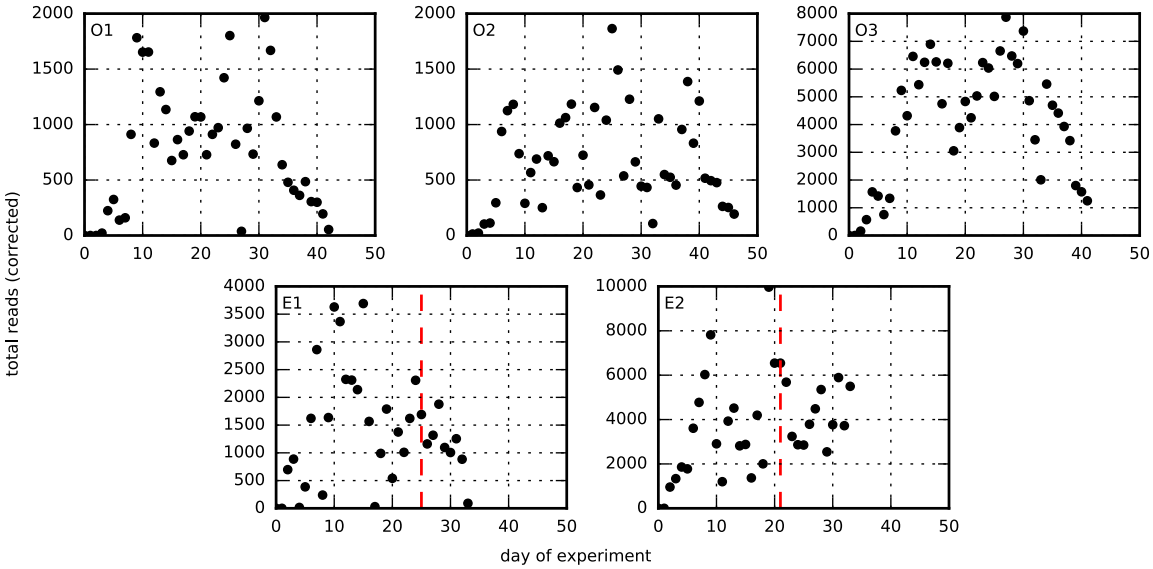

Figure 6.6: **Time series of total colony activity.** Activity is measured as the sum total of the reads for each bee on a given day, corrected by its detection rate. The red dashed line for colonies E1 and E2 indicates the day on which the perturbation took place.

zero activity are accounted for by weather: on rainy days, there is no foraging activity. Aside from that, it is not clear what controls the fluctuations in activity, or the size in the fluctuations.

In order to correct for colony-wide variation in activity by day, we normalized the number of reads for a bee on a given day by the total number of reads on that day. This allowed us to compare activity across different days of each experiment, despite variations in a colony's overall flight activity caused by weather, availability of forage, and other unknown factors. Thus the activity of a bee on a given day is effectively the fraction of reads that it contributes to the total reads for the colony on that day.

The data show almost no correlation between the reader success rate of an individual, and the activity level of that bee. This indicates that even when uncorrected for missed detections, an individual bee's activity level is not an artefact of tag placement or other factors affecting detection rate.

A bee's day of death was defined as the last day on which reads were detected for either of her tags. Activity measurements before the age at which foraging behaviour began (see subsection 6.3.4) and after death were not included in further analyses.

6.3.4 Detection of orientation flights

To focus our analyses on foraging activity, it was necessary to define and omit flight activity likely to be related to orientation flights, which occur for several days prior to the onset of foraging [16]. Based on personal observations, we knew that in our locality, most orientation flights occur in the afternoon, and so we defined a bee's first day of foraging to be the first day on which it has at least six reads, more than 25% of which occur before 12:00. This heuristic generated estimates of age at first foraging that agreed closely with those produced by a human observer. Any flight activity on or after this day was defined as foraging activity; corrected and normalized daily reads, as defined in subsection 6.3.3, were used as a proxy measure of foraging activity.

6.4 Results

6.4.1 Orientation flights and onset of foraging

We first confirmed that the flight behaviour detected by the PharmaSeq system fit with established norms of honeybee foraging behaviour by examining the age at onset of foraging in each experimental colony, using the method described in subsection 6.3.4 to detect onset of foraging. Because our colonies were composed initially from one-day-old worker bees, we expected to observe precocious foraging behaviour [59].

Figure 6.7 shows the distribution of age at onset of foraging for all five experimental colonies. We do

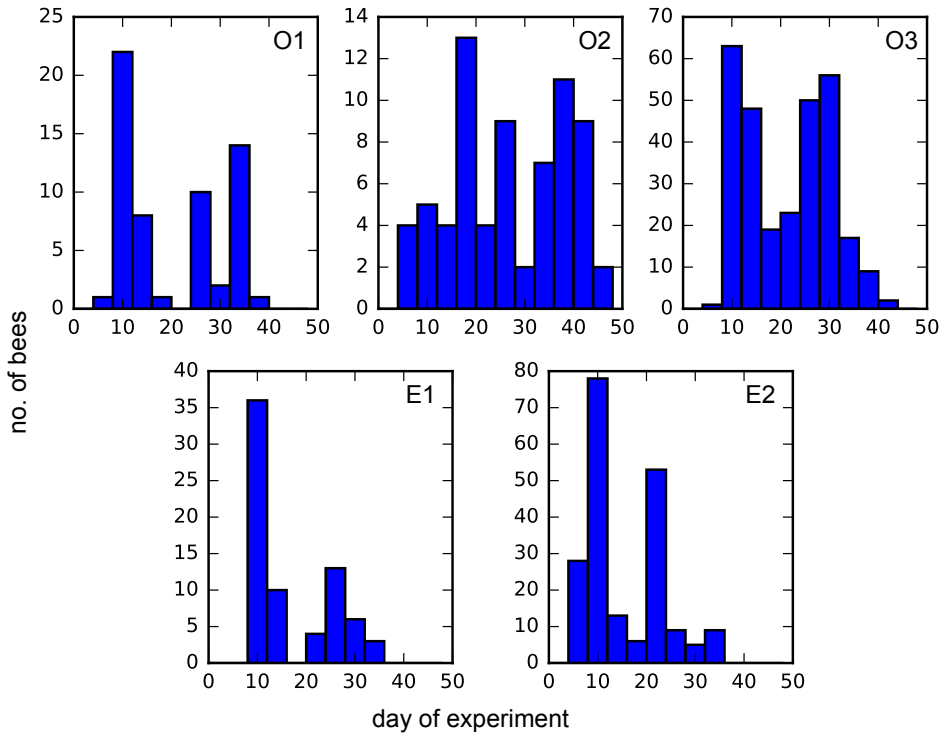

Figure 6.7: **Distribution of age at onset of foraging.** These histograms reveal bimodal distributions, with precocious foragers appearing around 10 days into the experiment, and a second cohort peaking 2–3 weeks later. The vertical axis gives counts in bins of width 4 days.

indeed see that there are two peaks in this distribution: the first, around 10 days, corresponds to the emergence of precocious foragers, while the second at 3–5 weeks corresponds to foragers of age typical of regular colonies. Interestingly, the two colonies in enclosures that were provided with an infinite food supply appear to have a particularly sharp first peak, suggesting perhaps that the number of precocious foragers that participate in foraging activity is modulated by the availability of forage.

Unlike most behavioural maturation experiments in bees, which rely on human observation of flight activity, we were able to use the monitoring technology to collect data from tagged bees continuously over several weeks to build a profile of the age of onset of foraging. This profile is consistent with previous, less complete reports of age at first foraging based on human observation [58, 133], indicating that automated monitoring system did not produce observable disruption of normal fight behaviour, and that the recordings produced by the system could be used to accurately infer flight behaviour.

6.4.2 Statistics of lifetime foraging activity

In order to quantify the distribution of activity across bees in the colony, we use the definition mentioned in subsection 6.3.3: that is, a bee's daily activity is her contribution of detectors reads to the total reads of the colony on that day, correcting for errors in detection rate. To study an individual's activity over her lifetime, we measured both her mean daily activity (ignoring days on which she had no activity, such as after her death) and her total activity (i.e. her daily activity summed over all days following orientation). The former is generally preferable because it accounts for the fact that bees have different lifespans. We also excluded individuals with detection rate less than 20%; for such individuals, their likelihood of not being detected at all during a round trip (i.e. passing each detector twice) is rather high: it is at least $(1 - 0.2)^4 = 41\%$.

Figure 6.8 shows the distribution of mean daily activity across bees for each dataset. We see a distribution with a long tail, in which a small number of bees have a very high activity compared to the others. The distribution of activity is continuous rather than, say, bimodal, indicating that there is a continuum of activity levels among individuals within the society. Note that it is difficult to compare mean activity levels across datasets, because in its definition, the normalization of daily activity implicitly depends on the number of tagged foragers, and this is different for each dataset—this is presumably why the large colonies O3 and E1 have a smaller range in the horizontal axis of Figure 6.8 compared with the other three, smaller colonies.

To quantify the degree of inequality in this distribution, we plotted Lorenz curves based on bees' mean daily activity (Figure 6.9. Each such curve displays the share of foraging activity (y-axis) contributed by the bottom x-percentage of foragers in the colony. Thus, a perfectly equitable distribution, in which all foragers have the same activity, corresponds to the line $y = x$, and the upward concavity relative to this

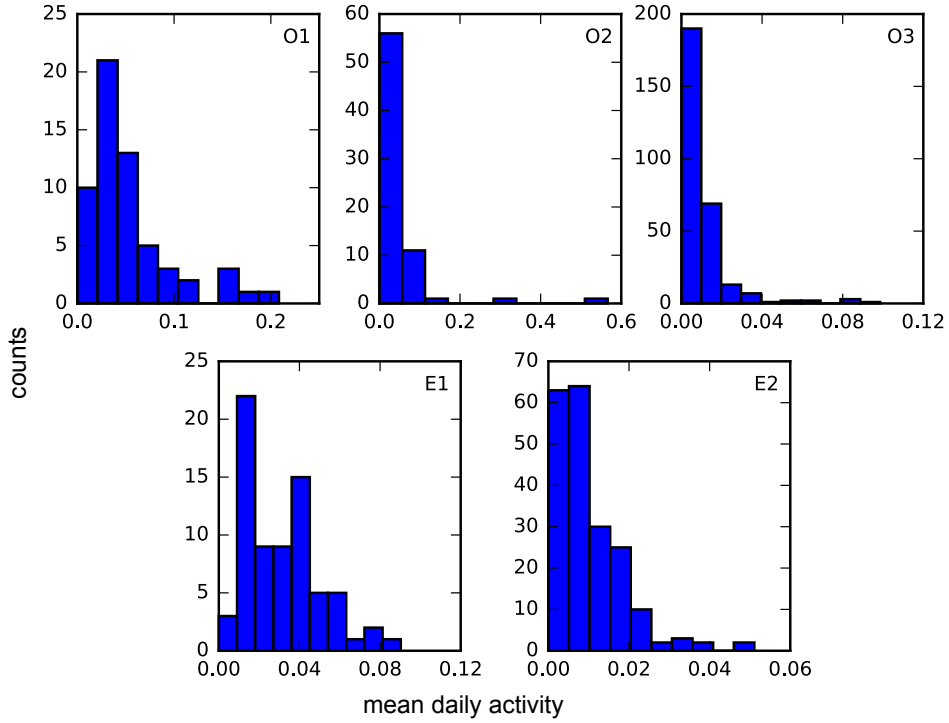

Figure 6.8: **Distribution of mean foraging activity.** Histograms for mean foraging activity across bees for each dataset. We used 10 bins spaced evenly over the range of activity values.

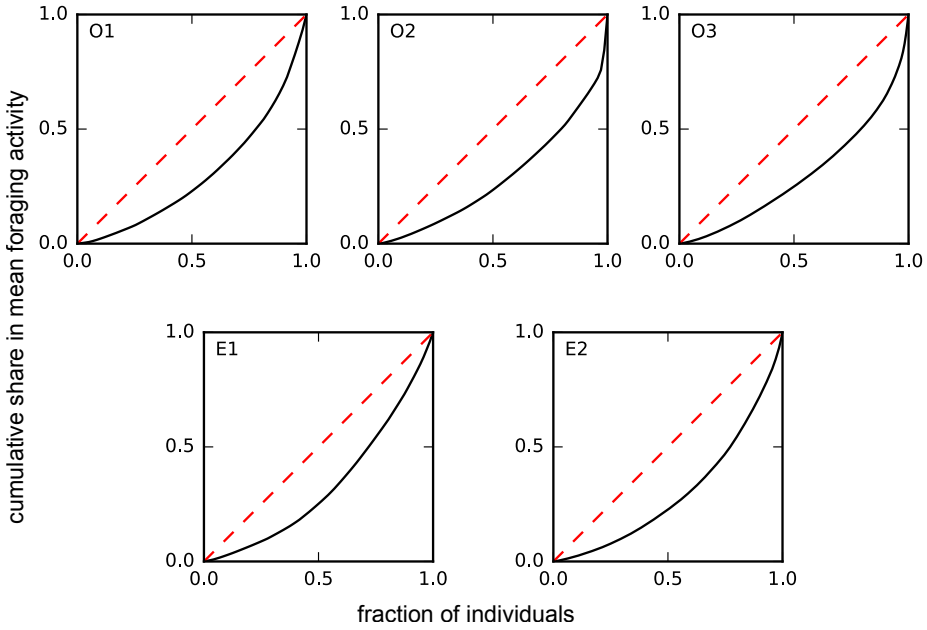

Figure 6.9: **Lorenz curves for bee activity.** The diagonal line $y = x$ (in red) corresponds to a completely equal distribution of activity. The Gini coefficient is equal to twice the area in between the red dashed line and the solid curve.

dataset	Gini coefficient	
	mean daily activity	total daily activity
E1	0.336	0.488
E2	0.398	0.541
O1	0.403	0.501
O2	0.425	0.479
O3	0.407	0.454

country	Gini coefficient	date
Brazil	0.519	2012
India	0.336	2012
South Africa	0.625	2013
Sweden	0.249	2013
USA	0.450	2007

Table 6.3: **Gini coefficients for bee and human societies.** For the bee datasets, the Gini coefficient is calculated for the distribution of both the mean daily activity and total daily activity. For comparison, the table on the right lists the Gini coefficient for the distribution of family income in five countries around the world, together with the date of the estimate, using data from *The World Factbook 2013-14* published by the US Central Intelligence Agency at https://www.cia.gov/library/publications/the-world-factbook/index.html.

provides a measure of the degree of inequality. Specifically, the Gini coefficient, which is proportional to the area between the Lorenz curve and the line $y = x$, is one common way of measuring the degree of inequality represented by the distribution. In Table 6.3, we list the Gini coefficient for each dataset, based on both the distribution of mean daily activity (Figure 6.8 and Figure 6.9), and the distribution of total daily activity.

6.4.3 Variation in foraging activity over lifetime

The previous section quantitatively established inequalities in foraging activity among members of a honeybee colony by looking at the mean (and total) daily activity. The data also allow us to address the question of how a forager's activity changes over time. Behavioural parameters such as foraging efficiency and foraging strategy of honeybee workers are known to vary over their lifespan [13, 102]. Our data allow us to explore if there are any consistent ways in which a worker's activity changes over time.

In this section, we will display time series for each individual bees in the form of a heatmap, and here we briefly describe how we chose to arrange the rows of the heatmap. Suppose the time series is for the daily activity $A(b, t)$ for each bee b; the data can be represented as a matrix A_{ij} with i mapping to a bee b and j indexing the day of the experiment. Let $\vec{a}_i$ represent row i of matrix $\boldsymbol{A}$, i.e., $\vec{a}_i$ is the discrete daily activity time series for the bee corresponding to index i. In order to help discern patterns, our goal will be to sort the set of vectors $\{\vec{a}_i \ : \ i = 1, 2, \ldots, N_b\}$ by similarity. The heuristic algorithm we use works as follows: we first define $\vec{a}_0$ to be simply a vector of zeros. Then we fix $i_0 = 0$, find $i_s \equiv \text{argmin}\,\{\|\vec{a}_{i_0} - \vec{a}_i\|_1 : \ i > i_0\}$, and then swap rows (i_0+1) and i_s in the matrix $\boldsymbol{A}$. We then repeat this process for each $i_0 \in \{1, 2, \ldots, (N_b-1)\}$. This algorithm ensures that similar rows are placed next to each other, making it easier to discern groups of bees with similar activity patterns, and to discern patterns within these groups.

In Figure 6.10, we plot the time series of daily activity for each bee after its orientation period, in the

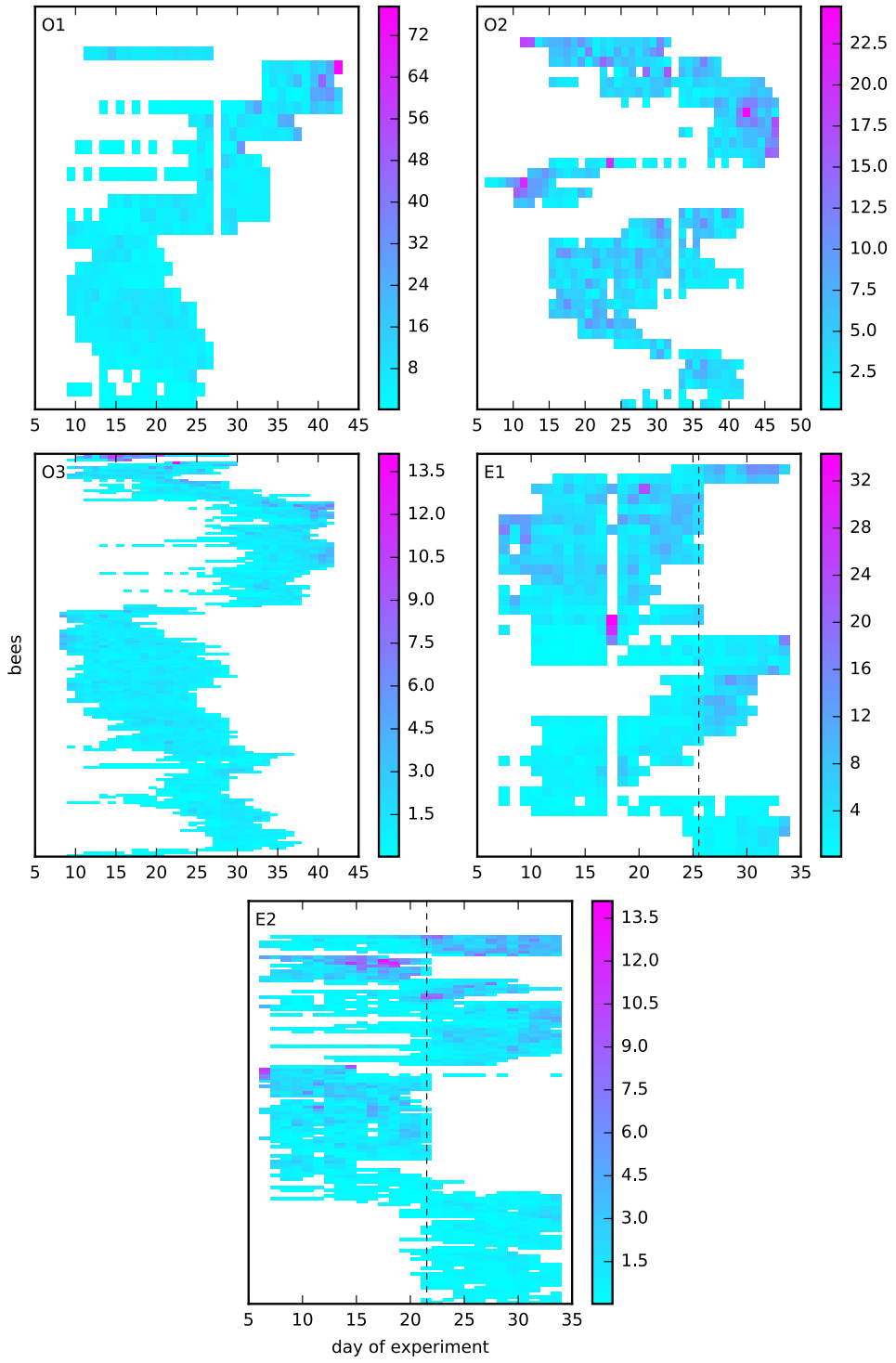

Figure 6.10: **Time series of daily activity for each bee.** Each row corresponds to a single bee, and the colour represents the percentage of reads contributed by that bee to the total reads of the colony on a given day. The rows are sorted roughly by similarity, and the forager-removal days for colonies E1 and E2 are marked by a dashed line.

form of a heatmap. The colour represents the percentage of reads contributed by a particular bee to the colony's total reads on a given day. Days of zero activity are in white—these occur because the bee did not leave the hive on that day, the bee has not completed orientation yet, or because the bee has died. Colonies O1, O2 and E1 have days with no almost activity (i.e. predominantly white vertical stripes): these were due to rain, which prevented the bees from wanting to leave the hive. These time series also provide an explanation for the distribution of age at onset of foraging, discussed in subsection 6.4.1. Precocious foragers show high activity early in the experiment (roughly days 10–25) and then die; the second cohort of foragers—the second peak in the histograms of Figure 6.7—then takes over.

In order to look for patterns in activity over a bee's lifespan, we also plotted the deviation of a bee's daily activity relative to the mean over its lifetime: that is, if bee b has activity $A(t, b)$ on day t, we plot

$$\text{relative fluctuation in daily activity} \equiv \frac{A(t,b) - \langle A(t,b)\rangle_t}{\langle A(t,b)\rangle_t}, \tag{6.3}$$

where the average $\langle \cdot \rangle_t$ is taken over all days t with non-zero activity for bee b. Figure 6.11 shows the time series of this quantity for each bee in the form of a heatmap. The two different colours represent positive (red) and negative (blue) deviations, and the shade indicates the magnitude.

The heat maps suggest that rather than a single shared pattern in activity level over lifespan, individual bees in fact exhibit a diversity of activity levels over their lifespan. For some individuals (e.g. those actively foraging on days 10–25 in colony O3) the daily activity remains fairly constant, with relatively small fluctuations up or down. For other, particularly older individuals, activity tends to increase in time. In particular, the individuals foraging after day 32 in colony O3 show this trend, as well as the foragers in colonies E1 and E2 that begin foraging after the perturbation. This is partly because of mortality: by the end of the experiment, there are fewer foragers still alive, and so an individual forager's contribution to the colony's total activity becomes larger. It is also interesting to note that every colony shows at least one large positive fluctuation (say > 3), but negative fluctuations are always relatively small in magnitude.

6.4.4 Colony resilience and plasticity in foraging

Recall that in the enclosure-housed colonies E1 and E2, removal experiments were conducted on days 26 and 22 respectively. In these experiments, all bees—both tagged and untagged—that arrived at either the pollen or the nectar feeder we captured and killed during a 1-hour period. Although this perturbation does not specifically target any particular type of forager, it is more likely that a high-activity forager is removed in this perturbation, simple because of their increased likelihood of being at the feed at that time. It was

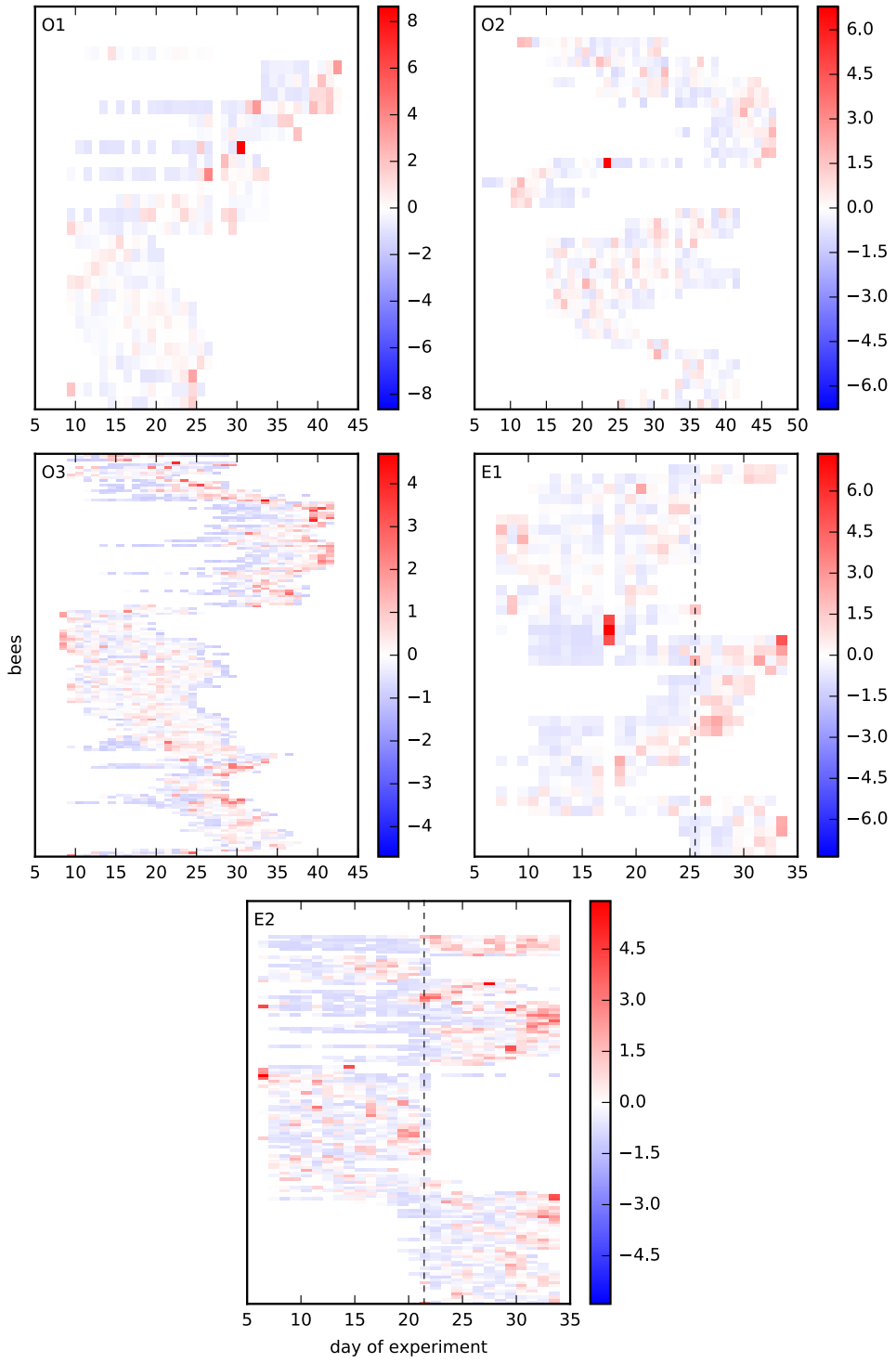

Figure 6.11: **Time series of fluctuations in activity for each bee.** Each row corresponds to a single bee, and the colour represents the fluctuation in daily activity relative to the mean for a given bee on a given day. The rows are sorted roughly by similarity, and the forager-removal days for colonies E1 and E2 are marked by a dashed line.

essential to target both tagged and untagged bees to ensure that that colony's workforce was sufficiently effected.

