

**Master Thesis, Institute of Computer Science, Freie Universität Berlin**

**Biorobotics Lab, Intelligent Systems and Robotics**

# **Temporal Analysis of Honey Bee Interaction Networks Based on Spatial Proximity**



*Alexa Schlegel*

Matriculation number: 4292909

[alexa.schlegel@fu-berlin.de](mailto:alexa.schlegel@fu-berlin.de)

Supervisor: Prof. Dr. Tim Landgraf, Freie Universität Berlin

Second Supervisor: Dr. Philipp Hövel, Technische Universität Berlin

Berlin, May 7, 2017

## **Eidesstattliche Erklärung**

Ich versichere hiermit an Eides Statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Berlin, den 20. April 2017

Alexa Schlegel

## **Abstract**

The BeesBook system provides high-resolution data about bee movements within a single colony by automatically tracking individual honey bees inside a hive over their entire life. This thesis focuses on the process of designing and implementing a network pipeline to extract interaction networks from this data. Spatial proximity is used as an indicator for interactions between bees. Social network analysis methods were applied to investigate the static and dynamic properties of the resulting social networks of honey bees on a global, intermediate and local level. The resulting networks were characterized by a low hierarchical structure and a high density. The global structure of the colony seems to be stable over time. The local structure is highly dynamic, as bees change communities as they age. Communities in the honey bee network are formed by age groups that show a high spatial fidelity. The findings are in line with the established state of research that colonies are organized around age-based task division. The results of the analysis validate the implemented pipeline and the inferred networks. Consequently, this work provides an excellent foundation for future research focusing on temporal network analysis.

***Keywords*** – social insects, spatial proximity network, interaction network, *Apis mellifera*, community detection, social network analysis

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	2
1.2	Research Goal and Method . . . . .	2
1.3	Outline . . . . .	3
<b>2</b>	<b>Theoretical Background for Network Analysis of Insect Colonies</b>	<b>4</b>
2.1	Social Network Analysis . . . . .	4
2.1.1	Network Measures and Metrics . . . . .	5
2.1.2	Community Detection . . . . .	6
2.1.3	Temporal Networks . . . . .	8
2.2	Related Studies . . . . .	9
2.2.1	Static Network Analysis of Honey Bee Colonies . . . . .	9
2.2.2	Temporal Network Analysis of Insect Colonies . . . . .	10
<b>3</b>	<b>Methodology</b>	<b>12</b>
3.1	The Experimental Setup and Dataset . . . . .	12
3.2	Data Quality and Data Cleansing . . . . .	13
3.2.1	Data Scheme . . . . .	13
3.2.2	ID Probabilities, Confidence Level, and Quality . . . . .	14
3.2.3	Detection Frequency Filter . . . . .	15
3.2.4	Time Series of Bees and Bee Pairs . . . . .	16
3.3	Inferring Spatial Proximity Networks . . . . .	17
3.3.1	Defining the Network and its Parameters . . . . .	17
3.3.2	Implementing the Network Pipeline . . . . .	18
3.3.3	Choosing Parameter Values for Network Analysis . . . . .	19
3.3.4	Summary . . . . .	20
3.4	Methods for Analyzing Spatial Proximity Networks . . . . .	21
3.4.1	Investigating the Topology and Network Characteristics . . . . .	22
3.4.2	Detecting Communities . . . . .	23
3.4.3	Development of Community Members . . . . .	26
<b>4</b>	<b>Results of Network Analysis</b>	<b>28</b>
4.1	Static Perspectives of Honey Bee Networks . . . . .	28
4.1.1	Properties of the Bee Colony . . . . .	29
4.1.2	Characteristics of Bees . . . . .	29
4.1.3	Functional Groups within the Colony . . . . .	32
4.2	Temporal Perspectives of Honey Bee Networks . . . . .	35
4.2.1	Stability of Functional Groups . . . . .	35

4.2.2	Dynamic of Individual Bees . . . . .	36
4.3	Discussion of Results . . . . .	37
4.3.1	Network Topology and Properties of Honey Bee Colonies . . . . .	37
4.3.2	Characterization of Functional Groups and its Dynamics . . . . .	39
<b>5</b>	<b>Conclusions</b>	<b>41</b>
5.1	Limitations . . . . .	41
5.2	Recommendations . . . . .	42
5.3	Outlook . . . . .	43
5.4	Closing Remarks . . . . .	44
<b>Bibliography</b>		<b>45</b>
<b>List of Figures</b>		<b>48</b>
<b>List of Tables</b>		<b>50</b>
<b>A Literatur Review</b>		<b>51</b>
<b>B Additional Information about the Dataset</b>		<b>55</b>
<b>C Network Analysis</b>		<b>57</b>