From the time of removal to the end of that day's observation period, the number of visits to both the pollen and nectar feeders was negligible (fewer than ten visits were observed experimentally). However, by the next morning, foraging activity and visits to the feeders had recovered, as seen in Figure 6.6. Moreover, Figure 6.10 shows that existing foragers adjust their activity levels to compensate for the loss of those removed. The number of new foragers observed the day after the removal (4 in E1, 6 in E2) was not much higher that on the few days before removal.

Overall these measurements suggest that activity levels do not remain constant across the foraging lifespan, and "elite" performance is at least partially environmentally determined. In other words, individuals may be able to adapt their activity level in response to external environmental factors, such as an increase or decrease in the profitability of a food source to which an individual forager is loyal [121], or the appearance or disappearance of a resource exploited by a subset of the foraging population [124]. It is also clear that internal factors such as changes in colony demography—such as the loss of foragers—also affect individual activity levels.

6.5 Discussion

In summary: we used RFID tagging technology to automatically track flight activity for a large number of individuals over their entire lifespans in five different colonies. The extensive dataset generated by this automatic monitoring technology revealed a skewed distribution in foraging activity across individuals, as measured both by their average flight activity and by their total activity over their lifetime. On a given day, there is a small number of bees with high activity, resembling so-called "elite" workers that have been reported in a number of other species. However, the distribution of activity across bees show that here is a continuum of activity levels, and elite workers are not necessarily a distinct subcaste of foragers, but simply those at the extreme of a range of activity levels. We quantified the skew in the distribution of activity using the framework of Lorenz curves and Gini coefficients.

Our data also allowed us to look at the temporal structure of a bee's activity over time, and the individual-level time series of activity indicate that a bee's activity level is flexibly adjusted over its lifetime, perhaps responding to environmental cues. Three aspects of our results support this idea of flexibility. First, the fact that the distribution of activity levels is continuous, rather than (for example) bimodal. Second, bees in our study showed a variety of activity level patterns throughout their lives, indicating that there is no set

developmental progression of foraging activity level; this suggests that individuals may adjust their foraging activity levels according to the instantaneous individual experience of social and environmental conditions. Third, as described in the final section of the results, removal of a fraction of the foraging population was associated with an almost five-fold increase in the activity level of previously low-activity foragers. Taken together, these data support the view that individual workers continuously adjust their activity level to ensure that colony nutritional needs are being adequately and efficiently met, and that the net activity of the whole foraging population is likely to be one of the factors that influences this decision.

Part III

Interaction networks and power laws in honeybee societies

Chapter 7

High-resolution spatiotemporal tracking of an entire society

Part III consists of this and the next two chapters. In this chapter, we introduce a system that allows us to track all individuals in a honeybee society, which forms the basis for this part of the dissertation. Using the data generated by this system, we can detect particular behaviours taking place inside the hive, and measure the activities of all individuals. We study the timeseries of these activities, and try to make predictions about the social caste of an individual on the basis of her activities.

In chapter 8 and chapter 9, we study the social networks generated by one particular behaviour, namely trophallaxis. We study the statistics of time-aggregated networks, the statistics of trophallaxis interactions, and probe the temporal structure of the interactions using the framework of temporal networks.

7.1 Introduction

In the late eighteenth century, the English social theorist Jeremy Bentham conceived of the *Panopticon*, a prison building designed such that all the inmates could be observed without their knowledge by a single watchman [10].[1] Bentham saw it as as means to affect the moral reformation of criminals, a "new mode of obtaining power of mind over mind, in a quantity hitherto without example: and that, to a degree equally without example." Although the Panopticon's architectural design was commissioned and completed, it ended up never being built due to various bureaucratic setbacks. This vision of constant and omnipresent surveillance of a section of society offends our modern sensibilities, but one wonders what the observer might have seen. Like any set of humans interacting socially in some fixed context, prisoner communities have a complex social structure—with features such as the emergence of a currency—that have long been studied by sociologists and social psychologists.

Today, in the era of so-called Big Data, tracking human behaviours and their social context is easier than ever before, and can occur at unprecedented scales, and sometimes without our knowledge. As a simple example, anyone who owns an Android smartphone today and has enabled Google's "location services"

[1] Thanks to Nigel Goldenfeld for pointing out the similarity, in abstraction, between the Panopticon and the bee tracking apparatus described in section 7.2.

feature ends up uploading their every movement continuously in time to Google's servers; they can then view their life's spatial trajectory as a timeline on Google Maps. Google presumably has these data for a substantial part of the US—indeed, world—population, and can study not just individual trajectories but also their interactions. In this way, companies such as Google and Facebook have access to sociological data on large human populations at spatiotemporal scales that previous generations of social scientists could only dream about.

The availability of Big Data has the potential not just to improve scientific understanding, but also to change—or extend—the traditional, hypothesis-driven method of doing science. As Larry Smarr has pointed out, "In a data-poor world, you have to come up with a topic you can focus enough to get a little bit of data, but if you're in a data-rich world, the data generates its own hypotheses, which then have to be verified." This is quoted in an article [107] about the planned Kavli Human Understanding through Measurement and Analysis (HUMAN) Project, which will track the lives of 2500 volunteer households in New York City, measuring different aspects of individual lifestyles to try to discover connections with health and behaviour, integrating both individual and environmental data using the city's urban infrastructure.[2]

In this chapter, we introduce a project to track every individual in a fully-functional animal society, namely a honeybee colony. We begin with a description of the experimental setup, and then move on to deriving individual behaviours from the tracking data. We also study the time variation of activity in the society. The goal of this chapter is to provide a descriptive, qualitative understanding of the dynamics and interactions that we can track in the honeybee colony. In chapter 8 and chapter 9, we will study interaction networks and their statistics in more depth.

7.2 Experimental design

7.2.1 Tagging and colony setup

Bees were obtained from source colonies maintained according to standard techniques at the University of Illinois Bee Research Facility. One-day-old adult worker bees were obtained by removing honeycomb brood frames containing sealed pupae from the source colonies, placing them in a Styrofoam box, and transferring them to an incubator where they were kept in specialized emergence cages at 34°C and 50% relative humidity.

Colonies were established with 1000–1200 one-day-old worker bees and one unrelated, naturally mated queen. All bees in the colony were tagged with a custom matrix barcode, called the "bCode," to enable a computer to automatically identify bees and measure their position and orientation (see Figure 7.1). The

[2]The author is grateful to Yoshi Oono for bringing the Kavli HUMAN project to his attention.

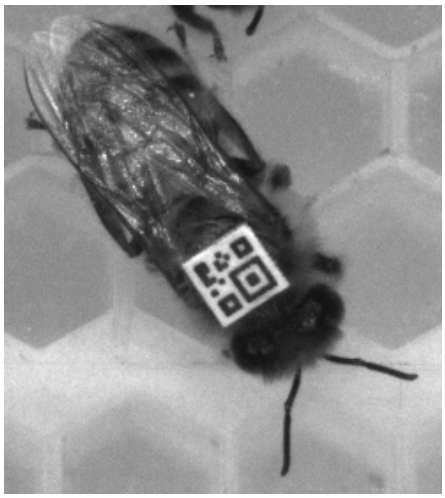

Figure 7.1: **bCodes.** A bee tagged with a bCode that uniquely identifies it.

bCode consists of a identifier pattern overlaid on a template pattern; the former is used to uniquely identify the bee, while the latter allows for detection of the bCode in an image and also provides information about the orientation of the tag (and thus of the bee) in an image. The template pattern consists of a "finder pattern" on the top-left, and two smaller "alignment patterns" on the top-right and bottom-left of each bCode. Like QR codes, bCodes incorporate Reed-Solomon error correction to ensure that the code is read correctly, even if a portion is damaged, dirty or otherwise obscured.

Each colony was provided with the same amount of honey and artificial "bee bread" (a mixture of 10 parts pollen, 9 parts honey, and 1 part water). Workers were barcoded in two groups of approximately 700 bees each: the first group emerged over 9 h at night prior to the day the colony was set up, while the second emerged over 9 h the next morning. During barcoding, each bee was first anaesthetized on ice. She was then picked up with a pair of blunt featherweight forceps (BioQuip Products, Rancho Dominguez, California, USA), and a small drop of Loctite Super Glue Gel Control (Henkel, Düsseldorf, Germany) was applied to her thorax. Next, a bCode was picked up with a moistened Plain Wood Applicator (Fisher Scientific, Waltham, Massachusetts, USA) and centered on the thorax. The bCode was positioned so that its orientation vector was parallel to the anteroposterior axis of the bee. The entire barcoding procedure for a single bee lasted 1–2 min, and marked bees were placed individually in a petri dish until they recovered.

Both groups of barcoded bees were simultaneously moved into the observation hive. Next, the queen was barcoded and added to the colony. The hive was then placed into the observation hive room and temporarily warmed to about 37°C to encourage the bees to spread out. Bees were allowed to acclimate until sunrise on the next day, at which point the workers were, on average, 24 h old. At sunrise we started recording images; in some cases, the recording began the previous night.

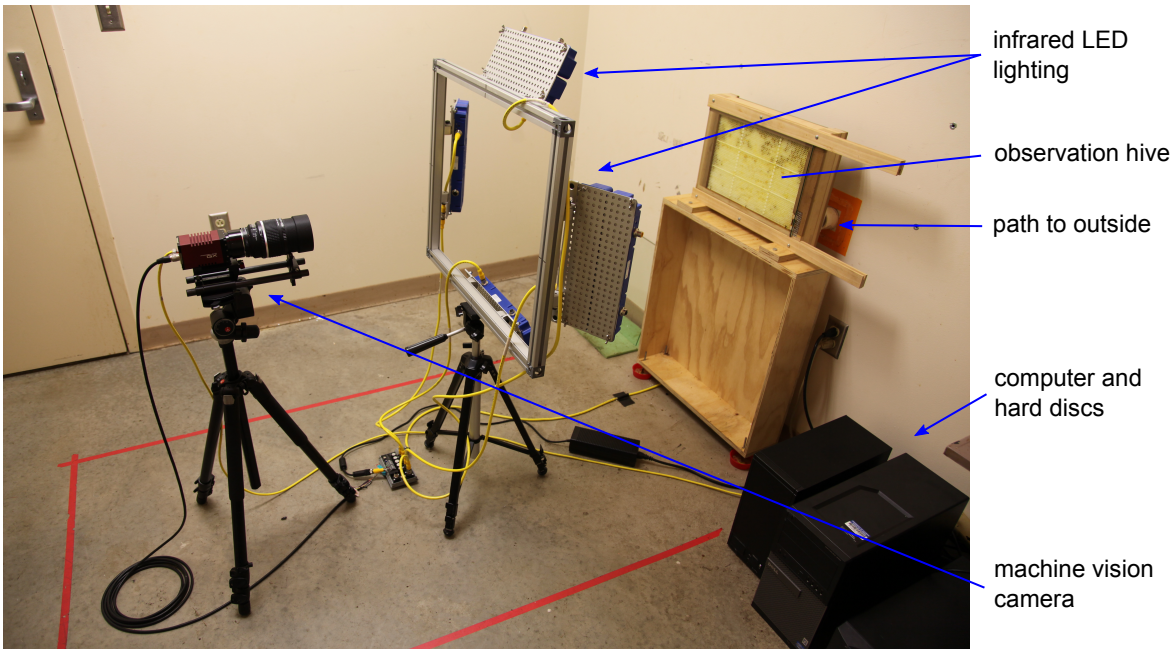

Figure 7.2: **Experimental setup.** The bee colony is set up in a glass-walled hive, which is illuminated by infrared LED lights mounted on a frame and imaged using a high-resolution machine vision camera at 1 Hz. Photograph courtesy of Tim Gernat.

7.2.2 Observation room and imaging

In each experiment, bees were housed in an observation hive with a single glass wall. The hive frame was made of wood in pieces that fit tightly around a white plastic honeycomb. The honeycomb was covered by a glass window; the distance between the honeycomb and glass was 8 mm, ensuring that bees would not crawl over and obscure each other's tags. The window was attached with a set of clamps and rails such that a dirty window could be exchanged daily with a clean one, with minimal disturbance to the bees. The lower-right corner of the hive frame was cut out to allow bees to move from the honeycomb onto a small ramp that led to an entrance tunnel that passed through the hive frame. This entrance tunnel passes through the room wall and into the outdoors, giving bees the ability to forage outside (see Figure 7.2).

The observation hive was maintained in a dedicated room in the University of Illinois Bee Research Facility. The room was kept completely isolated from visible light during the experiments, and the observation hive was illuminated only by 850 nm infrared light (L300-850W linear lights from Smart Vision Lights, Muskegon, Michigan, USA), which is invisible to honeybees. These lights were mounted on a frame positioned in front of the hive, and to further increase the contrast between the honeycomb and the bees, the back of the honeycomb was also illuminated with infrared light. The room was maintained at a temperature of about 32°C using a space heater.

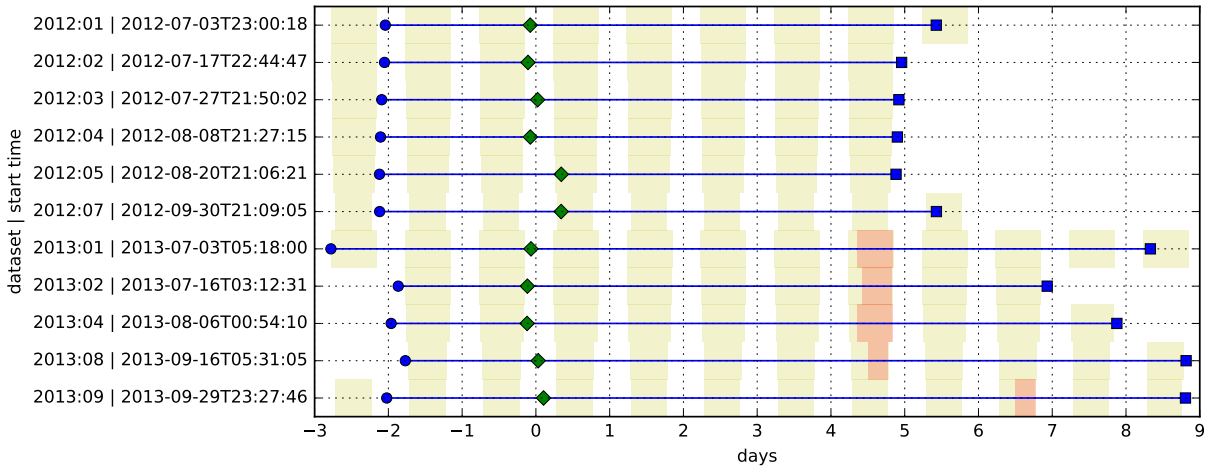

Figure 7.3: **Summary of datasets.** Each row corresponds to one dataset; the vertical label includes the time at which images began to be recorded, and this is indicated on the timeline by a blue ball. The time at which the hive entrance was first opened is indicated by a green diamond; the end of the experiment is marked by a blue square. The yellow shading indicates periods of daylight, i.e. sunrise to sunset each day, while red shading indicates periods of forager removal. The timelines here are aligned (at day 0) to midnight closest to the time at which the entrance was opened (green diamonds).

Images were recorded using a monochrome Prosilica GX6600 machine vision camera (Allied Vision, Exton, PA, USA), fitted with a Nikkor AF 135mm f/2 D DC prime lens (Nikon, Chiyoda, Tokyo, Japan). The lens was covered by a B+W 72mm IR Dark Red (092M) filter (Schneider Optics, Van Nuys, California, USA), which blocks visible light and thus ensured that images were not overexposed when the deep-red overhead light was turned on during periods of manual observation. The camera was mounted on a tripod and placed 1.3 m in front of the hive so that the entire honeycomb was visible. At the camera's resolution of 6576×4384 pixels, bCodes had a side length of approximately 38 pixels.

7.2.3 Summary of experiments

Figure 7.3 shows the timelines for six preliminary experiments performed in 2012 (lasting a week) plus five experiments performed in 2013 (lasting up to 11 days). Each year's set of experiments followed the same general design, with slight modifications due to weather conditions or improvements to the experimental protocols. The hive entrance was initially kept closed for 2–3 days to ensure that the bees did not attempt to fly out before being mature enough to do so. The timelines in Figure 7.3 are aligned so that day 0 corresponds to midnight closest to the time at which the entrance was first opened.

In the five experiments from 2013, a perturbation was performed on the fifth day after the hive entrance was opened. (For the dataset 2013:09, this was delayed by 2 days due to rain.) As many foragers as

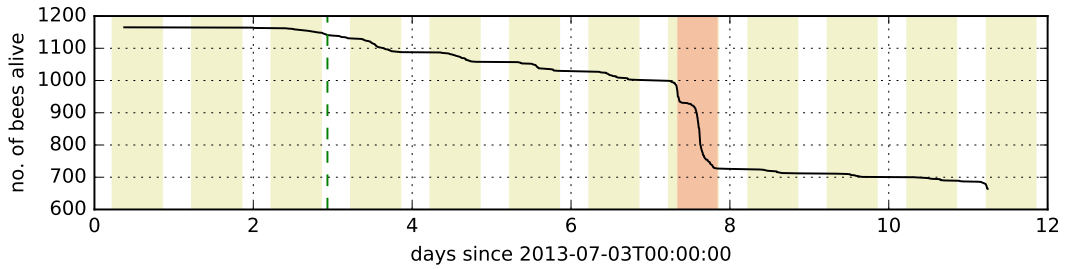

Figure 7.4: **Colony size vs. time.** In dataset 2013:01, the forager removal perturbation occurred on the eighth day, resulting in the loss of a large fraction (about 27%) of the colony.

possible were removed from the hive, by collecting bees returning from a foraging trip using a portable vacuum cleaner. If certain bees were observed to perform orientation flights—which occur before the onset of foraging in those bees—this collection process was delayed to allow them to finish. This was a significant perturbation to the colony demographic, as seen in Figure 7.4 for dataset 2013:01.

7.2.4 Image data processing

Figure 7.5 shows a sample image collected in the experiment 2012:01. Each such image had to be processed to try to recover the position and orientation of each bCode visible in the image. The images were first preprocessed by resizing to 140% of their original resolution and sharpening by digital unsharp masking. Each image was then converted to binary form by setting all pixels below a certain threshold value b to black, and all others to white. The pixels were then scanned, first to detect the finder pattern, and then to find alignment patterns in the vicinity of the potential finder-pattern locations. In this way, candidate bCode images were isolated in the complete image, and the identifier pattern could be read. The procedure was repeated for several empirically-determined thresholds b to provide a consensus identifier for each detected bCode. Further filtering step were performed to correct for errors in tracking:

1. All records of bCodes in which the identifier could not be correctly decoded were deleted.

2. Records of bCodes that were too small or too big were removed.

3. Records of bCodes that occured twice per time point were removed.

4. Records of bCodes that moved faster than 5 cm/s in successive time points were filtered out, as this far exceeds the typically-observed walking speed of a bee, estimated to be around 3 cm/s.

5. Records of bCodes after a bee's time of death were removed; this was necessary because dead bees may fall to the hive floor with their bCode facing the camera.

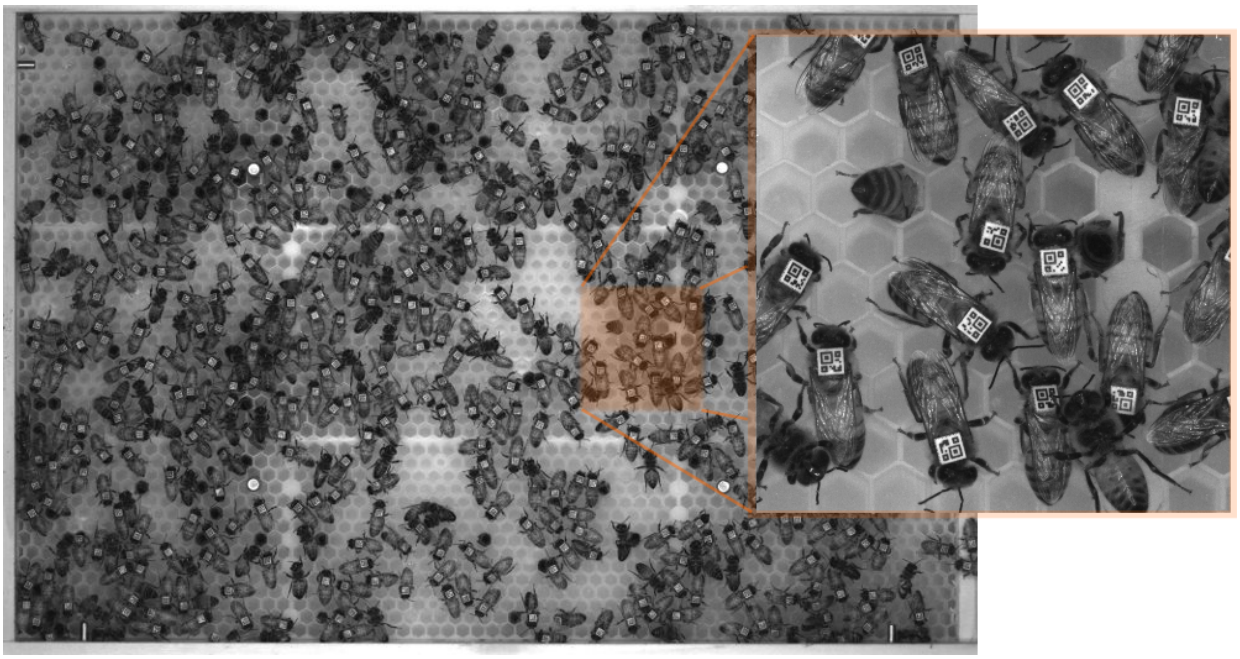

Figure 7.5: **Sample image of hive.** This image comes from dataset 2012:01. In each experiment, such images were acquired at a rate of 1 Hz. As the zoomed-in cut-out shows, the resolution of the image is high enough to distinguish and decode individual bCodes. Notice that bees are not always positioned so that their bCodes face the camera: they sometimes walk on the glass rather than the honeycomb and sometimes also go inside a honeycomb cell.

6. Finally, we deleted all bCodes that were outliers in their detection rate, as measured by the interquartile range rule.

These filtering steps resulted in the removal of $5.9\% \pm 0.9\%$ of records across the 2013 datasets.

Figure 7.6 shows the number of bCodes decoded in each image as a function of time. Because bCodes were not always successfully decoded from the images, and because in some cases (see Figure 7.5) bees were not displaying their bCodes in the direction of the camera, the number of bCodes detected is smaller than the actual number of bees present in the colony. Additionally, after the entrance was opened, bees could make foraging trips outside the hive, in which case they naturally would not show up in the images. In summary, the data extracted from the images at each timepoint t consists of a set of positions (x, y) and orientations $(\Delta x, \Delta y)$ for each bCode b detected: $\{(t, b, x, y, \Delta x, \Delta y)\}$. Since the imaging frequency was 1 Hz, we obtained one such set of points for each second of each experiment. The files recording the positions and orientations ranged in size from 15–20 GB.

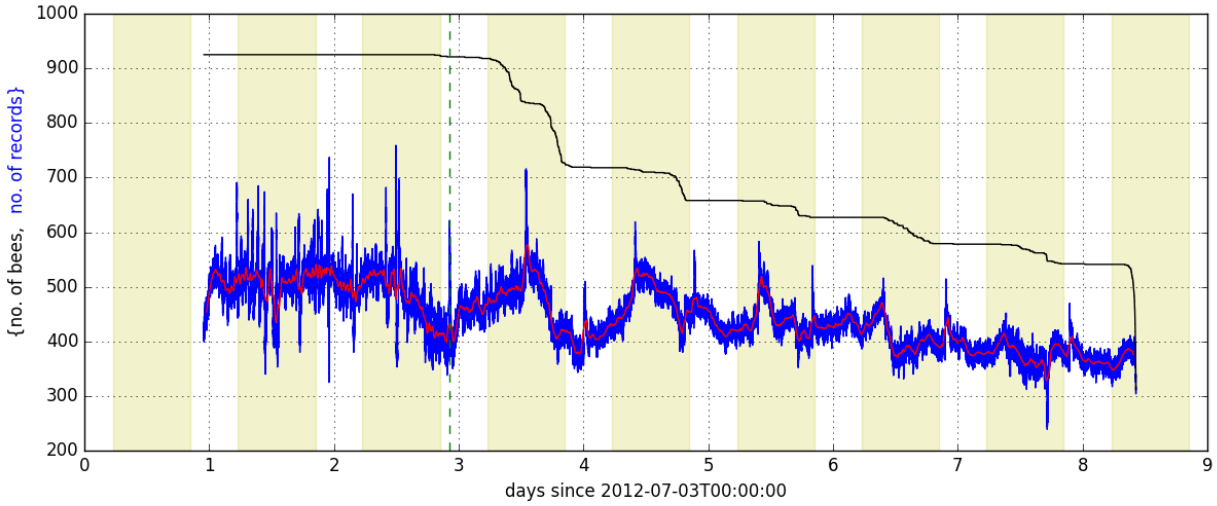

Figure 7.6: **Number of bCodes detected as a function of time.** The black line shows the number of bees alive in the colony, while the number of bCodes detected is shown in blue (with a moving average over 1 h in red). The latter is smaller than the number of bees alive, because bees could be positioned in a way that the bCode does not face the camera, or bees could be outside the hive, or some bCodes are not always successfully decoded.

7.3 Detection of individual behaviours

7.3.1 Trophallaxis detection

Trophallaxis, or liquid-food exchange, is an important social interaction in bees, in which bees touch each other's antennae while one bee transfers food to another bee. Trophallaxis occurs more often than expected if its sole purpose were the distribution of food among workers [86], and has been implicated in communication [133] and in the spread of disease [29]. In trophallaxis, one bee may "beg" for food from another by protruding her proboscis (tongue) towards the mandibles of another bee, keeping her antennae directed at that bee [27]. The other bee may accept and offer a droplet of food, which would then be sucked up via the proboscis. Alternatively, donor bees may spontaneously offer food to other bees without being asked. During the food transfer, the antennae of both bees touch each other frequently, and the success of food transfer is reduced if, for example, there is an amputation of one bee's antenna [27].

Trophallaxis was detected in the raw images using computer vision algorithms designed and implemented by Tim Gernat. The detection procedure first isolated pairs of bees that were in close proximity to each other, and whose orientation-vector geometries allowed for the possibility of trophallaxis (i.e. the pair could not be facing away from each other); we refer to this as the face-to-face configuration. Once such pairs were isolated, a particle swarm optimization algorithm was used to fit a path, composed of four Bézier curves, to the bee heads in the region above each bee's barcodes (Figure 7.7, outlined in pink). After this process

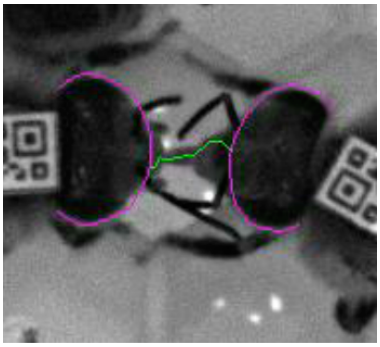

Figure 7.7: **Detection of trophallaxis using computer vision.** Algorithms identify the bees' heads and trace a proboscis connecting them in the image. Photograph courtesy of Tim Gernat.

completed successfully, the final step in the procedure involved searching for evidence of touch between the bees: either antennation, or extension of one bee's proboscis towards the other bee's head. Thus the final algorithm searched for a thin connection between the two heads (shown in green in Figure 7.7).

This detection procedure was applied to every image in the dataset. As a final post-processing step, trophallaxis contacts between the same pair of bees that were detected in successive images were merged into one single trophallaxis event. Furthermore, such interaction events between the same pair bees, but separated by less than 60 seconds, were then merged, provided that the bees remained in the "face-to-face" geometry during the separation interval. This prevented long trophallaxis events from being broken up due to missed detections, but ensured that distinct trophallaxis events—in which the bees move away from each other—were kept separate. In addition, trophallaxis interactions lasting less than 3 seconds were excluded, as such interactions do not always result in the transfer of food [46].

In order to set the parameters for the computer vision algorithms, and to quantify the error rate in detection, a library of 39,863 images was created by sampling pairs of bees in dataset 2013:04 that were sufficiently close for trophallaxis to occur. Each image in this library therefore consisted of a just pair of bees, and each image was manually annotated to indicate whether or not trophallaxis was actually occurring. There were 1,045 images in the library that contained trophallaxis. The performance of the automatic trophallaxis detector was then quantified using the annotations. We could measure the number of true positives, false positives, true negatives and false negatives by comparing the output from the automatic trophallaxis detector against the manual annotation, for all pairs of bees in the library that were sufficiently close for trophallaxis to occur.

Table 7.1 shows the four performance measures that are typically used to quantify the performance of this binary test. These are calculated both for trophallaxis interactions in individual images, and for trophallaxis events derived from merging interactions as described above. For comparison, the metrics are also computed

metric	definition	with computer vision		without computer vision	
		individual images	after post-processing	individual images	after post-processing
sensitivity	$TP/(TP+FN)$	0.68	0.52	0.76	0.65
precision	$TP/(TP+FP)$	0.55	0.80	0.21	0.29
specificity	$TN/(TN+FP)$	0.99	0.99	0.92	0.96
negative predictive value	$TN/(TN+FN)$	0.99	0.99	0.99	0.99

Table 7.1: **Performance of trophallaxis detector** Different metrics based on comparing true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN).

for encounters detected without the use of computer vision algorithms, i.e. only based detecting the face-to-face configuration.

The sensitivity indicates that over two-thirds of trophallaxis events are correctly detected by the automatic detector. Of all trophallaxis detections in individual pictures, 45% are false positives, reducing the precision of the detector; however, this is substantially reduced by the post-processing step, in which trophallaxis detections isolated in time (i.e. not spanning a series of images, and therefore of short duration) are discarded. The post-processing step has the tradeoff that the false negative rate goes up, reducing the sensitivity. This means that we miss some trophallaxis events, but this is less of an issue than having many false positives, as it means that we are effectively sampling the complete set of trophallaxis events and can still draw statistical conclusions.

There is a special issue for trophallaxis detection when it comes to the one queen in the colony. As the single female reproductive member, the queen receives special attention from the worker bees in the colony, who frequently encircle and interact with her, both with their antennae and with their proboscises. The trophallaxis detector therefore cannot distinguish between a true food-exchange event and these other interactions between the queen and the workers, and so that false positive rate for the queen is much higher for interactions involving the queen. Therefore the queen was excluded from the set of trophallaxis interactions generated by the detector.

The end result from the trophallaxis detection procedure is a set of trophallaxis events $\{(b_1, b_2, t_i, t_f)\}$; each event records an interaction between two bees with IDs b_1 and b_2, starting at time t_i and lasting continuously until time t_f.

7.3.2 Dance detection

Another behaviour that might be detected from movement data is the honeybee waggle dance, described previously in section 5.1. A preliminary study on dance detection was carried out by Weibing Deng. His method used compared the orientations of a bCode in two successive time points, $\hat{n}(t)$ and $\hat{n}(t+1)$, where

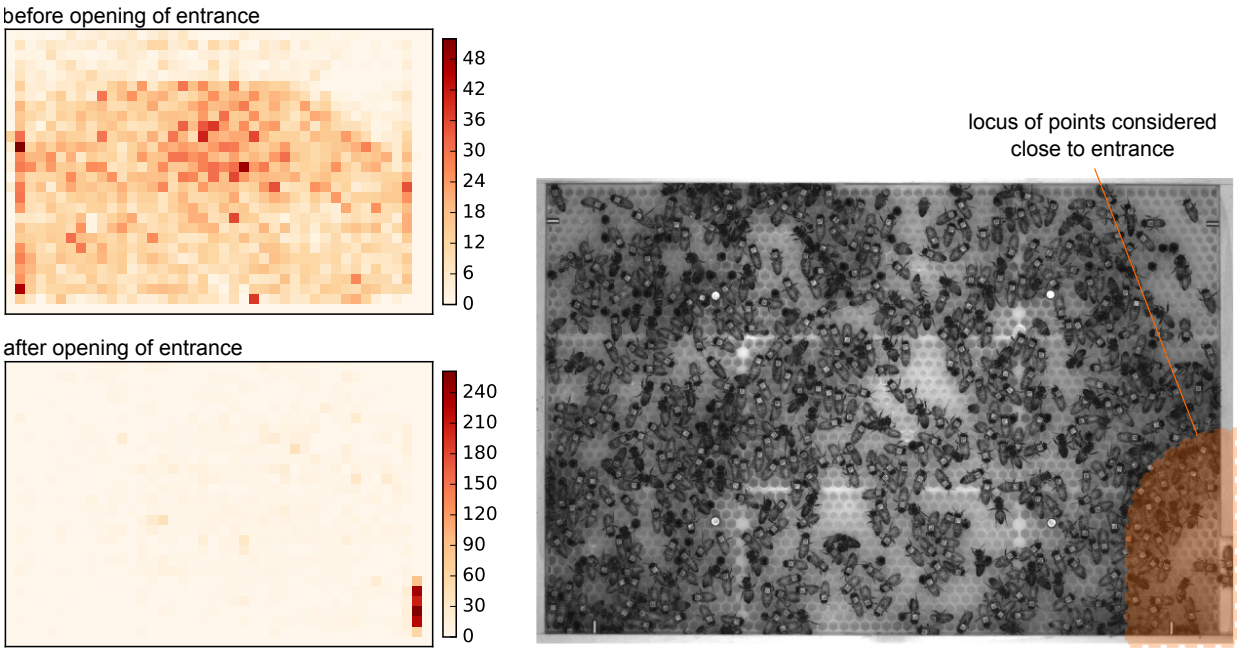

Figure 7.8: **Spatial locations of disappearances.** Left: heatmaps showing relative spatial density for locations where bees disappear from view for at least 30 s, both before and after the entrance is opened (noon to 3 PM on 2012-07-04 and 2012-07-07, respectively). The hive space is broken up into small blocks, and the colour indicates the number of disappearances at each block's location. Right: highlighted is the area near the entrance used in detecting trips.

$\hat{n} = (\Delta x, \Delta y)$ is the unit orientation vector of the bCode at some instant in time. In particular, dances were recorded when two heuristic conditions were satisfied:

1. Rapid change of direction: the dot product $\hat{n}(t) \cdot \hat{n}(t+1)$ is negative at least four times out of five in a 5-second time window. This suggests that a bee is performing a waggle run during this period (see Figure 5.1).

2. Alternating right and left turns: the sign of the cross product $\hat{n}(t) \times \hat{n}(t+1)$ changes at least once, suggesting that at the end of two subsequent waggle runs, the bee alternates directions.

Based on these, we obtained a set of dances $\{(b, t_i, t_f)\}$ indicating that bee with ID b is performing a dance in the interval $[t_i, t_f]$. Since the accuracy of this predictor has not yet been quantified, we consider these results a stepping stone to future work in this aspect of tracking individual behaviours. We use the dance detections only for the exploratory analyses described in section 7.5, and did not use them in the work described in chapter 8 and chapter 9.

7.3.3 Trip detection

In Part II, we studied foraging trips made by bees tagged with RFID p-Chips. In principle, the same kind of information should be available—and more extensively, since every bee was provided with a bCode—in the experiments described here, if we are able to observe bees leaving the entrance and returning. To study the possibility of identifying these foraging trips, we marked the spatial location at which bees disappeared over a three-hour period on two days in dataset 2012:01: first for a day before the entrance was opened (meaning that the bees could not forage) and for a day after the entrance was opened. The heatmaps in Figure 7.8 show the relative spatial density of these disappearance events, for all disappearances that lasted at least 30 seconds.

As noted previously, a bCode may disappear from view of the camera due to several reasons: the bee may go inside the honeycomb; she may be walking on the glass and facing the wrong way; the bCode may be tilted so that it is not read correctly; or the bee may actually be outside the hive (if the entrance is open). The heatmaps show that disappearances do occur even when the hive entrance is closed, but these happen less frequently and are spatially delocalized, compared to the disappearances after the entrance is open. In the latter case, the disappearances are strongly concentrated at the entrance to the hive, suggesting that they are part of a trip outside.

Motivated by this finding, we detected putative trips by looking for bees that disappear, and reappear, near the entrance of the hive. Specifically, we measured the height h of the entrance, and look for trips in the locus of points a distance h away from the entrance (see Figure 7.8), requiring that these trips last at least 30 s. We have no way of quantitatively validating the accuracy of this metric, since we do not actually track bees once they leave the observation hive. However, in section 7.4 we plot the timeseries of the number of trips detected in this way, and that suggests that it works reasonably well (see Figure 7.12). Before the entrance is opened, a relatively small number of putative trips is detected; these, of course, are not real foraging trips. Once the entrance opened, the number of trips increases, particularly during the day, and we see a clear diurnal pattern, with the number of trips falling to background levels during the night. In Table 7.2, we compare the average trip rate before and after the entrance is opened for each dataset from 2013.

Thus, although we expect that the false negative rate for this trip detector is very low, the false positive rate may be substantial, but cannot be easily quantified. The detection procedure could be modified to include measurements of speed immediately before disappearance, and immediately after appearance, for a putative trip—the idea being that a bee heading out or returning will be moving quickly compared to a bee that is, say, emerging from the honeycomb.

dataset	mean rate of detected trips (per hour)		
	overall	before opening of entrance	foraging period
2013:01	200	57.6	244
2013:02	264	20.2	310
2013:04	113	7.44	142
2013:08	173	13.7	197
2013:09	103	14.8	130

Table 7.2: **Trip rate in different periods.** The average trip rate is calculated over three time periods: the whole experiment; the period when the entrance is closed; and the period following 24 h after the entrance is opened.

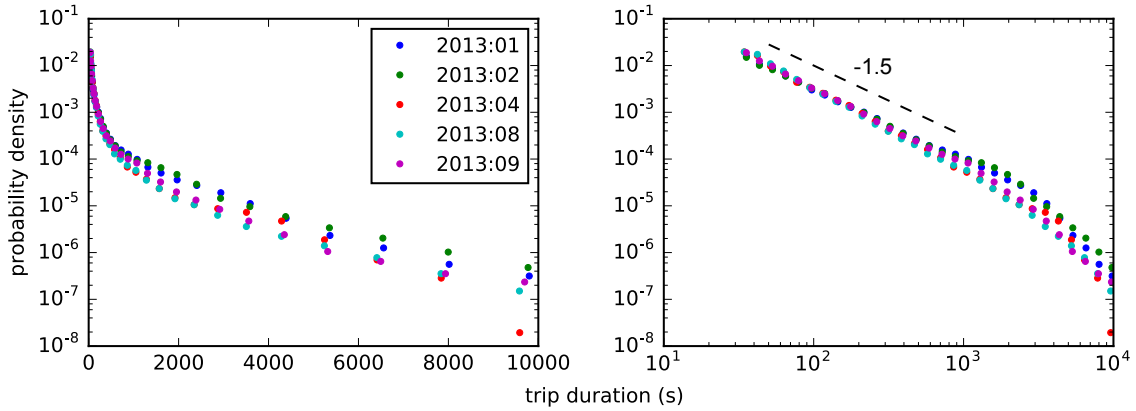

Figure 7.9: **Duration distribution of trips.** The probability density for trip duration is plotted on both semilogarithmic and double-logarithmic axes, using logarithmically-spaced bins on the abscissa. The distribution appears to be approximately follow a power law form (with exponent roughly -1.5) followed by an exponential cutoff.

The trip detection allows us to measure the distribution of the duration of trips—one of the measurements we wished to make in Part II, but were not able to due to the poor detection rate (see subsection 6.3.1). In Figure 7.9, we plot this distribution for those trips occurring at least one day after the hive entrance was opened (referred to as the foraging period in Table 7.2), and lasting at least 30 s. We also filtered out a very small number (on average, 0.05%) of putative trips lasting longer than 3 h. The distribution is consistent across datasets, and shows a broad distribution with power-law scaling that spans almost two orders of magnitude on the abscissa.

7.4 Resilient diurnal rhythms in colony activity

Like many other living organisms, including humans, the time variation in individual honeybee behaviour can be strongly entrained to the diurnal cycles of the sun. Honeybees are known to show sleep-like states when they are in the forager stage of development, with neuronal oscillations resulting in reduced visual

sensitivity at night [65]. However, inside the hive, larvae need to be fed at all times, and so the nurse bees responsible for this task work around the clock. Indeed, it has been shown that as worker bees age and progress from nurse to forager, they begin to develop rhythmicity by gradually reducing their nighttime activities [14, 81]. This development of rhythmicity has been linked to levels of a particular brain mRNA, which is not regulated by exposure to sunlight or age alone, suggesting a link between social context and regulation of circadian rhythms [120].

Previous studies have relied on sampling the activities of some individuals in the colony over short observation windows distributed throughout the day. The bCode tracking system, on the other hand, allows us to measure the activity of every individual in the colony continuously in time, and we can measure both individual timeseries of activity and the activity of the colony as a whole. Here we describe three metrics we used to capture colony-level activity. First, to measure activity manifested in movement, we define the instantaneous kinetic energy of an individual to be the square of its speed; the kinetic energy of the colony is then the sum of this quantities over all individuals present. We estimate the speed at time t as $|\vec{x}(t) - \vec{x}(t-1)|$, where $\vec{x}$ is the position of the bee in the image in pixels. This yields the kinetic energy in units of pixels-squared per second, which we rescale by $(250\text{ px})^2$ (the length of a typical bee in the images is about 250 px); Figure 7.10 shows the resulting timeseries for one dataset. We have not added a rotational component to the kinetic energy, but this is potentially important, given the relatively high packing fraction of the bees in the 2-dimensional hive space.

Second, we measured the rate of trophallaxis interactions in the colony (Figure 7.11). Trophallaxis occurs throughout the experiment, though before the entrance is opened, it involves the exchange of food already provided to the hive, while later on, foragers bring in additional food from external sources. Finally, we plotted the timeseries for the rate of trips in the colony. As discussed in subsection 7.3.3, not all of these are foraging trips (in particular, the ones before the entrance is opened).

All these timeseries show that 24–48 hours after the entrance is opened (and foraging activity begins) the overall colony activity is strongly entrained to the diurnal cycle, with a sharp rise in activity soon after sunrise, and a sharp drop in activity as sunset approaches. These patterns of activity are resilient to the forager removal perturbation, in which a large fraction of the colony was removed (see Figure 7.4). In particular, the timeseries for rate of trips shows no obvious anomaly in the day after the perturbation, indicating that division of labour within the hive has shifted to compensate for the loss of foragers. In Figure 7.12, there is a sharp decline in number of trips soon after the forager removal perturbation begins; this would seem to make sense, but is in fact not replicated across the other datasets (see Appendix A, where these timeseries are shown for other datasets).

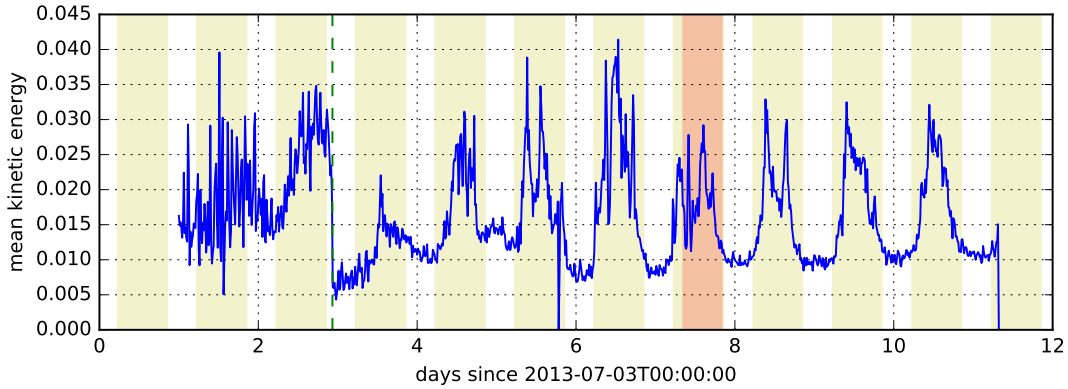

Figure 7.10: **Timeseries of mean kinetic energy.** Mean kinetic energy in 15-minute time windows for dataset 2013:01. The units for kinetic energy used here are $l_b^2 s^{-2}$, where $l_b = 250$ px is roughly the length of a bee in an image.

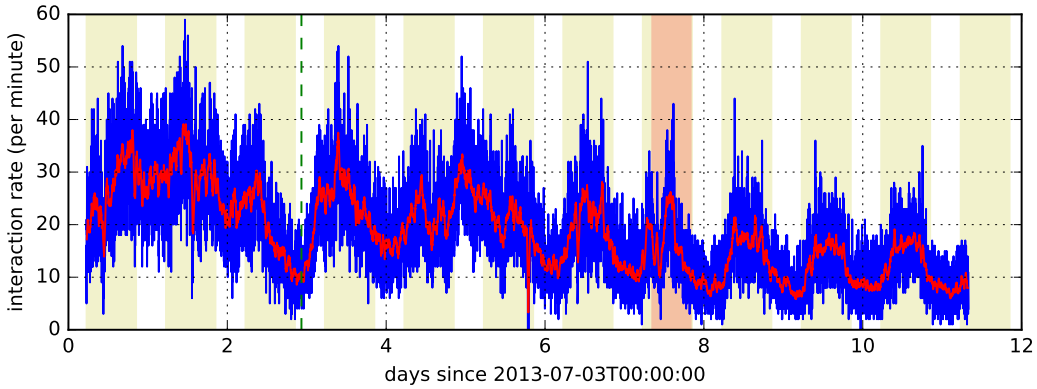

Figure 7.11: **Timeseries of trophallaxis interaction rate.** The number of trophallaxis events in each minute of dataset 2013:01 is shown in blue, while the red is a moving average over 60 minutes.

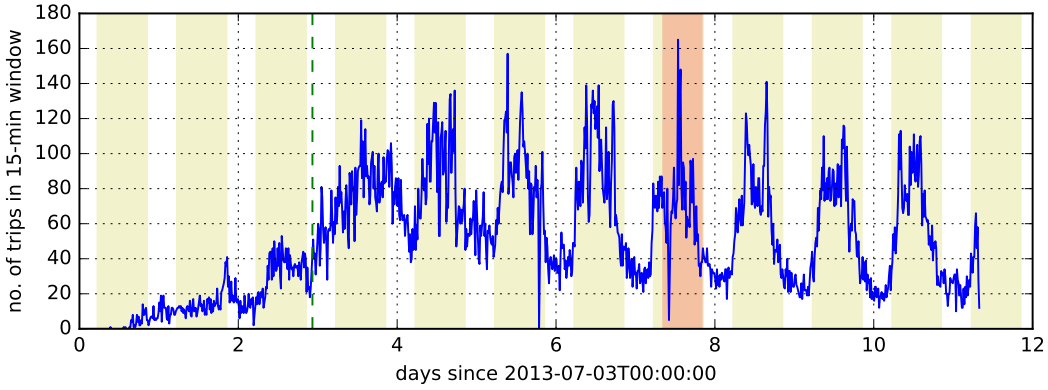

Figure 7.12: **Timeseries of number of detected trips.** The timeline for dataset 2013:01 was divided into 15-minute windows, and plotted here is the number of trips that began in each window.

Our work has focused on social networks, but these data offer the potential to study individual-level variation in activity (such as the onset of rhythmicity in the nurse to forager transition). Moreover, all the analyses developed in Part II could be applied to the trip data obtained here. There are promising avenues for future exploration.

7.5 Forager identification

In analyzing individual-level longitudinal data, it is often useful to be able to characterize the instantaneous state of an individual. One can then study how that state varies over time, how individuals of different states interact, and the relationship between the macroscopic state of the colony and the distribution of individual states. We would naïvely expect that the dimensionality of the state space of a bee is rather high, given the variety of activities that bees perform in the colony.

Rather than considering an abstract state space that is derived in some way from individual bee behaviours, we can study the subspace involving foraging: that is, at any instant in time, a bee is either or forager or a non-forager.[3] Based on the behavioural metrics developed in section 7.3, can we correctly assign a bee a place in this space? On the one hand, this problem may seem trivial: the bees that make trip and perform dances should be foragers. However, as we noted in section 7.3, both these metrics likely contain a significant number of false positives, and these are not easily filtered out.

We addressed this question from the viewpoint of *supervised learning*: we construct a feature vector $\vec{x}_i$ for each bee i and on this basis, try to make a classification $y_i \in \{0, 1\}$ for whether or not it is a forager. In the language of statistical learning, we are trying to learn the function (or classifier) f, where $y_i = f(\vec{x}_i)$. This is in contrast to unsupervised learning, discussed in subsection 3.3.2, where the goal is to try to discern clusters or patterns in the set of bees based on their feature vectors.

The classifier must be learnt using a labelled set of points $\{(\vec{x}_i,\ y_i) : i = 1, \ldots, N\}$, that is, we need some data in which it is known which bees are foragers and which are non-foragers, that we can use to train the classifier. In general, we do not have this information at an arbitrary instant in time—indeed, that is the goal of learning a classifier. However, the nature of the perturbation in the experiments from 2013 allows us to deduce (with only a small degree of uncertainty) which bees are foragers on the preceding day: in other words, the bees that were removed in the perturbation were foraging on that day of the experiment (and their identities are known from their time of death), so it is very likely that they were also foraging on the previous day.

[3]The author is grateful to Yoshi Oono for suggesting the idea of a general bee state space, going beyond the simple binary distinction between forager and non-forager. This way of thinking about the relationship between the macroscopic state of the bee society and the microstates of its constituent individuals provides fertile grounds for future exploration.

		predicted	
		0	1
actual	0	264	22
	1	31	93

sensitivity	0.711 ± 0.041
precision	0.797 ± 0.035
specificity	0.926 ± 0.015
negative predictive value	0.887 ± 0.019

Table 7.3: **Performance of the forager classifier.** Left: A sample confusion matrix showing the results of a single classification run. Right: Performance metrics calculated over 1000 cross-validation runs (values are mean ± one standard deviation). (See Table 7.1 for metric definitions.)

		predicted							
		0				1			
dataset →		02	04	08	09	02	04	08	09
actual	0	560	610	794	605	12	49	64	72
	1	214	172	83	127	254	177	99	120

Table 7.4: **Classifier validation across datasets.** Confusion matrices for each dataset.

Based on this idea, we partitioned the bees into foragers and non-foragers for the day before forager removal, in each of the datasets from 2013. For each bee, we constructed a feature vector, consisting of the following components derived from the period from 10 AM to 4 PM on the day preceding forager removal:

1. fraction of time spent in the fourth quadrant of the hive (the hive entrance lies in this quadrant)
2. total duration of time spend dancing
3. total duration of trips
4. mean of speed values in the 10 s before a trip begins and after a trip ends

Using these feature vectors and the labelled data, we performed two sets of numerical experiments. In both cases, we tried to learn the classifier using the support vector machine (SVM) model implemented in the `scikit-learn` machine learning module for the `python` programming language.

First, we focused on a single dataset, 2013:01. We randomly partitioned the set of bees into a training set and a validation set (in the ratio 60:40), and fit the SVM (with a linear kernel) on the training set. We then used the resulting classifier on the validation set to see how well it would perform. In particular, we can enumerate the so-called "confusion matrix" (see Table 7.3) and calculate performance metrics just as we did for the trophallaxis detector in subsection 7.3.1. To obtain bounds on the performance, we performed cross-validation, that is, we repeated the procedure described above 1000 times, repartitioning the datasets each time to give new training and validation sets. Table 7.3 shows how the classifier performs over these cross-validation runs.

Second, we trained the SVM on the complete labelled set from 2013:01, and tested in on the other

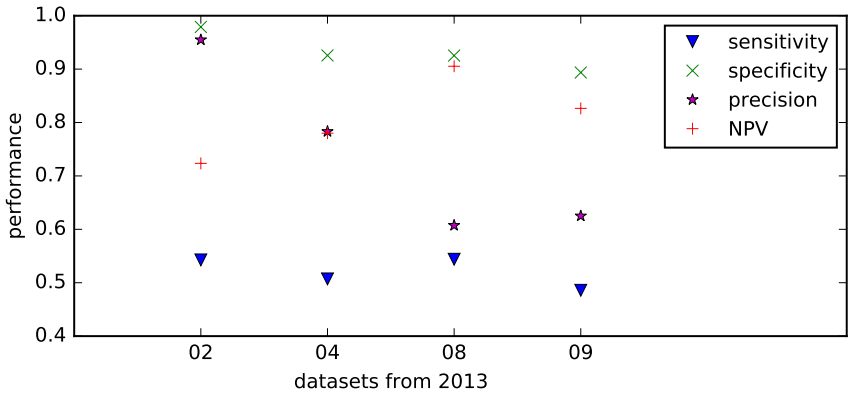

Figure 7.13: **Classification across datasets.** The classifier was trained on dataset 2013:01 and validated on the other four datasets from 2013. The performance metrics for these validation runs are plotted here. (See Table 7.1 for metric definitions.)

datasets. Table 7.4 shows the confusion matrices for each dataset in this case, and Figure 7.13 shows the performance metrics.

In both cases, the sensitivity are negative predictive value are relatively high (over 88% in the cross-validation runs for dataset 2013:01), indicating that the classifier performs well at identifying non-foragers. The specificity and precision are somewhat lower in both cases, due to false positives in identifying foragers.

7.6 Discussion

In this chapter, we have seen that we can capture the activities of all individuals in a real animal society with high spatiotemporal resolution, and we have explored the dynamics of the colony, detected individual behaviours, and tried to learn a classifier that assigns a biologically-relevant state to each individual. A promising future direction is to explore if individuals can be characterized by a well-defined state—not simply forager vs. non-forager—on the basis of data derived from the tracking data.

How can we discern structure in the society? One way is to use the framework of networks, which provide a compressed (but lossy) representation of the data. Real honeybee interactions occur in space and vary in time. A static network essentially ignores both these features and simply represents information about which bee interacts with which other bee (and how often, and so on). A temporal network includes time information, but still ignores the spatial aspect of interactions. In the next two chapters, we will study interaction networks for honeybee societies, particularly for the interaction of trophallaxis, to explore what structures are present in the society.

Chapter 8

Power laws and other structure in the honeybee trophallaxis network

Honeybees use multiple channels of communication inside the hive in coordinating their society, such as the dance, trophallaxis, pheromone signalling, antennation, and vibrations. Trophallaxis is but one channel of communication, but is believed to be important both for communication and in the spread of disease through the colony. In chapter 7, we described methods to track trophallaxis interactions inside a honeybee colony continuously over an extended period of time. This allows us to construct and study a social network based on trophallaxis. In this chapter, we will study some network properties ignoring the temporal aspect of the interactions. In chapter 9, we will incorporate the temporal properties into the social network using the framework of temporal networks. Here we will also study the statistics of one temporal feature of the interactions, namely their duration. We'll compare this against human communication networks, and show how both can be understood based on the movements in space that underlie the formation of these social interactions.

8.1 Basic static network properties

A social network is a network, or graph, in which individuals are represented by a set of nodes $\{i\}$, and connected by edges (i, j) when they interact. The edges may be directed if the interaction is asymmetric (for example, one individual is a donor and the other a recipient). The edges may also be weighted by some property of the interaction: for example, the edge for a pair of individuals who interact very frequently may be weighted higher than for a pair that do not interact much. Networks are useful when they can tell us something about the structure of the underlying interactions, and there is a whole body of literature devoted to metrics to capture network structure [8, 85]. In addition, one can create generative models of networks and compare their properties with real-world networks. Some well-known generative models are the Erdős-Rényi model for generating random graphs, the Barabási-Albert model to generate random scale-free networks, and the Watts-Strogatz model to generate small-world networks [8].

Two basic distributions that characterize network structure are the distributions of degree and edge

weight. The degree of a node is just the number of distinct interaction partners it has; the edge weight is a property of a link, and may be constructed in many different ways. For the honeybee trophallaxis networks, we will construct edge weights in two ways: first, we will simply count the frequency of interactions between two nodes i and j and call that weight c_{ij}; and second, we will measure the total duration of interactions and define that as the weight d_{ij}.

We have seen that honeybee trophallaxis interactions do not occur uniformly in time, but instead display nonstationary dynamics (see Figure 7.11). We can construct a static network simply by aggregating all interactions in a specified time window, and ignoring the time-ordering of the interactions—in effect, we are creating a compressed representation in which the detailed temporal information about the interactions is lost, and we simply focus on who interacts with whom, and how often. The static network that we create does, however, depend on the choice of time window over which we aggregate interactions, and so the distributions that we measure on the network also depend on this time window.

In Figure 8.1, we plot the degree distribution and the distribution of each edge weight c_{ij} and d_{ij}, for static networks created by aggregating over increasingly long time windows. Specifically, each time window extends from the start of the experiment through to each sunrise after the hive entrance was opened. The degree distribution is non-stationary: it is bell-shaped, with the mean and variance both growing in time. For the largest time window (i.e. upto sunrise on the last day), the typical bee has degree 335.2 ± 86.2 (mean $\pm$ SD); for comparison, the number of bees in the colony at the time the entrance is opened was 1088 in this dataset.

The bell-shaped degree distributions are typical of the Erdős-Rényi random-graph model, in which the network is constructed by adding each possible edge with some probability p, independent of the other edge or the node properties. This suggests that perhaps there is not much structure in the choice of interaction partners. In a honeybee society, there is differentiation of the nodes by task, but worker bees do not recognize each other as individuals, so this might be responsible for determining the form of the degree distribution.

On the other hand, both link-weight distributions show that there is structure in the interactions, and the distributions do not appear to change much in time. Both distributions follow a power law form, with the distribution for d_{ij} ranging over two orders of magnitude in duration. This suggests that there is some structure in the time-aggregated trophallaxis network: most links happen with low frequency and low duration, but there is a small number of links that are activated very frequently. This has interesting biological implications: given that bees do not recognize different individuals, they are not "choosing" to interact more with the particular individuals in the tails of the distribution. Rather, this might arise from division of labour—perhaps bees involved in the same task interact more frequently among themselves than

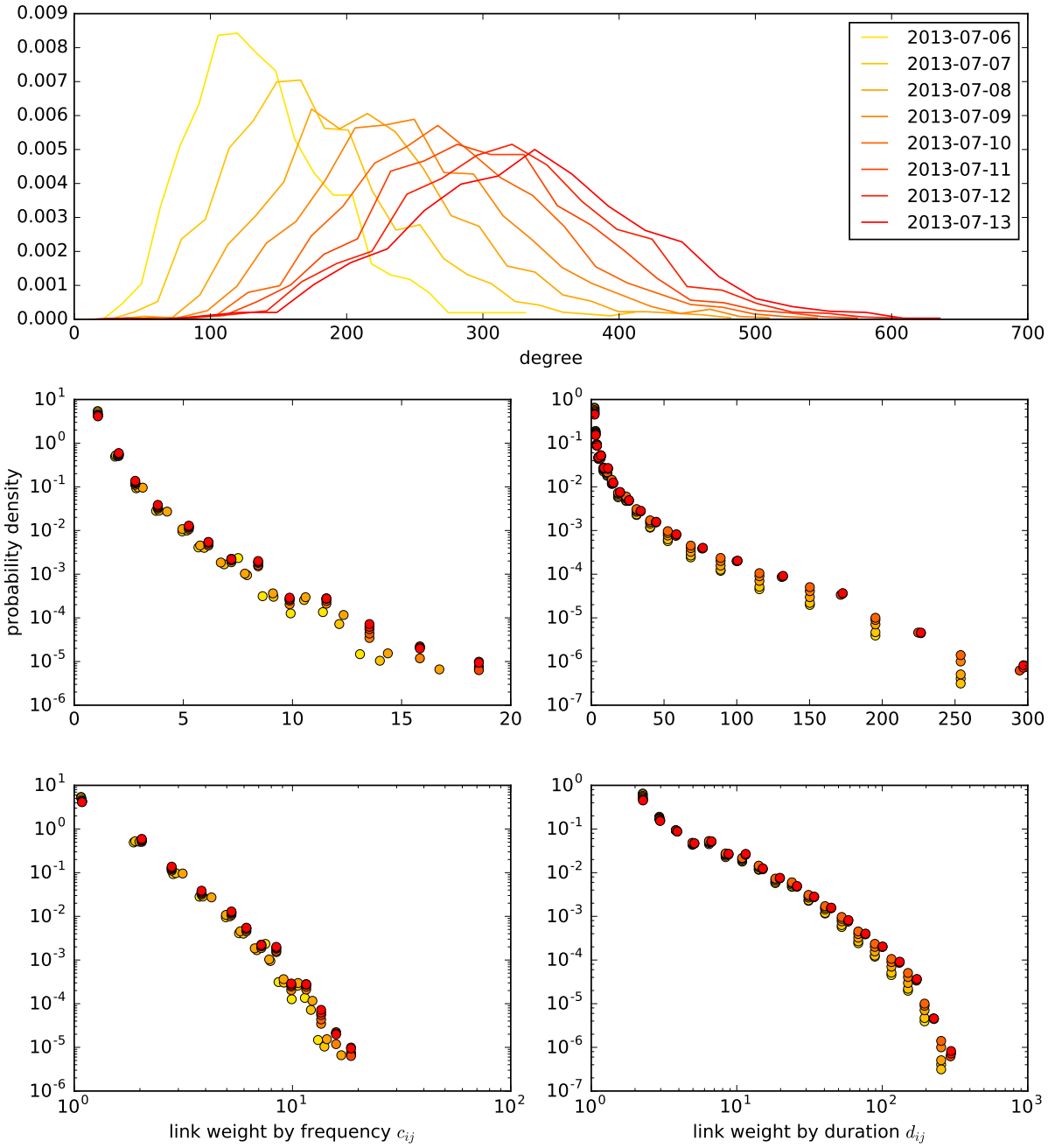

Figure 8.1: **Static network properties of the trophallaxis network.** The degree and link-weight distributions are calculated for several time windows, each beginning at the start of the experiment on 2013-07-03, and ending at sunrise on the date indicated in the legend. The weight distributions are plotted on both semi-logarithmic and double-logarithmic axes. The weight distributions appear to reach stationarity quickly while the degree distribution grows in mean and variance.

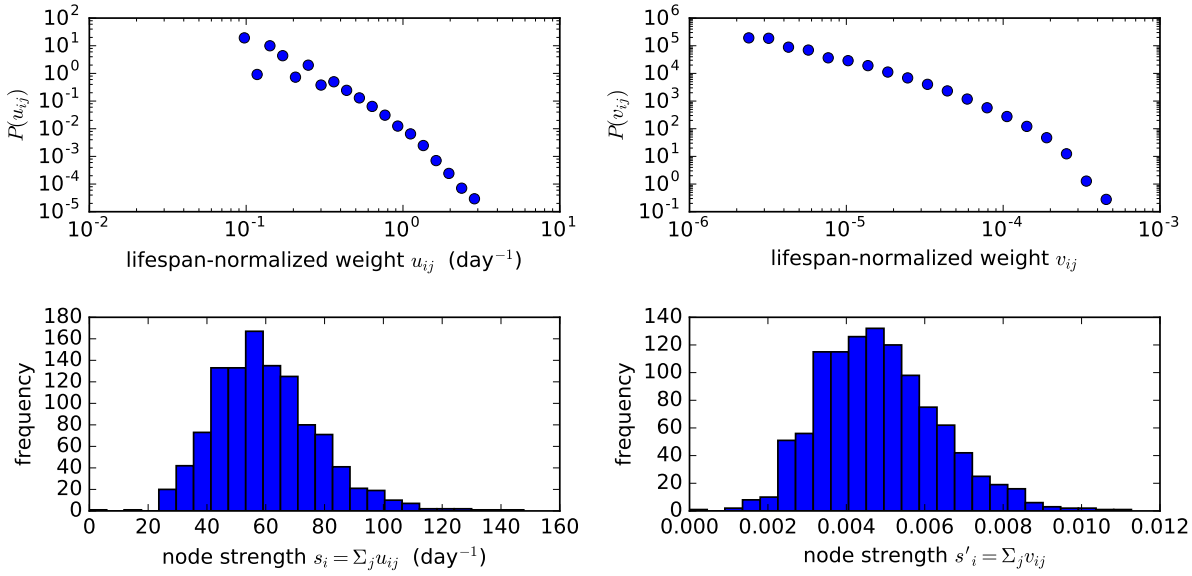

Figure 8.2: **Lifespan-normalized distributions of link weights.** Link weights are based both on frequency and duration. Also plotted are the distributions of node strength.

with other bees—or perhaps due to some underlying spatial heterogeneities in their interactions.

One feature of the trophallaxis network is that bees die over time (see Figure 7.4), both due to natural mortality and due to the forager-removal perturbation. This could confound our measurements of the link-weight distribution: for example, the total number of interactions between a pair of nodes might be short simply because one of the nodes died early. To correct for this, we normalize both link weights by lifespan by defining

$$u_{ij} \equiv \frac{c_{ij}}{\min\{T_i, T_j\}}, \qquad v_{ij} \equiv \frac{d_{ij}}{\min\{T_i, T_j\}}, \tag{8.1}$$

where T_i represents the lifespan of node i. Figure 8.2 shows the distribution for these normalized link weights. The form of the distribution remains basically the same as before. In addition, Figure 8.2 shows the node strength, calculated for each node by summing over all weights on links attached to that node. In the tail of these distributions for node strength are a small number of nodes with unusually high node strength, again a sign of social structure in the network.

8.2 Distribution of trophallaxis durations

We saw in the last section that the distribution of link weights based on total duration follows a power law distribution. What about the overall distribution of the durations themselves, and what is the total number of interactions across the nodes? In Figure 8.3, we answer this question by plotting two quantities. First,

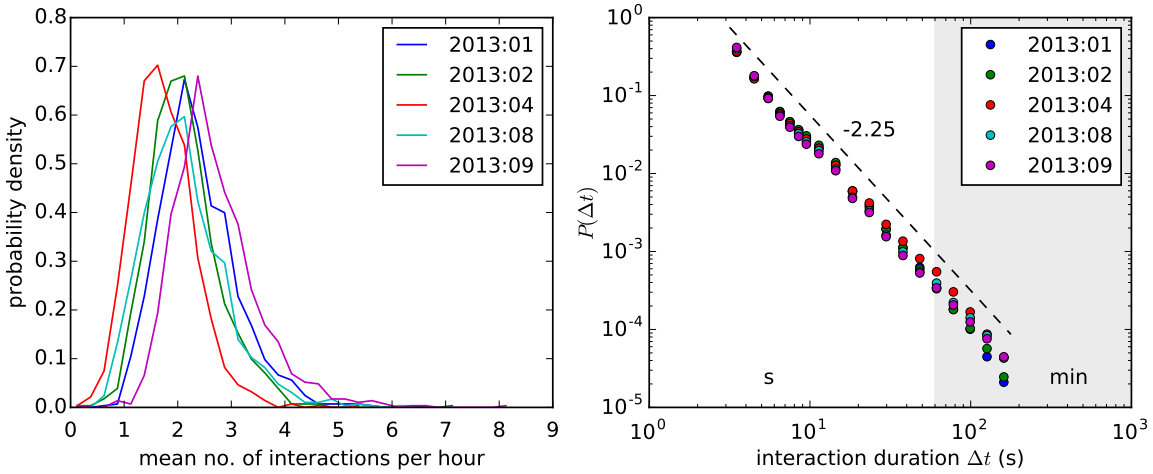

Figure 8.3: **Distributions of interaction rate across bees and interaction durations.** The interaction duration distribution is estimated using linearly-spaced bins below 10 s and log-spaced bins beyond 10 s, to avoid artefacts arising from the 1-second time resolution. To guide the eye, the dashed line represents a power-law scaling with exponent -2.25.

we plot the distribution of interaction rate across bees, that is, the mean number of interactions per bee, taking into account its lifespan; and second, we plot the distribution of the duration of all interactions.

The distribution of interaction rate across bees is not particularly surprising; the mean interaction rate in dataset 2013:01 is (2.38 ± 0.73) per hour (one SD). On the other hand, the duration distribution takes on a remarkably clean power-law form with exponent roughly -2.25, spanning the entire support, 3 s $\leq \Delta t \leq$ 180 s, of the distributions, and is consistent across datasets. Note that due to the post-processing step in the trophallaxis detection procedure (see subsection 7.3.1), interactions lasting more than 180 s and less than 3 s were discarded.

Next we wish to understand if the duration distribution arises from some sort of heterogeneity in the interaction durations across bees. When we visually plotted the duration distributions for random samples of bees, we found that they appeared to follow roughly the *same* power-law form as that for the aggregate interaction distribution shown in Figure 8.3; that is, the distribution of interaction durations Δt for each bee b took the form,

$$P_b(\Delta t) \sim [\Delta t]^{\alpha(b)}. \tag{8.2}$$

To test this quantitatively, we used a maximum-likelihood fitting procedure to obtain the power-law exponents $\alpha(b)$ for each bee b, and in each case we tested whether a power-law distribution is a better fit to the data than an exponential distribution using a goodness-of-fit test based on the Kolmogorov-Smirnov statistic [21]. This allows us to obtain a more accurate estimate of the exponents than by a simple linear

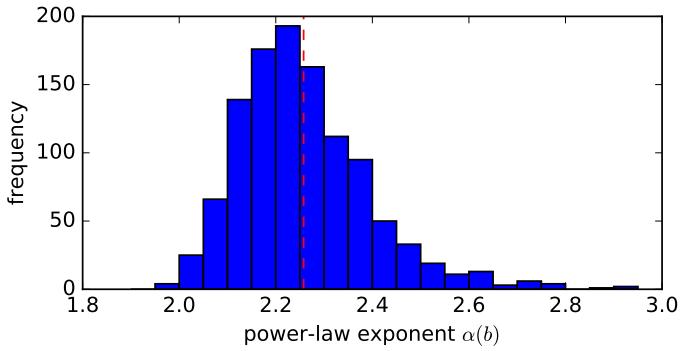

Figure 8.4: **Distributions of power-law exponents α across bees.** The histogram uses uniformly spaced bins of size 0.05.

regression on double-logarithmic axes—which is statistically unsound [21]—and second, allows us to verify if the distribution is truly different from an exponential distribution. We used these methods as implemented in the `powerlaw` package [1] for the `python` programming language.

For the 1164 bees with interactions in the dataset 2013:01, the goodness-of-fit test indicated that the power-law distribution was a better fit than an exponential distribution in 1115 cases (p-value $p < 0.01$ in a log-likelihood test). (The remaining 49 bees typically had fewer interactions, and a small maximum duration and thus smaller support for the estimated distribution of durations.) In Figure 8.4, we plot the distribution of $\alpha(b)$ across these 1115 bees. We see that the values cluster close to the mean, with exponents typically in the range 2.258 ± 0.136 (mean $\pm$ one SD).

These results suggest that the distribution is a universal feature of honeybee trophallaxis interactions. In the next section, we will see that this feature actually also spans different species, and we propose a simple explanation for the exponent of these power law distributions.

8.3 Modelling the duration distribution in face-to-face networks

8.3.1 Face-to-face interactions in honeybees and in humans

We have discovered that honeybee trophallaxis distributions follow a characteristic power-law form that spans two orders of magnitude in duration, with scaling exponent close to -2.25. What does this distribution tell us about honeybee biology? That is, is there some reason why bees appear to mostly spend relatively little time on exchanging food, but occasionally participate in exchanges of food that last two orders of magnitude longer? Is this really a feature intrinsic to the biological process of trophallaxis?

We address these questions by studying a second network for the bees, formed by considering face-

to-interactions alone, that is, by recording an interaction whenever a pair of bees is in the face-to-face configuration described in subsection 7.3.1. Since this is a prerequisite for a trophallaxis interaction to occur, the face-to-face network is a kind of superset of the trophallaxis network. Comparing the two enables us to check what effect the specific biology of trophallaxis has on any statistical property we measure—in this case, the interaction duration distribution.

For comparison, we also plotted similar face-to-face networks from the SocioPatterns project, which uses RFID sensors to construct face-to-face networks for human interactions in various settings [17]. The datasets from SocioPatterns are publicly available, and we analyzed four of them:

- `highschool_2011`: contacts, over a total of 4 days, among 126 students in a high school in Marseilles, France, in December 2011.
- `highschool_2012`: contacts over 7 days among 180 students from the same high school in November 2012.
- `hospital`: contacts among patients and health-care workers (75 individuals in total) in a hospital ward in Lyon, France, over four days in December 2010.
- `ht09`: collected at the ACM Hypertext 2009 conference, where 113 volunteers wore radio badges that monitored their face-to-face proximities.

Figure 8.5 shows the interaction duration distributions for the two bee networks and the four human networks described above; the durations have been rescaled by the sample mean, in order to be able to compare them. All the distributions roughly fall on top of one another over two orders of magnitude, with some noise at the top of the distribution (due to binning) and at the tail (due to the real cutoff in the scaling). In particular, the bee distributions for trophallaxis and face-to-face interactions are very similar, suggesting at first glance that the biology of trophallaxis does not necessarily manifest itself in this distribution. Judging by eye, the exponent $(-\alpha)$ of the power law appears to fall somewhere in the range $(-3.0, -2.0)$, and to measure the exponent more precisely, we make a set of *compensated plots*. Each such plot corresponds to a specific value of α, and we plot each empirical distribution $P(\theta)$ divided by the function $A\theta^{-\alpha}$ as a function of θ (where we denote the rescaled time as $\theta \equiv \Delta t/\langle \Delta t \rangle$ and A is a normalization factor that will not be relevant). If the plot for a given value of α flattens over some reasonably wide range, that indicates that the distribution in fact scales as $\theta^{-\alpha}$ over that range.

Figure 8.6 shows compensated plots for eight values of $\alpha \in [2,\ 3]$. First, we notice that the distribution for bee trophallaxis durations is actually slightly different than the other distributions for face-to-face contacts. Based on the compensation plots, we estimate that the magnitude of the exponent for trophallaxis falls in

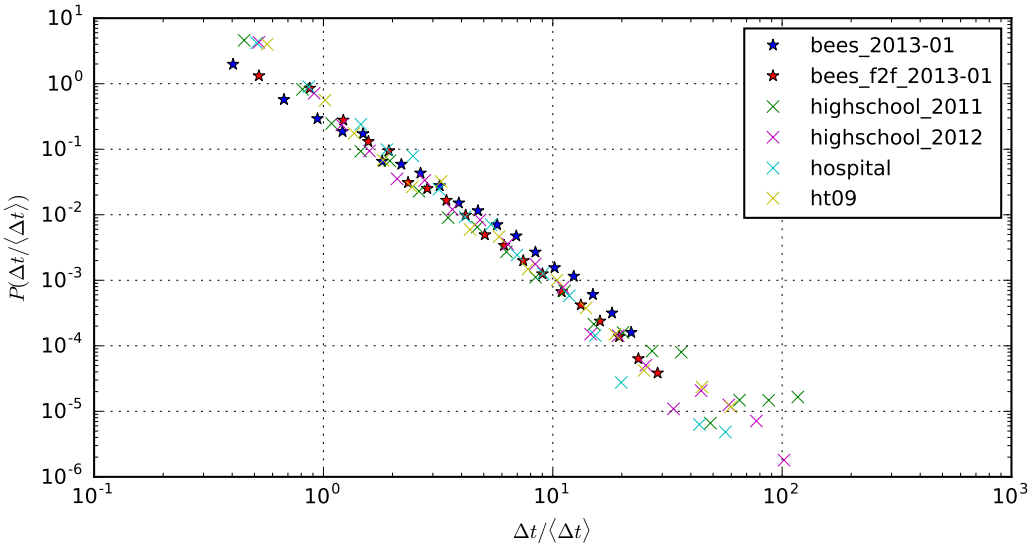

Figure 8.5: **Duration distributions in bee and human communication networks.** The two bee distributions correspond to trophallaxis durations and durations spent in the face-to-face configuration. The other four distribution are for human communications. All distributions have been rescaled by the mean to facilitate comparison.

the range (2.25 ± 0.15), since the compensation plots for these values in this range are flat over $10^0 \leq \theta \leq 10^1$. This estimate also agrees with the distribution of exponents calculated for the related, *individual* duration distributions (see Figure 8.4). The bee and human face-to-face duration distributions, meanwhile, have an exponent that is higher in magnitude, which we estimate from the compensation plots to fall in the range (2.65 ± 0.15). Since these two sets of distributions appear to be distinct, we plot them separately in Figure 8.7.

We are left with two questions: first, why are the face-to-face duration distributions universal for two very different organisms interacting in disparate social settings? And second, why does the trophallaxis distribution appear to scale with a smaller exponent than the face-to-face distribution in honeybees? We will propose an explanation for the values of these exponents in subsection 8.3.3, but first we review an existing model from the literature for human face-to-face interactions, as we will use this to test our explanation.

8.3.2 Modelling face-to-face interactions

Our goal in this subsection is to review a model proposed by Starnini et al. [113] in 2013 to explain features of human face-to-face interactions, such as those in the SocioPatterns datasets. This model was able to reproduce the interaction duration distribution and the distribution of inter-interaction times, as well as some other network features such as the weight distribution, the scaling of node strength with degree and

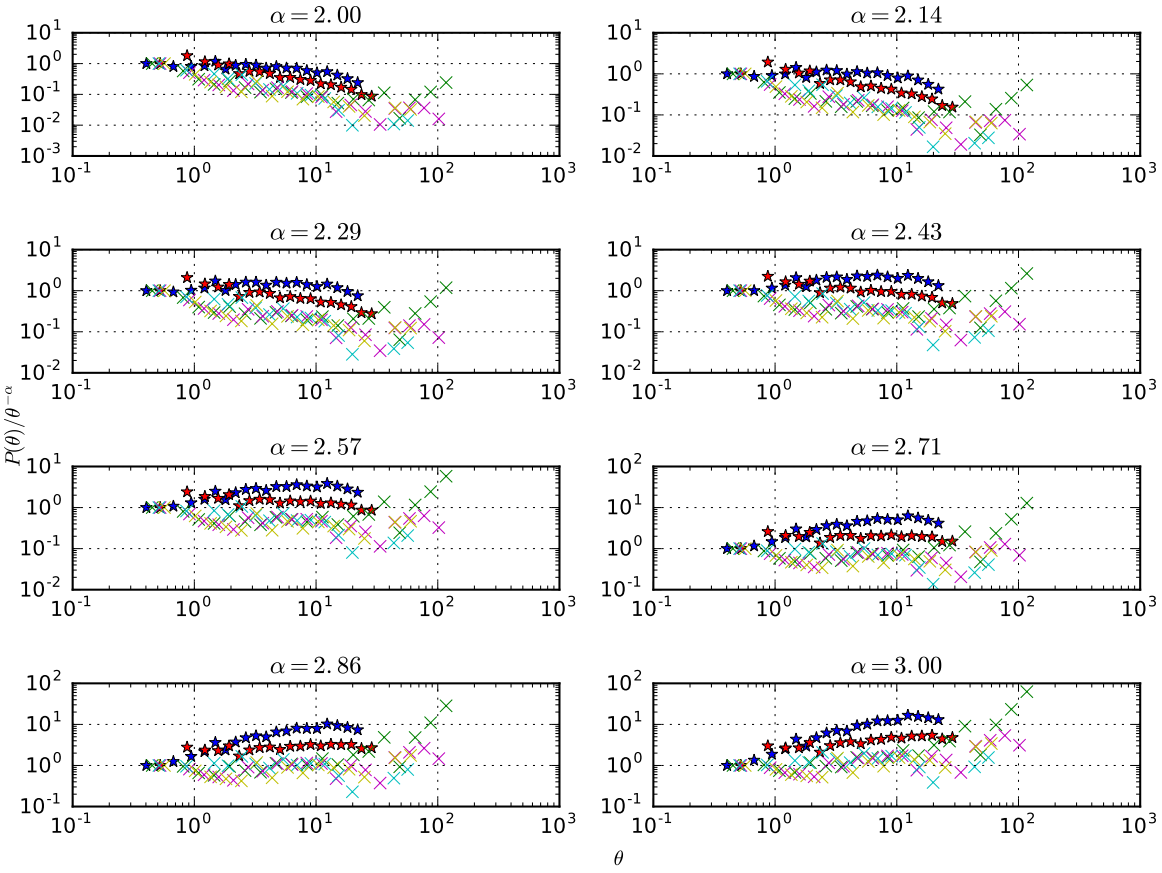

Figure 8.6: **Compensated plots of bee and human duration distributions.** We can determine both the scaling exponent and scaling range using the flat regions of compensation plots. The set of plot markers from Figure 8.5 is used again here.

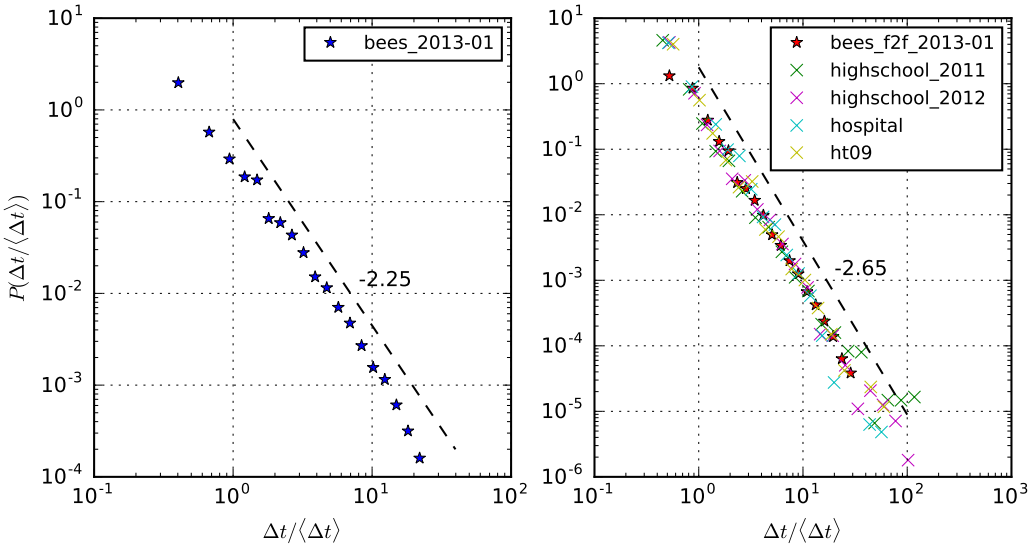

Figure 8.7: **Duration distributions: trophallaxis vs. proximity.** The scaling of the two distributions is slightly different, as revealed by the compensated plots.

the scaling of the degree with time; recently the authors published another paper [114] that studied more aspects of the same model, comparing other network properties, such as group sizes and path lengths.

In the model, agents move around in a two-dimensional space and interact with each other based on proximity: that is, when two agents are within some threshold distance of each other, they are said to interact. The model has three main features:

1. Each agent performs a random walk with a fixed step size when it is not interacting with any other agent; if the agent is interacting, then the random walk is biased, as discussed below.

2. Individual agents have different degrees of social appeal or *attractiveness* $\{a_i : a_i \in [0,1]\}$, which determines the movement of other agents that they interact with. In simulation, a_i was randomly sampled from a uniform distribution on $[0,1)$.

3. Not all agents are simultaneously present in the system; they can switch between active and inactive states, which determine whether or not they move or establish interactions.

The attractiveness of an individual affects the movement of other agents that it interacts with as follows: each agent i at a given instant in time t decides to take a step with probability

$$p_i(t) = 1 - \max_{j \in \mathcal{N}_i(t)} \{a_j\}, \tag{8.3}$$

where $\mathcal{N}_i(t)$ is the neighbourhood of i at that instant in time, i.e. the set of agents in its proximity. If

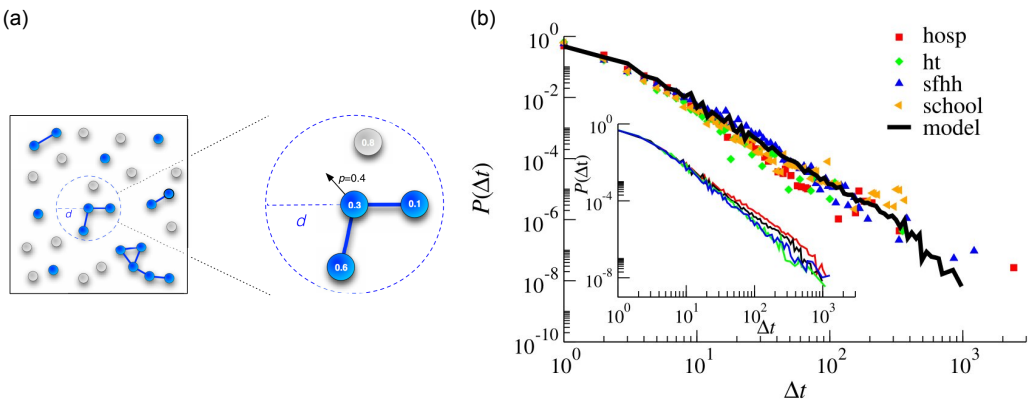

Figure 8.8: **Mobile agent model for proximity interactions.** (a) Inactive agents are shown in grey, active agents in blue. Each agent is assigned an attractiveness value. The focal agent's movements are determined the attractiveness of its active neighbours. (b) The model reproduces the duration distribution seen in human face-to-face networks. Both figures reproduced with permission from Ref. [113].

this neighbourhood is empty, then we have a regular random walk; if there are some agents present in the neighbourhood, then agent i slows down to a degree determined by the most attractive agent in the neighbourhood (see Figure 8.8a).

Simulations of this model show that the distribution of interaction durations matches those in four SocioPatterns datasets (see Figure 8.8b). Starnini et al. say that their model is not sensitive to the specifics of the slowing-down rule in Equation 8.3—one could, for example, use the average rather than the maximum over $\mathcal{N}_i(t)$. However, the heterogeneity in attractiveness values $\{a_i\}$ is essential; if all individuals were to have the same attractiveness, the model maps to a first-passage problem with $P(\Delta t) \sim (\Delta t)^{-3/2}$ [113]. We have also verified that the third feature (inactive vs. active agents) is not necessary to reproduce the duration distribution.

While the model appears to capture the essential physics that leads to this distribution, it does not tell us precisely what determines the power-law exponents (about -2.25 for the bee networks). In the next section, we will try to answer that question, and test our answer using random-walk proximity networks obtained from the model above and from two other models derived from the model above.

8.3.3 Scaling argument for the duration distribution exponent

The two essential features of the model in subsection 8.3.2 that appear crucial to reproducing the duration distribution were (1) the agent's speed reducing as it approached another agent, and (2) heterogeneities in the amount by which the speed is reduced during proximity. Based on these facts, we now provide an argument for how the tail of the distribution Δt should scale. Consider the situation in which some agent

has "trapped" another, focal agent within its proximity, so that the other agent is slowing down. The agents' neighbourhood is characterized by some length scale L, and during an interaction that lasts a duration Δt, the focal agent on average has to traverse this length L. Thus its average speed should be $\bar{v} = L/\Delta t$. Now, in this equation, the length scale L is a fixed quantity, while $\bar{v}$ and Δt are stochastic quantities for each interaction, and their probability densities are related as follows:

$$P(\Delta t)\ \mathrm{d}\Delta t = P(\bar{v})\ \mathrm{d}\bar{v} \tag{8.4}$$

$$P(\Delta t) = \left|\frac{\mathrm{d}\bar{v}}{\mathrm{d}\Delta t}\right| P(\bar{v}) \tag{8.5}$$

$$= L(\Delta t)^{-2} P(\bar{v}) \tag{8.6}$$

$$\sim (\Delta t)^{-2}\left[A_0 + A_1\bar{v} + A_2\bar{v}^2 + \mathcal{O}(\bar{v}^3)\right], \tag{8.7}$$

where in the last step, we are expanding the probability distribution for $\bar{v}$ as a Taylor series about $\bar{v} = 0$.

Equation 8.6 suggests why the interaction duration distributions for both bees and humans scaled with an exponent close to -2. If the probability distribution $P(\bar{v}) \to A_0 \neq 0$, then we should get a distribution with exponent -2. However, if the leading-order term in the asymptotic expansion is non-constant, then $P(\Delta t)$ is modified. In the case where the linear term is the leading-order term, we would get

$$P(\Delta t) \sim (\Delta t)^{-2}\bar{v} \sim (\Delta t)^{-2}\frac{L}{\Delta t} \sim (\Delta t)^{-3} \tag{8.8}$$

In general, if in the limit as $\bar{v} \to 0$ we have $P(\bar{v}) \sim \bar{v}^{\beta}$ (for some $\beta \geq 0$), then we obtain

$$P(\Delta t) \sim [\Delta t]^{-(2+\beta)} \qquad \text{as } \Delta t \to \infty. \tag{8.9}$$

Note in particular that the exponent $(2 + \beta) \geq 2$; if $P(\bar{v})$ scales linearly as $\bar{v} \to 0$, then the exponent is 3, while if it converges to a nonzero constant, the exponent is 2.

8.3.4 Testing the scaling argument in simulation

We tested the scaling argument described in the previous subsection using three models, all based on the mobile-agent model of Starnini et al. reviewed in subsection 8.3.2:

(A) The mobile-agent model with *constant* attractiveness across individuals, i.e. $a_i = a$ for all agents i.

(B) The mobile-agent model with a_i drawn from a uniform distribution—this is just the original model.

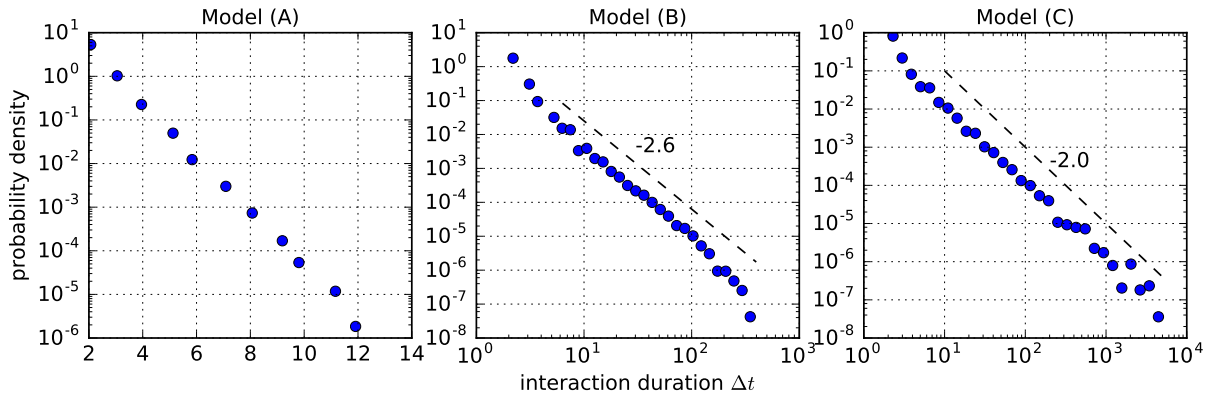

Figure 8.9: **Interaction duration distribution for each model.** The first distribution is exponential, while the other two are power laws. The dashed lines are added to guide the eye.

(C) The mobile-agent model with a_i drawn from a uniform distribution, but modified so that all but one of the agents are fixed in place. This model may be thought of as one in which a single agent moves in a space with randomly-placed "disorder" that slows it down. The interaction of the agent with the disorder is governed by Equation 8.3, as in the other models.

For each of these models, we performed simulations to measure the distribution of interaction durations and to measure the distributions of the average speed $\bar{v}$ during interaction. The average speed during a single, continuous interaction of a focal agent with any other agent was measured simply as the total number of steps taken by the focal agent, divided by the timespan of the interaction. Note that $0 \leq \bar{v} \leq 1$, since the step size is fixed (and set to 1).

Figure 8.9 shows the duration distribution obtained for each of the three models. The distribution for Model (A) decays exponentially and does not have a power-law tail, reflecting the earlier observation that the agents in the model have to be heterogeneous in order to obtain a fat-tailed distribution. Presumably the argument proposed in Equation 8.6 does not hold in this case, perhaps because the limit as $\bar{v} \to 0$ of $P(\bar{v})$ does not exist.

The distribution for Model (B) corresponds to the regular model, and this shows power-law scaling with an exponent higher in magnitude than 2. This suggests that $P(\bar{v}) \sim \bar{v}^{\beta}$ for some $\beta > 0$, and so the distribution must approach zero as $\bar{v} \to 0$. Finally the distribution for Model (C) shows power-law scaling with exponent very close to 2. This indicates that $P(\bar{v})$ tends to a non-zero constant as $\bar{v} \to 0$, so that by Equation 8.6, $P(\Delta t) \sim [\Delta t]^{-2}$.

Figure 8.10 shows histograms of $\bar{v}$ measured in simulation for each model, both over the entire range $\bar{v} \in [0, 1]$ and in the limit as $\bar{v} \to 0$. These histograms confirm the arguments made above about the

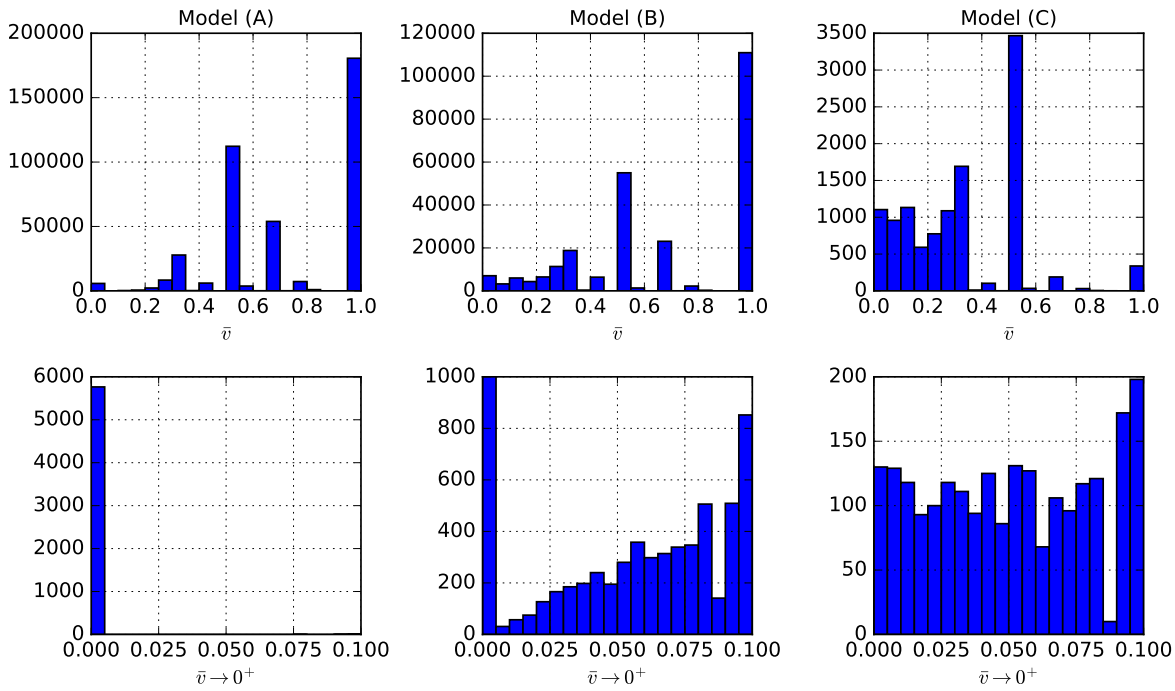

Figure 8.10: **Distribution of mean speed during interaction for each model.** Plotted on the first row are histograms for the distribution of $\bar{v}$ using bins of size 0.05. On the second row, we zoom in to the distribution near $\bar{v} = 0$, using bins of size 0.005.

asymptotics as $\bar{c} \to 0$. For Model (A), the sampled distribution does not appear to be smooth, which explains the lack of a heavy tail in the distribution. In Model (B), there is an isolated peak at $\bar{v} = 0$, but the distribution tends to zero in the limit as $\bar{v} \to 0^+$. The peak likely corresponds to short interactions in which the focal agents slows down due to another agent that has just appeared in its neighbourhood, but that the agent then quickly moves away. Finally, for Model (C), it is clear that the distribution is flat and non-zero in the asymptotic limit.

Recall that the scaling argument said that if $P(\bar{v}) \sim \bar{v}^\beta$ for some $\beta \geq 0$ in the limit as $\bar{v} \to 0$, then the duration distribution should follow the form $P(\Delta t) \sim [\Delta t]^{-(2+\beta)}$. In Model (B), the distribution tends to zero as $\bar{v} \to 0$, though we cannot measure the exact form of the asymptotics; if $P(\bar{v})$ is analytic at $\bar{v} = 0$, then the highest order term is linear. At any rate, we expect that the duration distribution should scale with some exponent of magnitude $(2 + \beta) > 2$ since $\beta > 0$ in this case. In contrast, in Model (C), $\beta = 0$ because the distribution of $\bar{v}$ is flat an nonzero as $\bar{v} \to 0$. This implies that $P(\Delta t) \sim [\Delta t]^{-2}$, which is what we see in simulation. Thus the scaling argument appears to work in these two cases, and provides a connection between the particular exponent observed for the duration distribution and the asymptotic form of the distribution of mean speed during interaction.

8.4 Discussion

In summary, we began the chapter by exploring whether honeybees actually form a network, that is, is there any structure in their interactions? The distribution of link weights suggests that there is: some pairs of bees interact far more frequently than the others, and this occurs despite the fact that worker bees do not recognize each other inside the hive. The also saw that unlike the degree distribution, the link-weight distribution is stationary.

Second, we explored the temporal structure of the duration of trophallaxis interactions. We found that surprisingly, bees share this distribution with human communication networks. We explored a model from the literature that captures the essential phenomena behind this distribution, and we extended our understanding of the phenomenology—in particular, the exponent of the distribution—using a scaling argument that relates duration Δt to mean speed $\bar{v}$ during interaction. This scaling argument says that the asymptotic form $P(\bar{v}) \sim \bar{v}^{-\beta}$ as $\bar{v} \to 0$ sets the duration distribution to be $P(\Delta t) = [\Delta t]^{-(2+\beta)}$. We tested this argument on simulation data for simple random-walk models, and found that the form of the measured $P(\bar{v})$ distribution does affect the exponent of the duration distribution. It will be interesting to also test this on simple random-walk models that begin to incorporate features of trophallaxis; one such class of models for ants has already been studied [53].

What does the scaling argument say for the exponents of the empirical distributions discussed in subsection 8.3.1? We found there that the bee trophallaxis duration distribution had an exponent that was smaller in magnitude than for face-to-face interactions in both bees and humans. We speculate that this difference arises from differences in the distribution of the mean speed during a trophallaxis interaction compared to that in a face-to-face interaction. Trophallaxis is qualitatively different in that when it occurs, the bees come to a complete halt and one bee may even lock its antennae with the other. This might change the asymptotic form of the distribution $P(\bar{v}) \sim \bar{v}^{\beta}$ such that β is close to zero, which pushes the exponent $(2 + \beta)$ of the duration distribution closer to 2. It would be interesting measure the distribution $P(\bar{v})$ using the tracking data, but at the present resolution of 1 second it is impossible to measure the asymptotics as $\bar{v} \to 0$.

Chapter 9

Rapid and robust spreading dynamics in the honeybee trophallaxis network

9.1 Introduction to temporal networks

9.1.1 Constructing temporal networks

A common method of studying social interactions is to create a network. The network representation is often chosen to be *static,* i.e. there is no time information encoded in the network. Individuals are represented by nodes in the network, and they are connected by an edge when they have some kind of social interaction. The edge may be directed if there is a corresponding asymmetry in the interaction; for example, if the social network corresponds to exchange of electronic mail, then the edge may be directed from sender to recipient.

A fundamental limitation of static networks in representing social interactions is the loss of time information. In order to study a time-dependent process occurring on a social network—such as the spread of information—the use of static networks can produce misleading results, as we discuss below. In order to address this shortcoming, temporal networks are used instead.

A natural way to extend static networks to temporal networks is by annotating the edges of a static network with time information. Specifically, each edge is annotated with a sequence of times for when that particular edge is activated (see Figure 9.1a). A second way of thinking of temporal networks is to picture the timeseries of contacts between nodes, by assigning each node its own timeline and marking contacts between timelines, as in Figure 9.1b. The two representations are logically equivalent but somewhat complementary in their utility, as their usefulness depends on the question being addressed.

Temporal networks respect the time-ordering of the underlying interactions between nodes, and this is especially important studying time-dependent processes occurring on the network, since temporal networks respect causality while the equivalent static network does not. Suppose we wish to understand if there is a path between nodes A and D in Figure 9.1. If we only considered the static network, then the answer is yes, because one can trace the paths A–B–D and A–B–C–D. Thus an information packet, for example, could propagate from node A to node D. However, the static network picture implicitly assumes that the

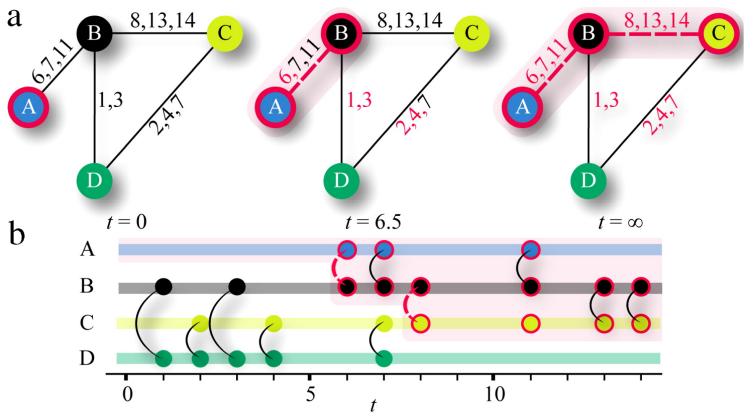

Figure 9.1: **Temporal network representations.** Two ways of thinking about temporal networks: (a) annotating each edge with a sequence of activation times; (b) picturing the timelines of each node and marking when a pair of nodes interact. Figure reproduced with permission from Ref. [56].

edges are available at all times, which is not the case in social interactions, where edges only have a transient existence: that is, the edge is activated for a certain period of time when the two nodes are interacting, and is inactive the rest of the time. In Figure 9.1, there is in fact no time-respecting path from A to D. This should be especially clear in the second representation, where we can trace an imaginary information packet travelling in the horizontal direction, branching out every time there is a contact between the node it is propagating along and another node.

In this chapter, we will use the framework of temporal networks to study the temporal properties of the honeybee trophallaxis network. The apparatus described in section 7.2 allows us to monitor all the individuals in honeybee colonies with 1-second time resolution over the course of about 10 days. The trophallaxis detection method described in subsection 7.3.1 is able to sample a large fraction of trophallaxis interactions that actually occur in the colony. This gives us a dataset of interactions that is unprecedented in its coverage of every individual in an animal society over an extended period of time. For comparison, we now review some existing studies of human social networks that have been analyzed as temporal networks. For a broad overview of the state of the art for temporal networks, we refer the reader to Holme and Saramäki's review article [56].

9.1.2 Review of human temporal networks

Temporal networks in human offer the opportunity to study sociological dynamics such as opinion formation [71, 7], the spread of rumours and fads [131], crowd dynamics [131], and disease spread. Electronic communication networks are one type of temporal network for which data exists. For electronic mail, networks have been constructed from e-mail log files from a university server [39, 122] as well as a commercial

dataset [122] and a viral marketing campaign [61]. These have been used to detect synchronous coherent structures in temporal activity [39] and the spread of computer viruses [122], among other aspects. Other communication datasets include those for mobile phone text messages [135, 139], phone calls [69] and instant messaging [72]. A common feature observed in these datasets is the burstiness of interactions, which we will discuss in more detail in section 9.3.

Another common way of constructing a network from human social interaction is based on *spatial proximity*: that is, an interaction is recorded when two individuals are near each other (and possibly facing each other). This is recorded either using mobile phones or providing participants with infrared- or RFID-based sensors that are activated when two participants are in close proximity to and facing each other. This has been used to study the predictability of interaction partners inside a company [117]. There are also two major projects that having been collecting human proximity data. The SocioPatterns project has collected several (ten at the time of writing) sets of human interactions in places such as schools, hospitals, technical conferences, workplaces, households and art galleries. These datasets are all published online, and have been used to study various aspects of networks, such as prediction of contact, assortative mixing (among genders), and the dynamics of processes occurring on the networks, such as random walks, diffusion, and spreading. An older project is the MIT Reality Mining project, which used cell phones to record human proximity interactions [38]. This has since been extended to a suite of projects under the title Reality Commons, which use the same technology to record human activity in different settings. In addition to these datasets, there also exist some idiosyncratic datasets that do not fall into these categories. One such dataset is the sexual network for internet-mediated prostitution, linking sellers and buyers of sex [99].

In some of these networks, space is an important aspect of the underlying interactions, but is disregarded in the creation of a temporal network (much like the way time is disregarded in creating a static social network). It is therefore important to keep in mind that the activation of edges is not just a time-dependent phenomenon, but also potentially depends on space—explicitly so, in the case of proximity networks. Our focus, however, is on the temporal aspects of the network, and in section 9.3, we study burstiness, a characteristic feature of social interactions. In the next section, we briefly review spreading processes, an important class of dynamical processes that occur on temporal networks.

9.2 Spreading processes and randomized reference networks

Epidemiologists have for a long time studied the spread of infectious disease through a population via social interactions. The history of mathematically modelling epidemic spreading processes dates all the way back

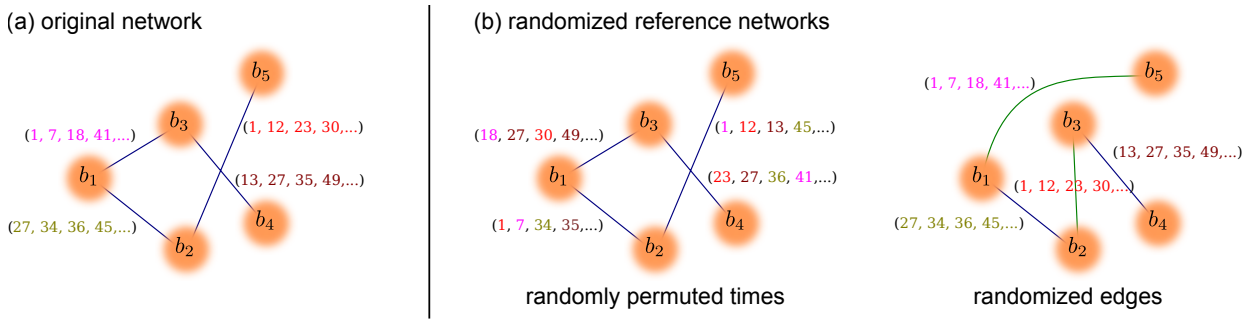

Figure 9.2: **Temporal network reference models.** (a) A temporal network represented as a regular network in which edges are annotated with timestamps. (b) In the randomly permuted times reference model, timestamps are permuted across all edges. In the randomized edges model, edges are rewired; shown here is the rewiring step for the edge (b_1, b_3), with the resultant rewired edges indicated in green.

to Daniel Bernoulli in 1760, who was trying to study the effects of immunization against the smallpox virus [55]. The framework of networks provides a natural setting in which to study epidemic processes, where each individual in a population is represented by a node in the network, and infections propagate along the edges.

One of the most commonly-used modelling frameworks is that of the *compartmental model,* where each node is in one of a discrete number of characteristic states. The simplest case is the so-called SI model, in which each individual is either susceptible (S) to infection, or has already been infected (I). Such a model can be extended in many ways; for example, a third, recovered (R) state could be included, to represent an individual that has recovered from the infection and is now immune to infection. Or an infected individual could recover but still be susceptible to a new infection (the SIS model). These models can be deterministic or stochastic, depending on the question being studied. Such models are of interest to physicists because they display phenomena that are similar to those seen in models of nonequilibrium statistical physics, such as phase transitions and macroscopic clusters [91].

In our work, we use the simplest, deterministic SI model operating on a temporal network. In this model, one node is initialized to the infected state I, while all other nodes are set to the S state. When the infected node comes into contact with a susceptible node, the infection propagates and replicates. This process repeats until an equilibrium is reached: if there are time-respecting paths from the initial infected nodes to every other node, then in the equilibrium state all nodes are infected. Otherwise, some fraction of nodes is infected in the equilibrium state while the rest remain susceptible.

We use this model as a means to probe the temporal structure of the network. In order to make sense of the dynamics of spreading observed on a real temporal network, we compare the dynamics against that on *randomized reference networks* (sometimes known as configuration networks). A randomized reference

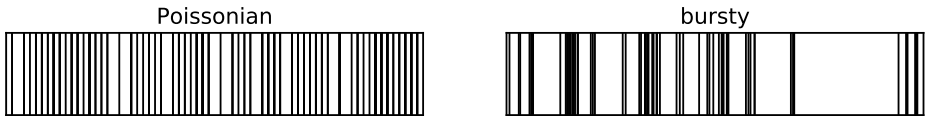

Figure 9.3: **Poissonian vs. bursty signals.** Simulated timeseries of activation events for a Poisson distribution of interevent durations τ and for a power-law distribution $P(\tau) \sim \tau^{-2}$.

network is created by taking the real temporal network and randomizing one or more particular features of the real network. By comparing the properties of (or dynamics on) the real network against those of the reference network, we can deduce the effect of the particular feature that is randomized on the properties (or processes) under study.

There are several different randomized reference models that may be used for this purpose [56]. Figure 9.2 shows the basic idea behind the two reference models that we adapted for use in our study (see subsection 9.4.1). Here we think of a temporal network in terms of edges annotated with timestamps. In the *randomly permuted times* model, the set of timestamps across all edges is permuted (while retaining the same number of timestamps on each edge). This model destroys the detailed ordering of events, but keeps intact the static network structure, the number of contacts per edge, and the overall times of activity for the network as a whole. A second reference model that we consider is the *randomized edges* model, which is constructed as follows: for each edge (i, j) in the network,

1. select another edge (i', j') uniformly at random,
2. rewire the edges into (i', j) and (i, j') with probability 1/2, or into (i, i'), (j', j'),
3. if step 2 created a self edge or a multi-edge, then undo and start over from step 1.

This randomization preserves timestamps on edges while rewiring the network, effectively randomizing the topology of the aggregate network. Although this randomization preserved the temporal properties of edges, it does not preserve the temporal properties of nodes. For instance, in the next section, we consider the distribution of times separating node activation events; this distribution is not conserved under randomized edges. On the other hand, the distribution of times separating edge activation events is preserved.

9.3 The waiting-time distribution and burstiness

Human social interactions are characterized by burstiness [56], or intermittency: there are bursts of interactions, in which subsequent interactions are separated by a relatively short duration, and these bursts themselves are separated by longer durations (see Figure 9.3). Burstiness manifests itself in the probability

distribution $P(\tau)$ of durations τ between events for a give node, which we call the waiting-time distribution, which spans a wide range of duration scales and may have a fat tail. One proposed measure of the degree of burstiness is the burstiness coefficient [50], which compares the mean μ_τ and the standard deviation σ_τ of the distribution:

$$\text{burstiness coefficient} \qquad B \equiv \frac{\sigma_\tau - \mu_\tau}{\sigma_\tau + \mu_\tau}. \tag{9.1}$$

This quantity is well defined for real-world signals in which the sample mean and sample standard deviation are well defined, and $B \in (-1, 1)$. The case $B > 0$ corresponds to burstiness, with higher values indicating a greater degree of burstiness (σ_τ is large relative to μ_τ). The case $B = 0$ corresponds to a neutral, Poissonian distribution, while $B < 0$ measures the regularity of the signal, with $B \to -1$ corresponding to a perfectly regular signal, i.e. $P(\tau)$ converges to a delta function.

In a temporal network, burstiness usually refers to the waiting-time distribution of node activation times (rather than edge activation times, which is sometimes distinguished as the interevent distribution). So how does this property of interactions affect the dynamics of spreading processes occurring on top of the network? In human temporal networks, burstiness is generally associated with a slowdown of spreading: that is, relative to a network that has been randomized using methods such as those described in section 9.2, spreading occurs slower in the empirical network than in the randomized network. This has been seen in the temporal networks for email activity, where bursty interactions lead to increased prevalence times for the outbreaks of computer viruses [122]. Similar effects have been observed in mobile phone call networks [69, 70], as well as in networks of ants making antennae-to-body contacts [15]. An exception is the sexual contact network, where simulated spreading was faster in the empirical network than on randomized reference models [100]; however, this network is somewhat different than the others mentioned above, because there is turnover among the nodes and there is a strong underlying geographic structure that partitions the network [99].

In general, the relationship between burstiness and spreading processes occurring on the network is not well understood [56, 57]. It is likely that the waiting-time distribution is not the only factor that determines the speed of spreading on the network (the network topology also modulates spreading), and analytic work has shown that if the processes underlying the activation of edges is non-stationary, then it is possible to observe a speedup of spreading (relative to a Poissonian interevent distribution) even with burstiness [57].

Figure 9.4 shows the distribution of waiting times for all five datasets from 2013. We see that the waiting-time distribution is broadly distributed, and follows a power law form (with exponent ≈ -0.83) for almost four orders of magnitude on the abscissa, and this is followed by a cutoff. Thus honeybee trophallaxis interactions share the feature of burstiness with human interactions. However, as we see in the next section, the presence of burstiness in the interactions does not slow down spreading in the honeybee trophallaxis

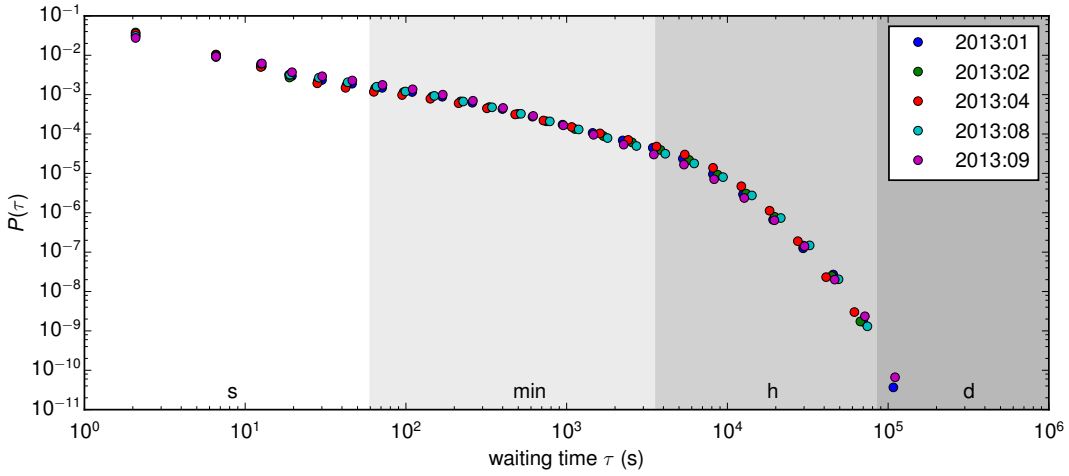

Figure 9.4: **Waiting-time distributions for trophallaxis interactions.** The distribution is estimated using logarithmically-spaced bins. The power law scaling occurs over almost four orders of magnitude on both axes.

network.

9.4 Rapid spreading dynamics on bee trophallaxis networks

We have seen that the honeybee trophallaxis network shares the temporal feature of burstiness with human communication networks. In this section, we study the dynamics of spreading on the trophallaxis network under the deterministic SI model as a means to further probe its temporal structure. We find that surprisingly, there is a speedup of spreading in the trophallaxis network despite the bursty nature of the interactions. We quantify this speedup and study how robust it is to both the real demographic perturbation of forager removal and to *in silico* removal of interactions from the trophallaxis network.

9.4.1 Choice of reference models

We ran spreading simulations on each dataset from 2013. For comparison, we created two ensembles of randomized reference networks:

- temporally randomized networks, generated using the randomly permuted time (RPT) model described in section 9.2.
- temporally and topologically randomized networks, generated by first implementing the randomized edges (RE) model and then applying the RPT model; we call this the RE+RPT ensemble.

In the basic form of the RE and RPT models, the randomizations are uniform: in the RPT case, the timestamps are permuted uniformly at random across edges, while in the RE case, each edge is rewired to another edge chosen uniformly at random. However, in the case of bee networks, the number of nodes is not constant in time, but is a monotone decreasing function due to mortality and the perturbation during the experiment (see e.g. Figure 7.4). In order to account for this, the RPT randomization was modified to ensure that in the permutation, a timestamp was not assigned to an edge joining two nodes if one of those nodes was already dead.

For each empirical trophallaxis network, an ensemble of 100 RPT networks and an ensemble of 100 RE+RPT networks were created using this procedure. We could then statistically compare any property of the empirical network again the equivalent property in each ensemble of randomized reference networks; this is sometimes known as a *conditional uniform graph (CUG) test*.

These particular models were chosen because the RPT model allows us to study the effect of temporal correlations and heterogeneities on the dynamics of spreading. The RE+RPT model tells us what, if any, additional effect the topology of the network has on spreading dynamics.

9.4.2 Spreading dynamics on trophallaxis networks

We used the deterministic SI model to simulate spreading on the empirical and randomized networks. A single simulation run consisted of initially setting all bees to the susceptible state, choosing an interaction uniformly at random, and setting the two bees involved in that interaction to the infected state. SI spreading dynamics were then simulated over a 10 h window of time (long enough for all the bees to be infected). The dynamics were measured by the prevalence $p(t) = n(t)/N(t)$, where $n(t)$ is the number of infected bees at time t after the first infection, and $N(t)$ is the total number of bees alive at that time. Note that due to mortality, $N(t)$ is monotone decreasing, but this means that $p(t)$ can either increase or decrease.

The deterministic SI model essentially provides an upper bound on the rate of spreading via trophallaxis interactions. To estimate this bound, we repeat each SI simulation run 1000 times for each network, with different initial conditions, and average over these to obtain the mean prevalence $\bar{p}(t)$ at time t after the infection. Each set of simulation runs was performed for both the empirical trophallaxis network and for each network in each randomized ensemble.

Figure 9.5a shows the prevalence curve $\bar{p}(t)$ for the empirical network for dataset 2013:01 and for each of the 100 networks in each randomized ensemble; the corresponding figures for the other datasets are in Appendix B (see Figure B.1). We see that spreading is faster on the empirical network than on the randomized networks for most of the spreading process. Following Ref. [69], we use the time $t_{20\%}$ it takes

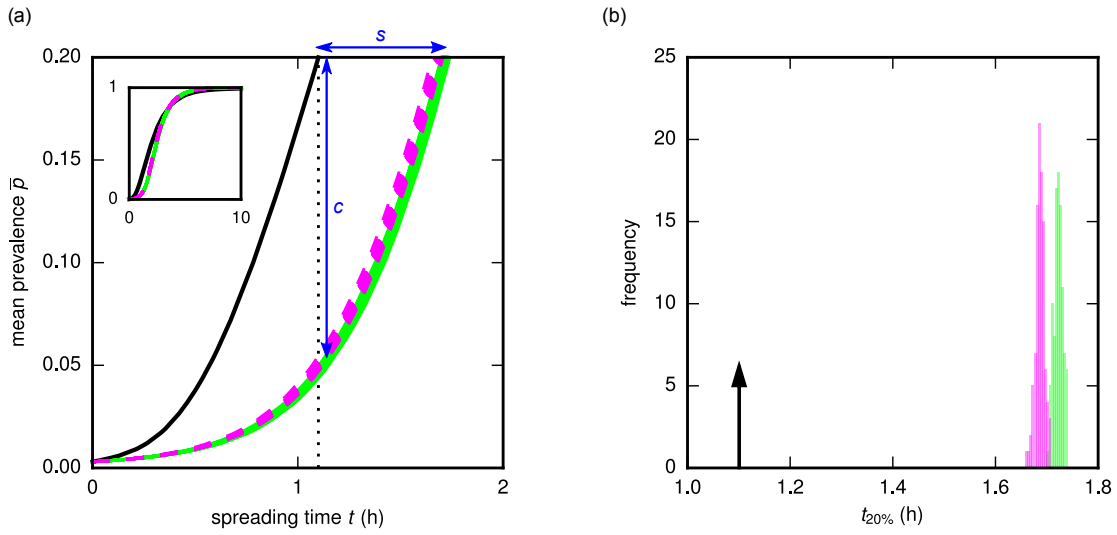

Figure 9.5: **Prevalence curves for empirical and randomized networks.** (a) Prevalence curves, until 20% prevalence, for the empirical trophallaxis network (black) and 100 randomized networks in each of the RPT (magenta) and RE+RPT (green) ensembles. The speedup metrics s and c are designed to measured the lengths indicated in blue. (Inset: complete prevalence curve over 10 h time period.) (b) Time at 20% prevalence for the empirical network (black arrow) and for the RPT (magenta histogram) and RE+RPT (green histogram) ensembles.

to reach 20% prevalence as a measure of the speed of spreading. Figure 9.5b compares the empirical value for $t_{20\%}$ against the set of values obtained for each randomized ensemble. This serves to show the degree of the speedup in spreading (which is quantified by s, defined below), and also confirms that the speedup is statistically significant (p-value $p < 0.1$ in a CUG test to reject the null hypothesis that $t_{20\%}$ is the same for the empirical and randomized networks).

We quantify the speedup of spreading in the empirical network relative to the RPT ensemble of randomized networks by defining two quantities. First, the speedup in terms of the time to 20% prevalence:

$$\text{speedup} \qquad s \equiv s_{20\%} = \left\langle \frac{t_{20\%}^{(k)}}{t_{20\%}^{emp}} - 1 \right\rangle_k \tag{9.2}$$

where $t_{20\%}^{emp}$ is the time at which the prevalence of the empirical network reaches 20%, and the average is taken over the set $\{t_{20\%}^{(k)} : k = 1, \ldots, 100\}$ of corresponding times for the ensemble of RPT reference networks. We also measured the speedup in term of the prevalence ratio, defined by

$$\text{prevalence ratio} \qquad c \equiv \left\langle \frac{0.2}{\bar{p}_k(t_{20\%}^{emp})} \right\rangle_k , \tag{9.3}$$

where $\bar{p}_k$ is the prevalence curve for the kth network in the randomized ensemble. What these quantities

dataset	speedup s	prevalence ratio c	crossover prevalence	burstiness B	exponent α
2013:01	0.53	3.99	0.83	0.33	−0.83
2013:02	0.44	3.21	0.76	0.32	−0.85
2013:04	0.24	1.92	0.72	0.27	−0.82
2013:08	0.57	4.21	0.84	0.34	−0.86
2013:09	0.56	4.06	0.88	0.39	−0.87

Table 9.1: **Measurements of spreading speed and burstiness** Listed are two metrics s and c of the speedup of spreading; the prevalence at the crossover point between the empirical and the average RPT prevalence curves; the burstiness coefficient B and the exponent of the power-law fit to the waiting-time distribution.

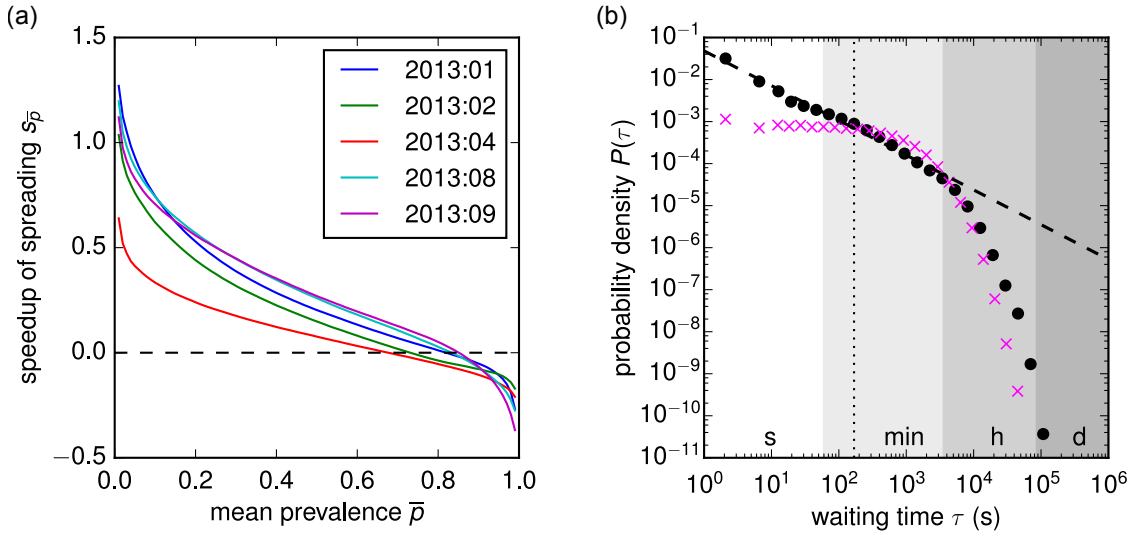

Figure 9.6: **Speedup as a function of prevalence and the destruction of burstiness.** (a) Speedup measured at different prevalence values. (b) The waiting-time distribution for the empirical trophallaxis network (black balls) and for the RPT networks (magenta crosses). The latter is approximately exponentially distributed, showing that the waiting times for interactions in the RPT networks are not bursty, but Poissonian.

measure is shown schematically in blue in Figure 9.5a.

We found that the speedup of spreading is consistent across datasets (see Table 9.1), with a mean speedup of $\langle s \rangle = 46.7\% \pm 13.7\%$ in the empirical network relative to the RPT network. Spreading was faster until a majority of individuals were reached in each case, as the crossover values in Table 9.1 show. We also measured the speedup as a function of the prevalence, that is, $s_{\bar{p}} = \langle t_{\bar{p}}^{(k)} / t_{\bar{p}}^{emp} - 1 \rangle_k$ (see Figure 9.6a). This also shows that there is a speedup until a majority of individuals are reached.

As Figure 9.5 shows, the additional topological randomization (i.e. the second RE+RPT ensemble) has the effect of slowing the spreading further, but the magnitude of the slowdown is small (on average, $\langle s \rangle = -1.9\%$ for the speedup of the RE+RPT networks relative to the RPT networks). This suggests that although topology does play a role in the enhancement of spreading, the predominant driver appears to be

the temporal structure that is destroyed by the RPT randomization. Burstiness is one temporal feature that is destroyed by the RPT randomzation process (see Figure 9.6b), so unlike human communication networks, we observe a speedup in the spreading dynamics *despite* the burstiness of interactions.

9.4.3 Resilience of the speedup of spreading

To test if the speedup persists on networks of shorter timescales, and to see if the speedup of spreading is resilient to the forager removal perturbation, we repeated spreading simulations on a temporal network constructed from the day after the forager removal experiment (24 h window starting at sunrise), and on an ensemble of RPT networks obtained by randomizing the empirical network. We observed that spreading continued to be faster on the empirical network relative to both the RPT and RE+RPT networks.

Given the dynamics underlying interaction activity, it is clear that the dynamics of spreading depends on the time period over which the network is constructed. The result discussed above suggests that the speedup of spreading can occur even on the short time window of 24 h. A future direction for exploration is to study the dependence of spreading properties as the time window is varied.

9.4.4 Rescaled dynamics in thinned networks

Due to the errors in trophallaxis detection discussed in subsection 7.3.1, and in particular the false negatives, the set of interactions that we measured is really a subset of the interactions that actually occurred in the colony. (False positives are a separate issue, but as discussed, in subsection 7.3.1, the post-processing step in the trophallaxis detection procedure greatly reduces the false positive rate, at the expense of increasing the number of false negatives.) To study how this subsampling of interactions affects our conclusions about the speedup of spreading, we performed the following numerical experiment.

Using the empirical trophallaxis network for dataset 2013:08, we sampled a fractions $\rho \in \{0.2,\ 0.4,\ 0.6,\ 0.8\}$ of the empirical interactions to create four "thinned" networks. We then repeated our analyses on each of these thinned networks, first creating ensembles of RPT and RE+RPT networks for each thinned network, and then running spreading simulations with the same protocols describe above.

In order to compare dynamics across the thinned networks, we used the following rescaling of time. Suppose the mean interaction rate in the original network is λ (i.e. simply the total number of interactions divided by the total time over which the temporal network is constructed). The subsampling process lowers the interaction rate in the thinned network to $\lambda_\rho = \rho\lambda$, and we expect that the quantity $1/\lambda_\rho$ sets the timescale for dynamics on the network. Therefore we plot the prevalence curves for the thinned networks as a function of the dimensionless time variable $\lambda_\rho t$. To verify the assumption that $1/\lambda_\rho$ alone sets the

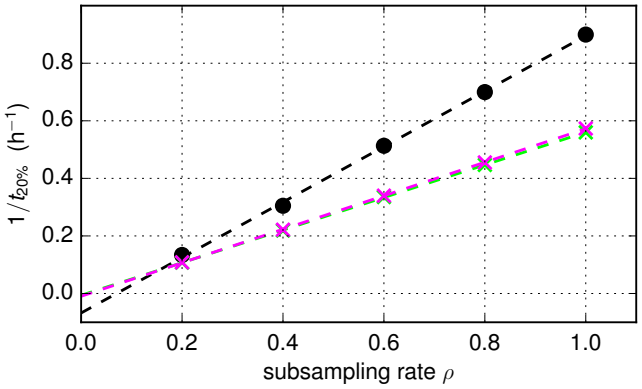

Figure 9.7: **Scaling of $t_{20\%}$ with subsampling rate.** To verify that (the reciprocal of) $\lambda_\rho \propto \rho$ sets the timescale for spreading, we see how $1/t_{20\%}$ changes with ρ, for the empirical network (black balls), the RPT networks (magenta crosses) and the RE+RPT networks (green crosses).

timescale for spreading dynamics, we plotted the reciprocal of the time to 20% prevalence against ρ (which is proportional to λ_ρ), and found that it scales linearly for both the empirical and randomized networks, as expected (see Figure 9.7).

We find that surprisingly, the rescaled prevalence curves for $\rho \in \{0.6,\ 0.8,\ 1\}$ collapse on top of each other while those for $\rho \in \{0.2,\ 0.4\}$ are closer to the randomized curves but still show a speedup of spreading (see Figure 9.8a). This suggests that the spreading dynamics are largely unaffected by the thinning process. On the one hand, this is good news for us in terms of addressing our worry about false negatives in trophallaxis detection, for this suggests that we are accurately capturing spreading dynamics on the network even though we only have a subset of the total interactions in the colony. It is, on the other hand, very interesting that the temporal structures responsible for the speedup in spreading appear to be still present even when only 20% of the interactions are retained. An interesting avenue for further exploration would be to subsample nodes rather than interactions, and study how the dynamics are affected by this version of subsampling of the network. Nodes could be either by subsampled uniformly at random, or the network topology could be perturbed on the basis of some static network property, such as the node strength (see section 8.1).

9.5 Discussion

We have discovered that although bees, like human, share the property of burstiness in their social interactions, the temporal network formed by these interactions is able to support faster spreading dynamics (under the SI model) relative to randomized reference models that destroy both temporal structure—including burstiness—and topology. The thinning analysis described in subsection 9.4.4 suggests that the temporal

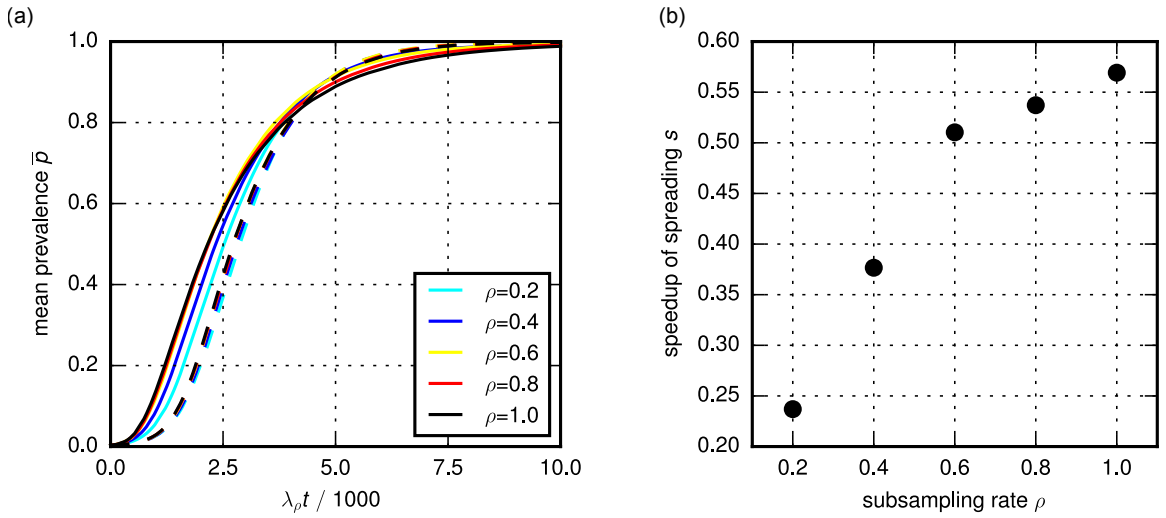

Figure 9.8: **Spreading on thinned networks.** (a) Prevalence curves in dimensionless time for the thinned networks. The solid prevalence curves are for the empirical thinned networks, and the dashed curves are the mean prevalence curves on the RPT ensemble. (b) Variation in speedup with subsampling rate ρ.

structures responsible for the speedup of spreading persist even when many interactions are removed uniformly at random from the network. We also observe the speedup of spreading on shorter time windows, including on the temporal network constructed from interactions aggregated over a 24 h time window following a demographic perturbation. We speculate that this resiliency is in part rooted in individual anonymity in the hive, as individual worker living in large colonies do not appear to recognize each other as individuals. This means that bees, unlike humans, may not seek out a particular interaction partner, but instead interact opportunistically, meaning that there is no single point of failure at the collective level.

Chapter 10

Conclusions and outlook

In Part I, we studied the relationship between the social caste of honeybees and the community of bacteria living in their guts. We showed that the microbiomes of queens and males—which had never been characterized before—shared the same bacteria as those in worker microbiomes, but the compositions were different enough that we can generally distinguish both drone and queen microbiomes from forager or nurse microbiomes. On the one hand, these differences are surprising given that the different social castes mix freely inside the nest, exchanging food with each other. On the other hand, the two reproductive members of the colony—males and queens—have different physiologies than workers, and our results suggest that these differences are sufficient to alter the composition of their microbiomes.

In Part II, we looked at the statistics of flight activity among foragers in honeybee colonies. We found that activity is not distributed evenly across different bees, but instead, the distribution is highly skewed, with a small number of individuals performing a large fraction of foraging flights. We quantified this distribution using the framework of Lorenz curves and Gini coefficients. We also studied the time-variation of lifetime flight activity, and found that there is no fixed developmental progression in activity as foragers grow older. Finally, we found that there is a reserve workforce of low activity foragers that increase their activity in response to a demographic perturbation in which foragers are removed from the colony. This confirms that the foraging workfoces behaves in a coordinated fashion, with individuals regulating their in response to the collective needs of the colony.

In Part III, we described a system to track individual bees as they go about their lives in a two-dimensional hive. We studied behaviours that can be derived from the tracking data, and used these to classify worker bees by their social caste. We also measured the diurnal rhythms in activity and saw that they are resilient to the demographics perturbation of forager removal. We then studied the structure of interactions based on the behaviour of trophallaxis. We used the framework of both static, time-aggregated networks and temporal networks. We found that trophallaxis interactions occur in a bursty manner, and that the durations of interactions follow a power-law distribution, just like in human face-to-face networks. We derived a scaling argument that explains the exponent of this distribution. Finally, we studied the dynamics of spreading

processes on trophallaxis networks, and discovered that surprisingly, the networks are structured in way that speeds up spreading. This is in contrast to networks of human communication, even though the networks for both species share the feature of burstiness.

The common thread that runs through these different projects is that they all leverage biological data obtained using cutting-edge technologies, and try to make quantitative statistical measurements, predictions and models. We saw that the high-dimensional sequence data obtained at the microbial level can be compressed into clusters of OTUs, and that these can be used to make predictions of social caste at the level of individual bees. The automated flight-monitoring technology allowed us to sample the distribution of flight activity among bees, going beyond simple anecdotal statements, perhaps saying in this case that 20% of the bees perform 80% of the flights. And the bCode monitoring system allowed us measure timeseries and several distributions, and to discover an interesting property of the honeybee temporal network that would have been impossible to guess from manual observations alone.

It seems that the march of technological progress will continue, and perhaps in a decade or two, there will be no need to distinguish quantitative biology as a separate subfield of biology, because the adjective will have become redundant. As we acquire more and more biological data, we will also need to update and improve our methods to analyze the data, and we will need to develop new methods to integrate disparate kinds of data.

In the realm of microbiomes, many challenges remain. We are now moving to a time when it is increasingly common to obtain longitudinal microbiome data, which reflect on the underlying non-equilibrium ecological dynamics of the microbial community. Moreover, metagenomics still does not know how to cope with the truly unknown bacterial strains: most studies take OTUs and map them to a database of known bacterial phylotypes. This does not help improve our understanding of the unknown taxa that are not present in the database. These issues are complicated further by horizontal gene transfer among bacteria, making it difficult to truly identify bacterial species in an ecological sense.

What about future directions enabled by the tracking system discussed in Part III? An important next step will be to study the honeybee social network formed by multiple channels of communication, rather than just trophallaxis. We discussed methods that can be used to detect dances and trips, and these could be integrated with trophallaxis data to obtain a more highly-resolved representation of the social network; indeed, experiments are already under way to improve the trip detection method by including an additional camera at the hive entrance. It may also be possible to extend the forager identification method discussed in chapter 7 such that each bee is assigned an instantaneous state. If such a well-defined state exists, then there are rich possibilities for studying the relationship between bee states and the colony state, and to measure

and model transitions in both bee states and in colony states in response to perturbations such as forager removal. Another important avenue for future exploration is looking at the response of the honeybee society to a modification to some of its individuals. We have already seen one such modification in Part III, when foraging individuals were removed from the society, but one can imaging milder modifications, in which the individuals still remain in the society but behave differently in some sense.

While all our studies used the honeybee as a model social organism, the issues studied here are far from irrelevant to human societies. The study of the human microbiome is exploding with association studies between the microbiome and health. In some limited areas, such as *C. difficile* infections, microbiome-based treatments have already proven effect. More generally, the field of personal medicine is being driven by the collection of individual-level data, following the same trends as in quantitative biology. Human social networks, meanwhile, might seem a lot more complex than those of the honeybee, but we have already seen in this dissertation that the two share some features, such as bursty interactions. Given the issues surrounding privacy and the ownership of personal data for humans, it seems all the more important to understand how much the social network can tell us about the individual.

Technological advances in data acquisition and computation also mean that divisions between different fields of scientific study appear to be more blurry, and a multidisciplinary approach often seems most effective in tackling many modern-day scientific problems. On a more personal note, I consider myself very fortunate to have had the opportunity at this time to work alongside scientists of different backgrounds, and I've come to appreciate more deeply a "problems-first" approach to science, rather than a "tools-first" approach. Of course, the two aren't exactly mutually exclusive, and there has to some compatibility between problems and tools for there to be progress. A multidisciplinary approach improves the chances that scientific problems are matched with the right tools, and perhaps this will come increasingly common in the years to come.

Appendix A

Timeseries of colony activity

In section 7.4, we discussed the measurements of colony activity for one dataset (2013:01). In this appendix, we include plots of the same measurements for the other four datasets from 2013. As a living system far from equilibrium, a honeybee colony undergoes dynamics that are never in a steady state, and are different across different datasets due to changes in the environment, such as season of the year, weather, food availability, etc. Comparing timeseries across datasets helps build an intuition for which features are intrinsic to the society, and which features are non-replicable. For example, in Figure 7.10 and Figure 7.12, there is a sharp drop in activity as soon as the forager-removal perturbation begins. This may seem to make sense and suggest further questions, e.g. what are the timescales for the response and for recovery? However, we see in the figures in this section that that particular feature is not actually replicated across datasets, and therefore does not appear to be an intrinsic feature of the bee colony.

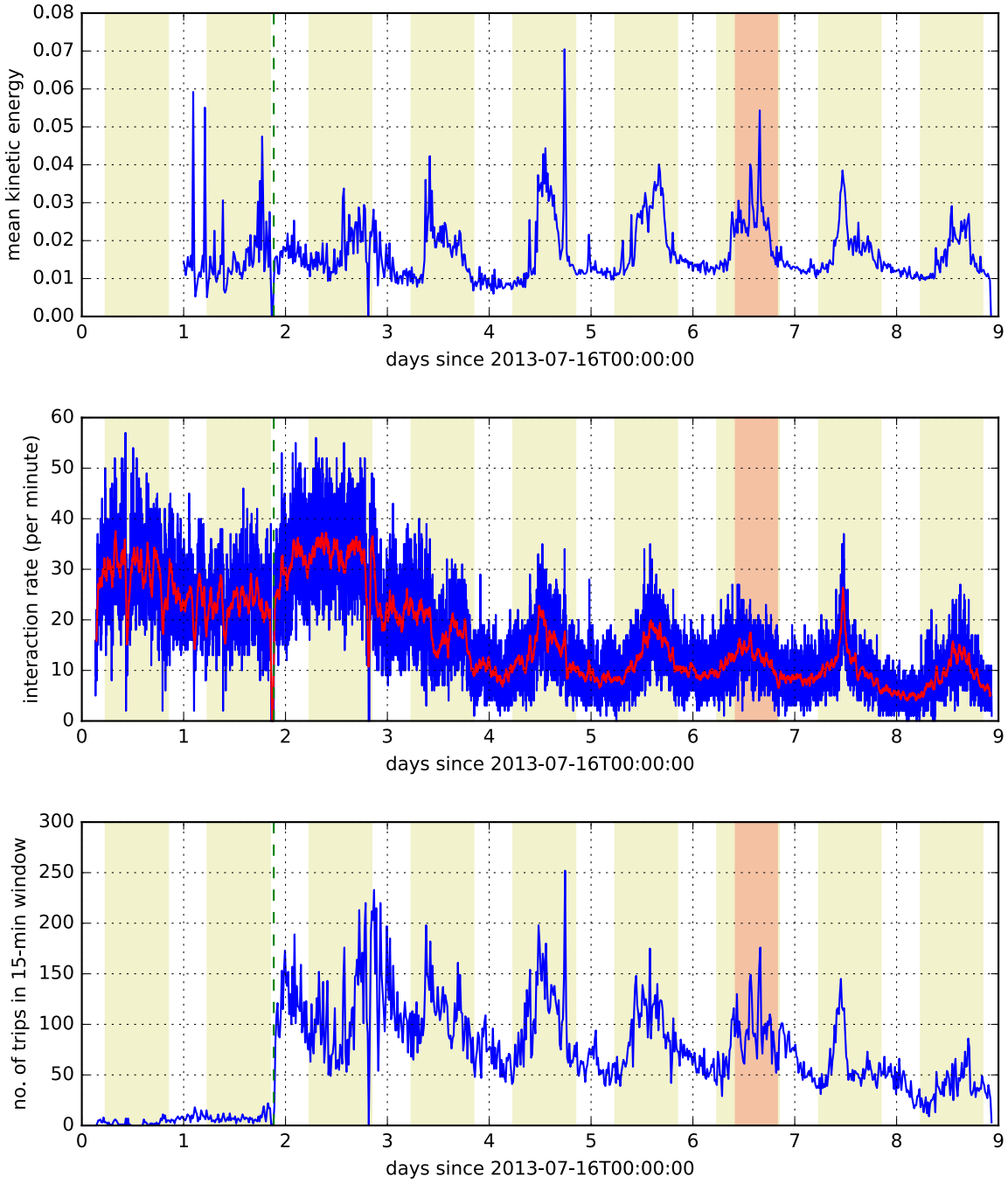

Figure A.1: **Activity timeseries for dataset 2013-02.**

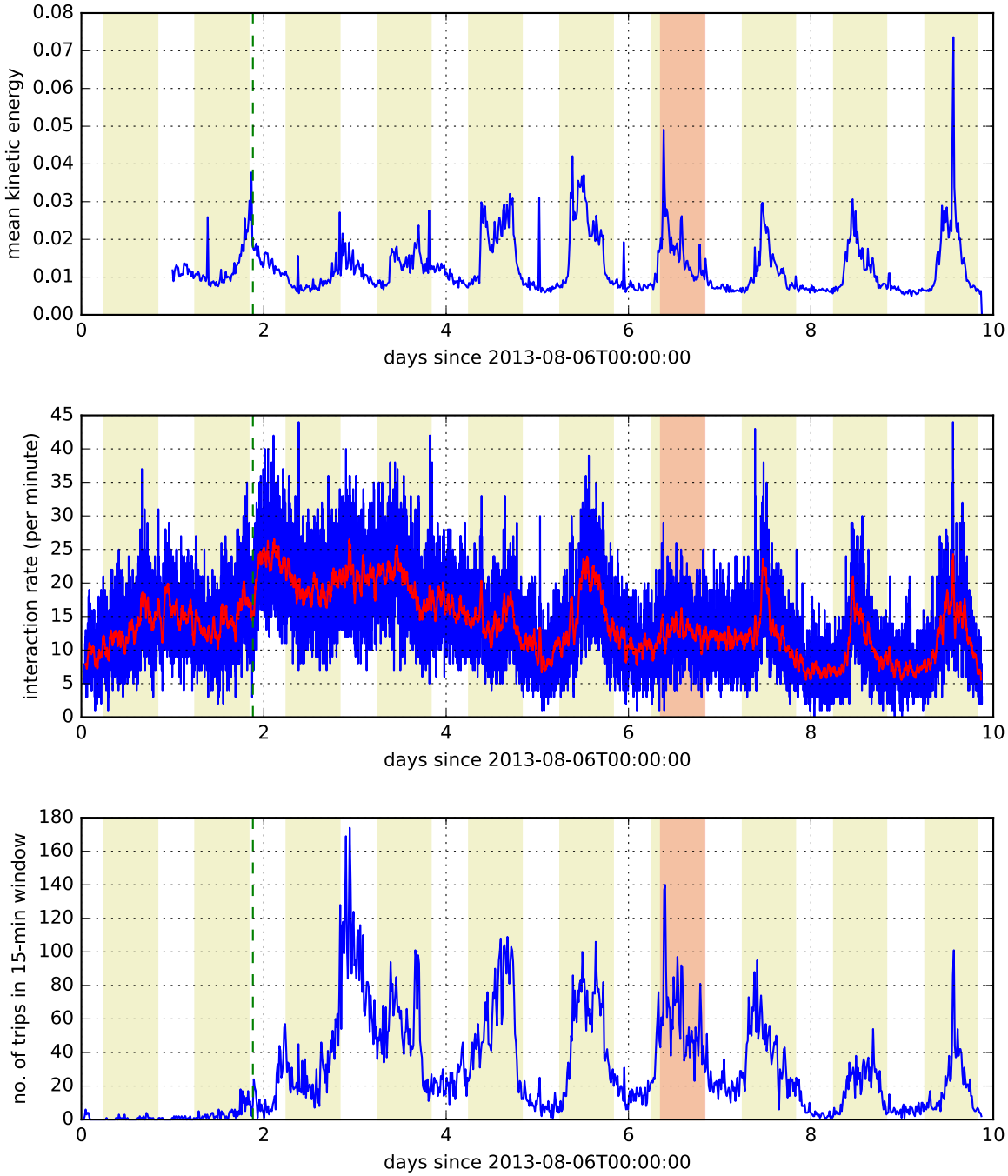

Figure A.2: **Activity timeseries for dataset 2013-04.**

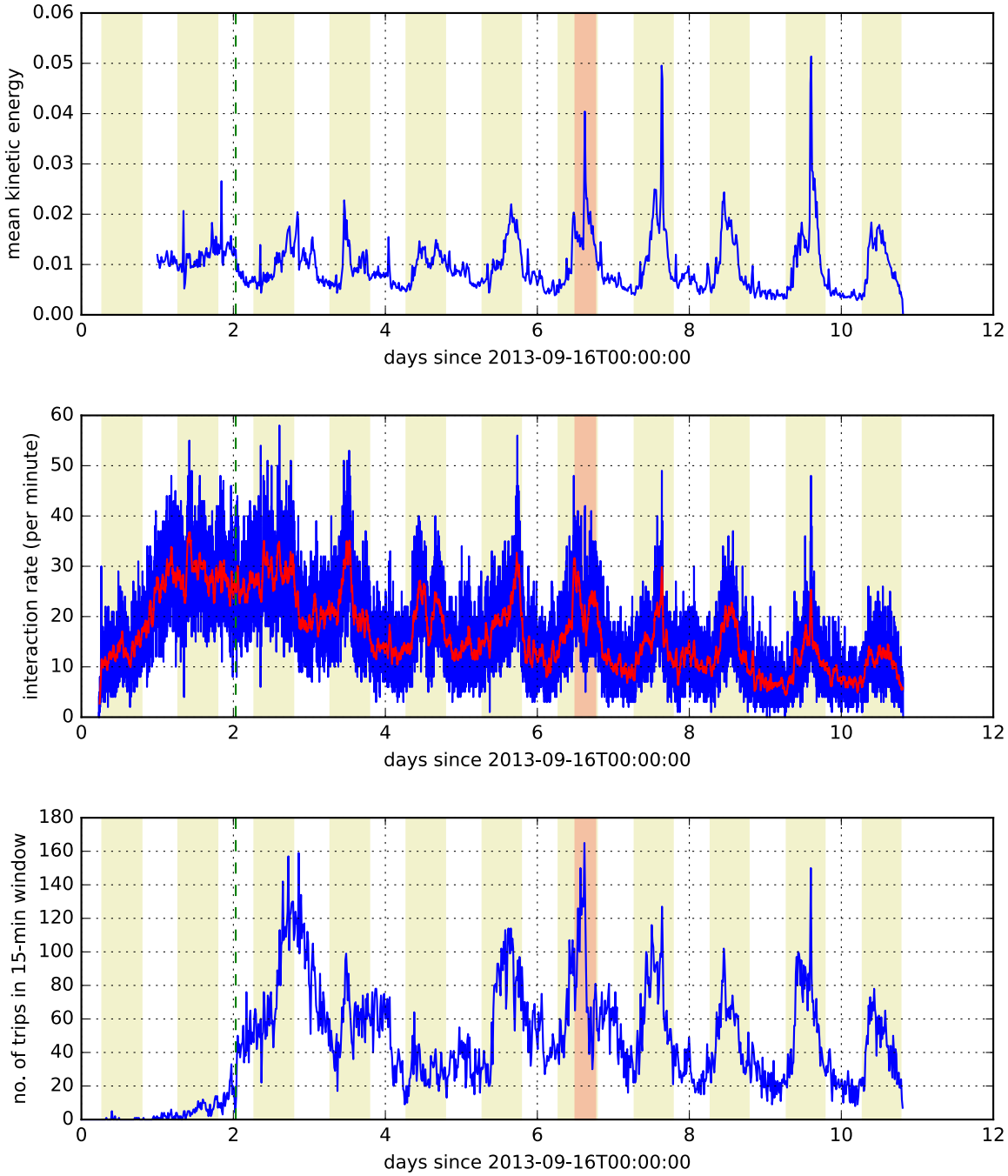

Figure A.3: **Activity timeseries for dataset 2013-08.**

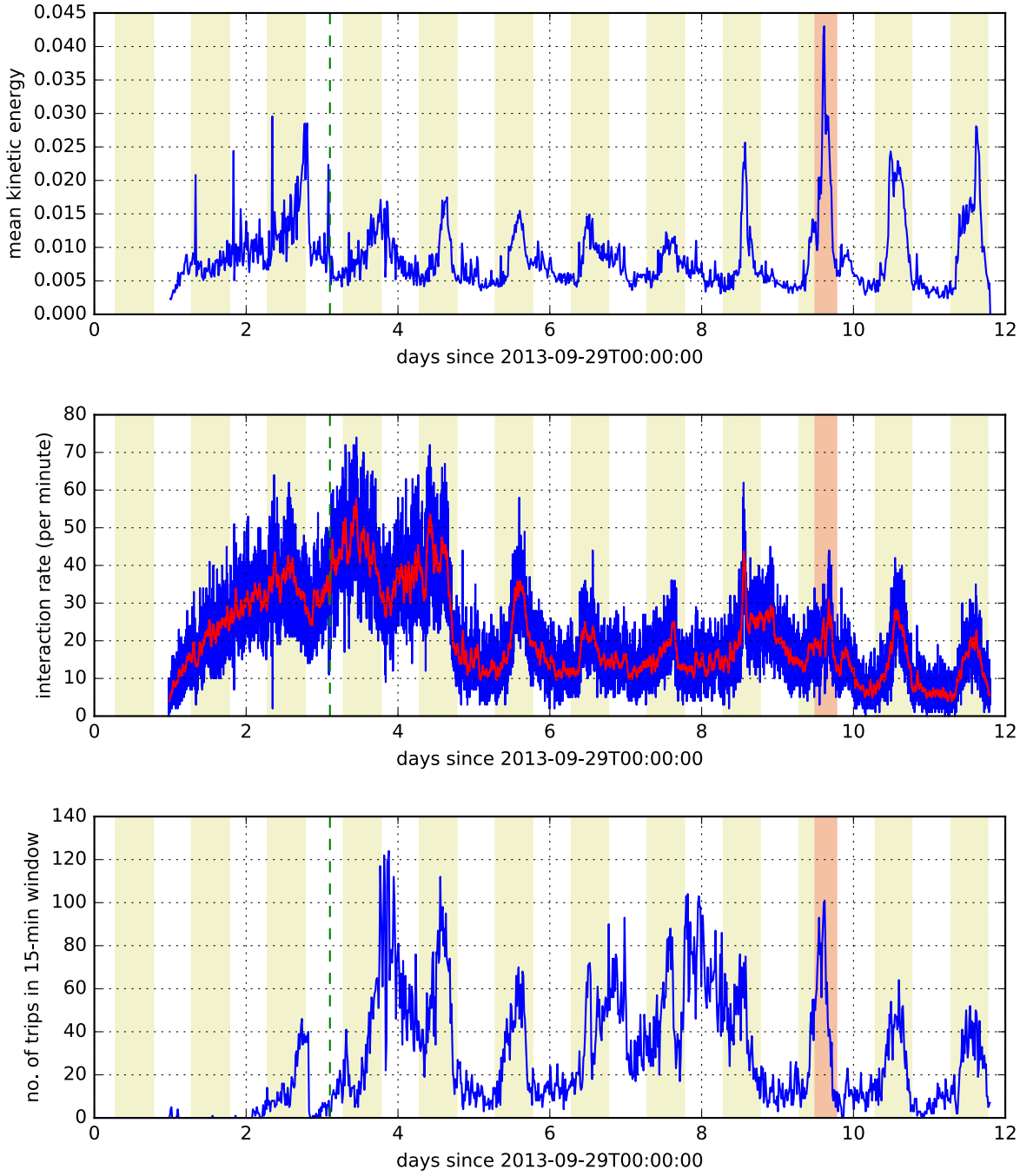

Figure A.4: **Activity timeseries for dataset 2013-09.**

Appendix B

Supplementary material on spreading dynamics

Figure B.1 shows the prevalence curves for the four other datasets, similar to Figure 9.5 in the main text. In all cases, there is a speedup of spreading, as quantified in Table 9.1.

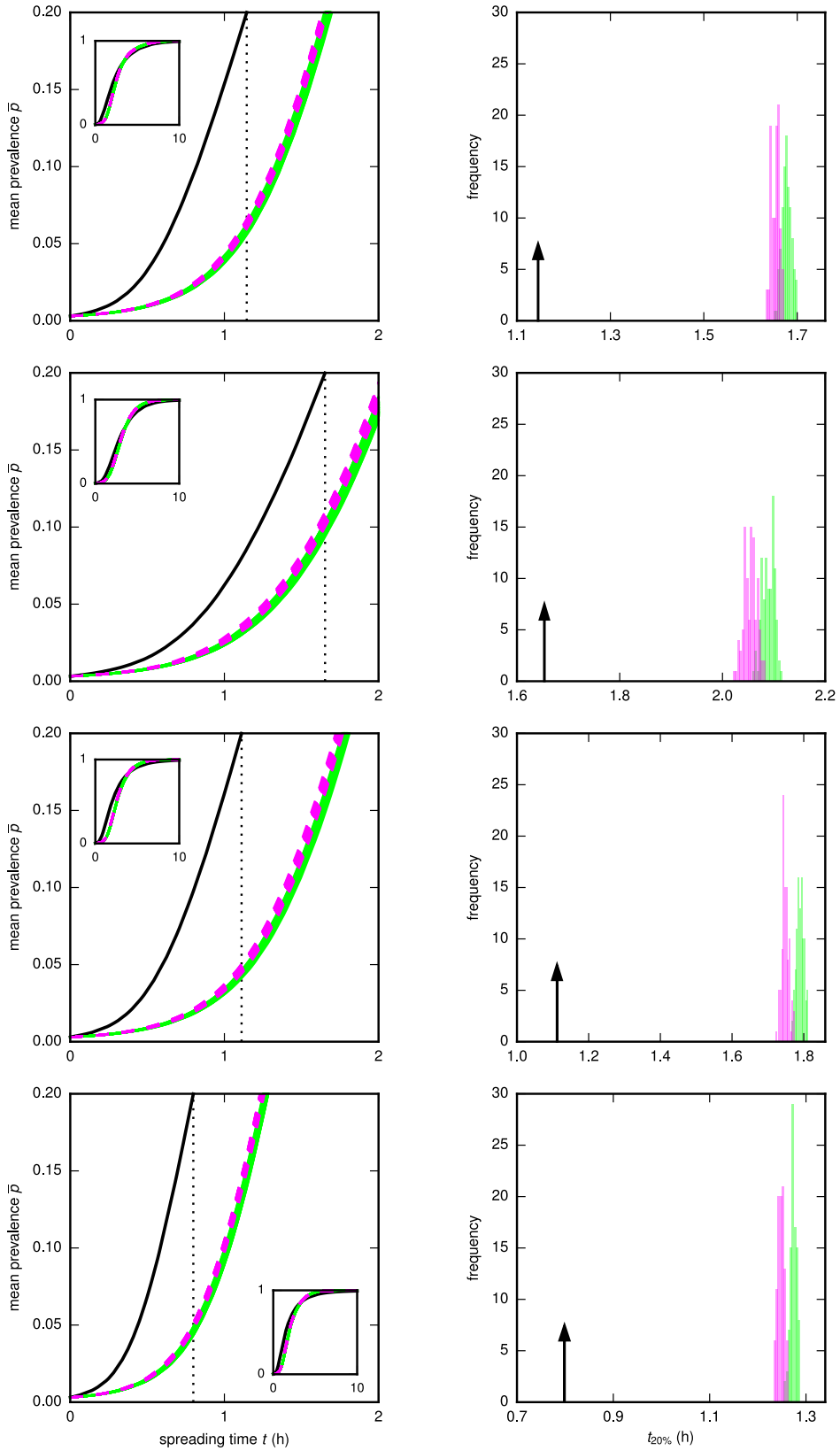

Figure B.1: **Prevalence curves in other datasets.**

References

[1] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777, 2014. [Cited on p. 91.]

[2] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997. [Cited on p. 28.]

[3] Marti J Anderson. A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46, 2001. [Cited on p. 35.]

[4] Marti J Anderson and Trevor J Willis. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*, 84(2):511–525, 2003. [Cited on p. 34.]

[5] Dirk Babendreier, David Joller, Jörg Romeis, Franz Bigler, and Franco Widmer. Bacterial community structures in honeybee intestines and their response to two insecticidal proteins. *FEMS microbiology ecology*, 59(3):600–610, 2007. [Cited on pp. 28 and 29.]

[6] Fredrik Bäckhed, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon. Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–1920, 2005. [Cited on p. 12.]

[7] David B Bahr and Eve Passerini. Statistical mechanics of opinion formation and collective behavior: Micro-sociology. *The Journal of mathematical sociology*, 23(1):1–27, 1998. [Cited on p. 102.]

[8] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016. [Also online at `http://barabasi.com/networksciencebook/`; accessed 2016-10-02]. [Cited on p. 86.]

[9] Richard D Bardgett, Chris Freeman, and Nicholas J Ostle. Microbial contributions to climate change through carbon cycle feedbacks. *The ISME journal*, 2(8):805–814, 2008. [Cited on p. 11.]

[10] Jeremy Bentham. *The Works of Jeremy Bentham Vol. 4*. Edinburgh: William Tait, 1838-1843. [Online; accessed 2016-09-16]. [Cited on p. 68.]

[11] SN Beshers, ZY Huang, Y Oono, and GE Robinson. Social inhibition and the regulation of temporal polyethism in honey bees. *Journal of Theoretical Biology*, 213(3):461–479, 2001. [Cited on p. 8.]

[12] Blair D Beverly, H McLendon, S Nacu, S Holmes, and Deborah M Gordon. How site fidelity leads to individual differences in the foraging activity of harvester ants. *Behavioral Ecology*, page arp041, 2009. [Cited on p. 47.]

[13] Jacobus C Biesmeijer and Thomas D Seeley. The use of waggle dance information by honey bees throughout their foraging careers. *Behavioral Ecology and Sociobiology*, 59(1):133–142, 2005. [Cited on p. 61.]

[14] Guy Bloch and Gene E Robinson. Chronobiology: reversal of honeybee behavioural rhythms. *Nature*, 410(6832):1048–1048, 2001. [Cited on p. 81.]

[15] Benjamin Blonder and Anna Dornhaus. Time-ordered networks reveal limitations to information flow in ant colonies. *PloS one*, 6(5):e20298, 2011. [Cited on p. 106.]

[16] Elizabeth A. Capaldi, Alan D. Smith, Juliet L. Osborne, Susan E. Fahrbach, Sarah M. Farris, Donald R. Reynolds, Ann S. Edwards, Andrew Martin, Gene E. Robinson, Guy M. Poppy, and Joseph R. Riley. Ontogeny of orientation flight in the honeybee revealed by harmonic radar. *Nature*, 403(6769):537–540, feb 2000. [Cited on p. 57.]

[17] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010. [Cited on p. 92.]

[18] Soumitesh Chakravorty, Danica Helb, Michele Burday, Nancy Connell, and David Alland. A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, 69(2):330–339, 2007. [Cited on p. 27.]

[19] Ilseung Cho and Martin J Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, 2012. [Cited on p. 13.]

[20] KR Clarke and RN Gorley. *PRIMER v6: User Manual/Tutorial*. PRIMER-E, 2006. [Cited on p. 31.]

[21] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. [Cited on pp. 90 and 91.]

[22] James R Cole, Qiong Wang, E Cardenas, J Fish, Benli Chai, Ryan J Farris, AS Kulam-Syed-Mohideen, Donna M McGarrell, T Marsh, George M Garrity, et al. The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic acids research*, 37(suppl 1):D141–D145, 2009. [Cited on p. 20.]

[23] Vanessa Corby-Harris, Patrick Maes, and Kirk E Anderson. The bacterial communities associated with honey bee (apis mellifera) foragers. *PLoS One*, 9(4):e95056, 2014. [Cited on p. 37.]

[24] Vanessa Corby-Harris, Lucy A Snyder, Melissa R Schwan, Patrick Maes, Quinn S McFrederick, and Kirk E Anderson. Origin and effect of alpha 2.2 acetobacteraceae in honey bee larvae and description of parasaccharibacter apium gen. nov., sp. nov. *Applied and environmental microbiology*, 80(24):7460–7472, 2014. [Cited on p. 37.]

[25] K Crailsheim et al. The protein balance of the honey bee worker. *Apidologie*, 21(5):417–429, 1990. [Cited on p. 26.]

[26] K Crailsheim, LHW Schneider, N Hrassnigg, G Bühlmann, U Brosch, R Gmeinbauer, and B Schöffmann. Pollen consumption and utilization in worker honeybees (apis mellifera carnica): dependence on individual age and function. *Journal of insect Physiology*, 38(6):409–419, 1992. [Cited on p. 26.]

[27] Karl Crailsheim. Trophallactic interactions in the adult honeybee (apis mellifera l.). *Apidologie*, 29:97–112, 1998. [Cited on p. 75.]

[28] Bernard J Crespi and Douglas Yanega. The definition of eusociality. *Behavioral Ecology*, 6(1):109–115, 1995. [Cited on p. 6.]

[29] Joachim R De Miranda and Elke Genersch. Deformed wing virus. *Journal of Invertebrate Pathology*, 103:S48–S61, 2010. [Cited on p. 75.]

[30] TZ DeSantis, Philip Hugenholtz, K Keller, EL Brodie, N Larsen, YM Piceno, R Phan, and Gary L Andersen. Nast: a multiple sequence alignment server for comparative analysis of 16s rrna genes. *Nucleic acids research*, 34(suppl 2):W394–W399, 2006. [Cited on p. 19.]

[31] Rod J Dillon, Gordon Webster, Andrew J Weightman, and A Keith Charnley. Diversity of gut microbiota increases with aging and starvation in the desert locust. *Antonie van Leeuwenhoek*, 97(1):69–77, 2010. [Cited on p. 36.]

[32] Maria G Dominguez-Bello, Elizabeth K Costello, Monica Contreras, Magda Magris, Glida Hidalgo, Noah Fierer, and Rob Knight. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107(26):11971–11975, 2010. [Cited on p. 13.]

[33] Maria G Dominguez-Bello, Kassandra M De Jesus-Laboy, Nan Shen, Laura M Cox, Amnon Amir, Antonio Gonzalez, Nicholas A Bokulich, Se Jin Song, Marina Hoashi, Juana I Rivera-Vinas, et al. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nature medicine*, 2016. [Cited on p. 13.]

[34] W Ford Doolittle and Olga Zhaxybayeva. On the origin of prokaryotic species. *Genome Research*, 19(5):744–756, 2009. [Cited on p. 12.]

[35] Anna Dornhaus. Specialization does not predict individual efficiency in an ant. *PLoS Biol*, 6(11):e285, 2008. [Cited on p. 47.]

[36] Fred C Dyer. The biology of the dance language. *Annual review of entomology*, 47(1):917–949, 2002. [Cited on p. 41.]

[37] Valeria DArgenio and Francesco Salvatore. The role of the gut microbiome in the healthy adult status. *Clinica Chimica Acta*, 451:97–102, 2015. [Cited on p. 25.]

[38] Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006. [Cited on p. 103.]

[39] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, 2004. [Cited on pp. 102 and 103.]

[40] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, jun 2011. [Cited on p. 19.]

[41] Robert C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010. [Cited on p. 28.]

[42] Robert C Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature methods*, 10(10):996–998, 2013. [Cited on p. 21.]

[43] Andrew M Edwards, Richard A Phillips, Nicholas W Watkins, Mervyn P Freeman, Eugene J Murphy, Vsevolod Afanasyev, Sergey V Buldyrev, Marcos GE da Luz, Ernesto P Raposo, H Eugene Stanley, et al. Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449(7165):1044–1048, 2007. [Cited on p. 44.]

[44] Philipp Engel, Vincent G Martinson, and Nancy A Moran. Functional diversity within the simple gut microbiota of the honey bee. *Proceedings of the National Academy of Sciences*, 109(27):11002–11007, 2012. [Cited on pp. 36 and 37.]

[45] Paul G Falkowski. Tracing oxygen's imprint on earth's metabolic evolution. *Science*, 311(5768):1724–1725, 2006. [Cited on p. 11.]

[46] WM Farina and AJ Wainselboim. Thermographic recordings show that honeybees may receive nectar from foragers even during short trophallactic contacts. *Insectes sociaux*, 48(4):360–362, 2001. [Cited on p. 76.]

[47] Karoline Faust, Leo Lahti, Didier Gonze, Willem M de Vos, and Jeroen Raes. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current opinion in microbiology*, 25:56–66, 2015. [Cited on p. 25.]

[48] Ana Fernández, Suiying Huang, Sherry Seston, Jian Xing, Robert Hickey, Craig Criddle, and James Tiedje. How stable is stable? function versus community composition. *Applied and Environmental Microbiology*, 65(8):3697–3704, 1999. [Cited on p. 25.]

[49] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, 8(9):e1002687, 2012. [Cited on p. 21.]

[50] K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008. [Cited on p. 106.]

[51] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016. [Cited on pp. 15 and 16.]

[52] Christoph Grüter and Walter M Farina. The honeybee waggle dance: can we follow the steps? *Trends in Ecology & Evolution*, 24(5):242–247, 2009. [Cited on p. 41.]

[53] Johannes Grwer, Henrik Ronellenfitsch, Marco G. Mazza, and Eleni Katifori. A trophallaxis inspired model for distributed transport between randomly interacting agents, 2016. arXiv:1607.06055. [Cited on p. 100.]

[54] Jo Handelsman, Michelle R Rondon, Sean F Brady, Jon Clardy, and Robert M Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–R249, 1998. [Cited on p. 12.]

[55] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000. [Cited on p. 104.]

[56] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012. [Cited on pp. 102, 105, and 106.]

[57] Dávid X Horváth and János Kertész. Spreading dynamics on networks: the role of burstiness, topology and non-stationarity. *New Journal of Physics*, 16(7):073037, 2014. [Cited on p. 106.]

[58] Zhi-Yong Huang and Gene E Robinson. Honeybee colony integration: worker-worker interactions mediate hormonally regulated plasticity in division of labor. *Proceedings of the National Academy of Sciences*, 89(24):11726–11729, 1992. [Cited on p. 59.]

[59] Zhi-Yong Huang and Gene E Robinson. Regulation of honey bee division of labor by colony age demography. *Behavioral Ecology and Sociobiology*, 39(3):147–158, 1996. [Cited on p. 57.]

[60] Christine R Hurd, Erik V Nordheim, and Robert L Jeanne. Elite workers and the colony-level pattern of labor division in the yellowjacket wasp, vespula germanica. *Behaviour*, 140(7):827–845, 2003. [Cited on p. 47.]

[61] José Luis Iribarren and Esteban Moro. Impact of human activity patterns on the dynamics of information diffusion. *Physical review letters*, 103(3):038702, 2009. [Cited on p. 103.]

[62] Patricio Jeraldo, Nicholas Chia, and Nigel Goldenfeld. On the suitability of short reads of 16s rrna for phylogeny-based analyses in environmental surveys. *Environmental microbiology*, 13(11):3000–3009, 2011. [Cited on pp. 18, 27, and 28.]

[63] Patricio Jeraldo, Alvaro Hernandez, Henrik B Nielsen, Xianfeng Chen, Bryan A White, Nigel Goldenfeld, Heidi Nelson, David Alhquist, Lisa Boardman, and Nicholas Chia. Capturing one of the human gut microbiomes most wanted: Reconstructing the genome of a novel butyrate-producing, clostridial scavenger from metagenomic sequence data. *Frontiers in Microbiology*, 7:783, 2016. [Cited on p. 14.]

[64] Patricio Jeraldo, Krishna Kalari, Xianfeng Chen, Jaysheel Bhavsar, Ashutosh Mangalam, Bryan White, Heidi Nelson, Jean-Pierre Kocher, and Nicholas Chia. IM-TORNADO: a tool for comparison of 16s reads from paired-end libraries. *PLoS One*, 9(12):e114804, 2014. [Cited on p. 18.]

[65] Walter Kaiser and Jana Steiner-Kaiser. Neuronal correlates of sleep, wakefulness and arousal in a diurnal insect. 1983. [Cited on p. 81.]

[66] Jens Kallmeyer, Robert Pockalny, Rishi Ram Adhikari, David C Smith, and Steven D'Hondt. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences*, 109(40):16213–16216, 2012. [Cited on p. 11.]

[67] Masaki Kamakura. Royalactin induces queen differentiation in honeybees. *Nature*, 473(7348):478–483, 2011. [Cited on p. 26.]

[68] Cheuk-Wai Kan, Christopher P Fredlake, Erin AS Doherty, and Annelise E Barron. DNA sequencing and genotyping in miniaturized electrophoresis systems. *Electrophoresis*, 25(21-22):3564–3588, 2004. [Cited on p. 14.]

[69] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E*, 83:025102, Feb 2011. [Cited on pp. 103, 106, and 108.]

[70] Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, Jari Saramäki, and Márton Karsai. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03005, 2012. [Cited on p. 106.]

[71] Balazs Kozma and Alain Barrat. Consensus formation on adaptive networks. *Phys. Rev. E*, 77:016102, Jan 2008. [Cited on p. 102.]

[72] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008. [Cited on p. 103.]

[73] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015. [Cited on pp. 21 and 22.]

[74] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235, 2005. [Cited on p. 22.]

[75] Wenfu Mao, Mary A Schuler, and May R Berenbaum. A dietary phytochemical alters caste-associated gene expression in honey bees. *Science advances*, 1(7):e1500795, 2015. [Cited on p. 26.]

[76] Vincent G Martinson, Bryan N Danforth, Robert L Minckley, Olav Rueppell, Salim Tingek, and Nancy A Moran. A simple and distinctive microbiota associated with honey bees and bumble bees. *Molecular Ecology*, 20(3):619–628, 2011. [Cited on p. 26.]

[77] Vincent G Martinson, Jamie Moy, and Nancy A Moran. Establishment of characteristic gut bacteria during development of the honeybee worker. *Applied and environmental microbiology*, 78(8):2830–2840, 2012. [Cited on p. 26.]

[78] Heather R Mattila, Daniela Rios, Victoria E Walker-Sperling, Guus Roeselers, and Irene LG Newton. Characterization of the active microbiotas associated with honey bees reveals healthier and broader communities when colonies are genetically diverse. *PLoS One*, 7(3):e32962, 2012. [Cited on p. 26.]

[79] Michael Möglich and Bert Hölldobler. Social carrying behavior and division of labor during nest moving of ants. *Psyche*, 81(2):219–236, 1974. [Cited on p. 47.]

[80] Kathrin I Mohr and Christoph C Tebbe. Diversity and phylotype consistency of bacteria in the guts of three bee species (apoidea) at an oilseed rape field. *Environmental Microbiology*, 8(2):258–272, 2006. [Cited on p. 26.]

[81] Darrell Moore, Jennifer E Angel, Iain M Cheeseman, Susan E Fahrbach, and Gene E Robinson. Timekeeping in the honey bee colony: integration of circadian rhythms and division of labor. *Behavioral Ecology and Sociobiology*, 43(3):147–160, 1998. [Cited on p. 81.]

[82] Nancy A Moran. Symbiosis as an adaptive process and source of phenotypic complexity. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8627–8633, 2007. [Cited on p. 26.]

[83] Nancy A Moran, Allison K Hansen, J Elijah Powell, and Zakee L Sabree. Distinctive gut microbiota of honey bees assessed using deep sampling from individual worker bees. *PLoS One*, 7(4):e36393, 2012. [Cited on pp. 26 and 36.]

[84] Jennifer G Mulle, William G Sharp, and Joseph F Cubells. The gut microbiome: a new frontier in autism research. *Current psychiatry reports*, 15(2):1–9, 2013. [Cited on p. 2.]

[85] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003. [Cited on p. 86.]

[86] HL Nixon and CR Ribbands. Food transmission within the honeybee community. *Proceedings of the Royal Society of London B: Biological Sciences*, 140(898):43–50, 1952. [Cited on p. 75.]

[87] Martin A Nowak, Corina E Tarnita, and Edward O Wilson. The evolution of eusociality. *Nature*, 466(7310):1057–1062, 2010. [Cited on p. 6.]

[88] Samuel A Ocko and L Mahadevan. Collective thermoregulation in bee clusters. *Journal of The Royal Society Interface*, 11(91):20131033, 2014. [Cited on p. 8.]

[89] Yoshitsugu Oono. *The Nonlinear World: Conceptual Analysis and Phenomenology*. Springer Science & Business Media, 2012. [Cited on p. 1.]

[90] George F Oster and Edward O Wilson. *Caste and ecology in the social insects*. Princeton University Press, 1978. [Cited on p. 47.]

[91] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015. [Cited on p. 104.]

[92] R Christopher Plowright and Catherine MS Plowright. Elitism in social insects: a positive feedback model. In Robert L Jeanne, editor, *Interindividual behavioral variability in social insects*. Westview Press, 1988. [Cited on p. 47.]

[93] Andrew M Reynolds, Alan D Smith, Randolf Menzel, Uwe Greggers, Donald R Reynolds, and Joseph R Riley. Displaced honey bees perform optimal scale-free search flights. *Ecology*, 88(8):1955–1961, 2007. [Cited on p. 44.]

[94] Andrew M Reynolds, Alan D Smith, Don R Reynolds, Norman L Carreck, and Juliet L Osborne. Honeybees perform optimal scale-free searching flights when attempting to locate a food source. *Journal of Experimental Biology*, 210(21):3763–3770, 2007. [Cited on p. 44.]

[95] Andy Reynolds. Liberating lévy walk research from the shackles of optimal foraging. *Physics of life reviews*, 14:59–83, 2015. [Cited on p. 44.]

[96] C Ronald Ribbands. *The behaviour and social life of honeybees*. Dover Publications, 1964. [Cited on p. 47.]

[97] JR Riley, AD Smith, DR Reynolds, AS Edwards, JL Osborne, IH Williams, NL Carreck, GM Poppy, et al. Tracking bees with harmonic radar. *Nature*, 379(6560):29–30, 1996. [Cited on p. 44.]

[98] Gene E Robinson, Robert E Page, Colette Strambi, and Alain Strambi. Colony integration in honey bees: mechanisms of behavioral reversion. *Ethology*, 90(4):336–348, 1992. [Cited on p. 50.]

[99] Luis EC Rocha, Fredrik Liljeros, and Petter Holme. Information dynamics shape the sexual networks of internet-mediated prostitution. *Proceedings of the National Academy of Sciences*, 107(13):5706–5711, 2010. [Cited on pp. 103 and 106.]

[100] Luis EC Rocha, Fredrik Liljeros, and Petter Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol*, 7(3):e1001109, 2011. [Cited on p. 106.]

[101] Zakee L Sabree, Allison K Hansen, and Nancy A Moran. Independent studies using deep sequencing resolve the same set of core bacterial species dominating gut communities of honey bees. *PLoS One*, 7(7):e41250, 2012. [Cited on pp. 26 and 36.]

[102] M-P Schippers, R Dukas, RW Smith, J Wang, K Smolen, and GB McClelland. Lifetime performance in foraging honeybees: behaviour and physiology. *Journal of Experimental Biology*, 209(19):3828–3836, 2006. [Cited on p. 61.]

[103] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009. [Cited on pp. 19 and 28.]

[104] Erwin Schrödinger. *What is life?* Cambridge University Press, 1944. [Cited on p. 1.]

[105] Thomas D Seeley. The wisdom of the hive: the social physiology of honey bee, 1995. [Cited on p. 6.]

[106] K Sekiguchi and SF Sakagami. Structure of foraging population and related problems in the honeybee, with considerations on the division of labour in bee colonies. *Hokkaido Natl Agric Exp Sta Rep*, 69:1–65, 1966. [Cited on p. 47.]

[107] Kelly Servick. Proposed study would closely track 10,000 new yorkers. *Science*, 350(6260):493–494, 2015. [Cited on p. 69.]

[108] Gil Sharon, Daniel Segal, John M Ringo, Abraham Hefetz, Ilana Zilber-Rosenberg, and Eugene Rosenberg. Commensal bacteria play a role in mating preference of drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 107(46):20051–20056, 2010. [Cited on p. 37.]

[109] A Christian Silva and Victor M Yakovenko. Temporal evolution of the thermal and superthermal income classes in the usa during 1983–2001. *EPL (Europhysics Letters)*, 69(2):304, 2004. [Cited on p. 46.]

[110] Maksim Sipos. *Phase transitions in fluids and biological systems*. PhD thesis, University of Illinois at Urbana-Champaign, 2013. [Cited on pp. 11 and 13.]

[111] Maksim Sipos, Patricio Jeraldo, Nicholas Chia, Ani Qu, A Singh Dhillon, Michael E Konkel, Karen E Nelson, Bryan A White, and Nigel Goldenfeld. Robust computational analysis of rrna hypervariable tag datasets. *PLoS One*, 5(12):e15220, 2010. [Cited on pp. 18 and 19.]

[112] Se Jin Song, Christian Lauber, Elizabeth K Costello, Catherine A Lozupone, Gregory Humphrey, Donna Berg-Lyons, J Gregory Caporaso, Dan Knights, Jose C Clemente, Sara Nakielny, et al. Cohabiting family members share microbiota with one another and with their dogs. *Elife*, 2:e00458, 2013. [Cited on pp. 13 and 26.]

[113] Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. Modeling human dynamics of face-to-face interaction networks. *Physical review letters*, 110(16):168701, 2013. [Cited on pp. 93 and 96.]

[114] Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. Model reproduces individual, group and collective dynamics of human contact networks. *Social Networks*, 47:130–137, 2016. [Cited on p. 95.]

[115] Eric J Stewart. Growing unculturable bacteria. *Journal of bacteriology*, 194(16):4151–4160, 2012. [Cited on p. 12.]

[116] Maria Julia Szolderits and Karl Crailsheim. A comparison of pollen consumption and digestion in honeybee (apis mellifera carnica) drones and workers. *Journal of Insect Physiology*, 39(10):877–881, 1993. [Cited on p. 26.]

[117] Taro Takaguchi, Mitsuhiro Nakamura, Nobuo Sato, Kazuo Yano, and Naoki Masuda. Predictability of conversation partners. *Phys. Rev. X*, 1:011008, Sep 2011. [Cited on p. 103.]

[118] Kevin R Theis, Thomas M Schmidt, and Kay E Holekamp. Evidence for a bacterial mechanism for group-specific social odors among hyenas. *Scientific reports*, 2:615, 2012. [Cited on p. 37.]

[119] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics-a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):1, 2012. [Cited on pp. 13 and 14.]

[120] Dan P Toma, Guy Bloch, Darrell Moore, and Gene E Robinson. Changes in period mRNA levels in the brain and division of labor in honey bee colonies. *Proceedings of the National Academy of Sciences*, 97(12):6914–6919, 2000. [Cited on p. 81.]

[121] John M Townsend-Mehler, Fred C Dyer, and Kim Maida. Deciding when to explore and when to persist: a comparison of honeybees and bumblebees in their response to downshifts in reward. *Behavioral ecology and sociobiology*, 65(2):305–312, 2011. [Cited on p. 65.]

[122] Alexei Vazquez, Balazs Racz, Andras Lukacs, and Albert-Laszlo Barabasi. Impact of non-poissonian activity patterns on spreading processes. *Physical review letters*, 98(15):158702, 2007. [Cited on pp. 102, 103, and 106.]

[123] Claudia Vilo and Qunfeng Dong. Evaluation of the RDP classifier accuracy using 16s rRNA gene variable regions. *Metagenomics*, 1:1–5, 2012. [Cited on p. 27.]

[124] P Kirk Visscher and Thomas D Seeley. Foraging strategy of honeybee colonies in a temperate deciduous forest. *Ecology*, 63(6):1790–1801, 1982. [Cited on p. 65.]

[125] Gandhimohan M Viswanathan, V Afanasyev, SV Buldyrev, EJ Murphy, PA Prince, H Eugene Stanley, et al. Lévy flight search patterns of wandering albatrosses. *Nature*, 381(6581):413–415, 1996. [Cited on p. 43.]

[126] Gandhimohan M Viswanathan, Marcos GE Da Luz, Ernesto P Raposo, and H Eugene Stanley. *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge University Press, 2011. [Cited on p. 42.]

[127] Gandimohan M Viswanathan, Sergey V Buldyrev, Shlomo Havlin, MGE Da Luz, EP Raposo, and H Eugene Stanley. Optimizing the success of random searches. *Nature*, 401(6756):911–914, 1999. [Cited on p. 43.]

[128] Karl Von Frisch. Die tänze der bienen. In *Tanzsprache und Orientierung der Bienen*, pages 3–330. Springer, 1965. [Cited on p. 40.]

[129] Ying Wang, Thomas M Gilbreath III, Phanidhar Kukutla, Guiyun Yan, and Jiannong Xu. Dynamic gut microbiome across life history of the malaria mosquito anopheles gambiae in kenya. *PloS one*, 6(9):e24767, 2011. [Cited on p. 36.]

[130] James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. [Cited on p. 11.]

[131] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002. [Cited on p. 102.]

[132] Robert H Whittaker. Evolution and measurement of species diversity. *Taxon*, pages 213–251, 1972. [Cited on p. 21.]

[133] Mark L Winston. *The biology of the honey bee*. Harvard University Press, 1987. [Cited on pp. 6, 26, 39, 48, 59, and 75.]

[134] Carl R Woese and George E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977. [Cited on pp. 11 and 13.]

[135] Ye Wu, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber. Evidence for a bimodal distribution in human communication. *Proceedings of the national academy of sciences*, 107(44):18803–18808, 2010. [Cited on p. 103.]

[136] Victor M Yakovenko and J Barkley Rosser Jr. Colloquium: Statistical mechanics of money, wealth, and income. *Reviews of Modern Physics*, 81(4):1703, 2009. [Cited on p. 46.]

[137] Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. *Nature Reviews Microbiology*, 12(9):635–645, 2014. [Cited on p. 17.]

[138] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, et al. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012. [Cited on p. 13.]

[139] Zhao Zhi-Dan, Xia Hu, Shang Ming-Sheng, and Zhou Tao. Empirical analysis on the human dynamics of a large-scale short message communication system. *Chinese Physics Letters*, 28(6):068901, 2011. [Cited on p. 103.]