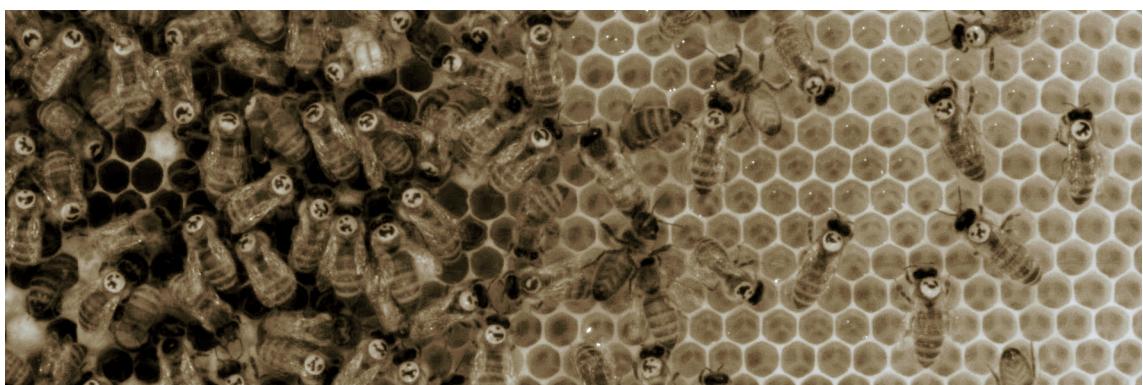
# Chapter 1

## Introduction

A social insect society is formed by thousands of individuals that continuously interact with each other inside a dark nest. Honey bees are organized in colonies, which are a complex and dynamic system. Observing individual honey bees and their interactions with each other is vital for understanding collective behavior and the organization of tasks within the colony.

Within the BeesBook project of the Biorobotics Lab of Freie Universität Berlin, Wario, Wild, Couvillon, *et al.* [1] developed technologies to automatically track all individuals of a honey bee (*Apis mellifera*) colony that are inside the honeycomb. Shortly after hatching, each bee is marked with a circular 12-bit tag (Figure 1.1) on their thorax and then added to the observation colony. Several cameras observe the colony over a period of nine weeks. An image analysis pipeline evaluates each frame automatically. The resulting dataset contains the exact position of each detected bee on the honeycomb and its age for each frame.

In this thesis, worker-worker interaction networks, based on spatial proximity, are derived from the described dataset. Each node in the network is a bee and a link between two nodes results if two bees are located close to each other for a specified period. The networks are time-aggregated, which means that one network represents the data of multiple frames. After extracting the static networks, social network analysis methods are applied to determine the characteristics of the resulting networks and its social structure.



**Figure 1.1:** Tagged bees inside the observation hive.

## 1.1 Motivation

Manual insect tagging and tracking are widely applied in the behavioral sciences. Insects are marked using colored paint or numbered tags to distinguish individuals, and then they are observed using a video recorder or by taking photos. The interaction data is obtained by repeatedly watching the video files and manually extracting events. Labeling only a small group of the colonies' individuals, short observation periods, a low sampling resolution, or limiting the observation to only a small area of the hive is very common.

Consequently, most insect related studies, in the field of animal social network analysis, examine only a limited subset of the life of a colony. Due to technical limitations, the majority of social insect interaction network studies focus on static aspects. Recently, automated tracking of insects has become technically feasible [1]–[3]. Using automated high resolution tracking data, which includes all individuals of the entire hive over an extended period, allows for more advanced analysis focusing on temporal dynamics.

Automatic tracking data of ant colonies [4] has already been investigated with network analysis methods. Mersch, Crespi, and Keller [4] used a dataset, obtained by long-term observation of six ant colonies including all individuals, to investigate temporal aspects of the ants' social network. Their study was able to provide new insights into the dynamics of the colonies' functional units.

Applying network analysis methods on data that was obtained by automatic tracking of honey bees offers the possibility to investigate temporal aspects of bees' social networks. Therefore it holds the potential to reveal new insights in the area of behavioral sciences. My work contributes to lay the foundation for following this new path of research.

## 1.2 Research Goal and Method

The aim of this thesis is to investigate whether the BeesBook data of tracked honey bees is useful for creating worker-worker interaction networks using spatial proximity as an indicator for interactions between bees. Thus, I will implement a pipeline to extract networks out of the given data. Furthermore, I will investigate if the resulting networks are suitable for social network analysis.

I want to achieve my research goals by answering the following questions:

1. *Is it possible to infer temporal networks with the provided honey bee tracking data?*  
 What challenges and limitations arise from using this data?  
 What pipeline parameters are necessary?
2. *What kind of worker-worker interaction networks emerge and how are they structured?*  
 What is their topology?

What are the properties of the networks and how do they differ from randomly generated networks?

3. *Does the network display a meaningful community structure?*

How are the identified communities characterized?

Do they reflect already known colony behavior concerning age and spatial distribution?

4. *How do these communities develop over time?*

Do the communities have stable properties?

How do members move between communities?

This work is meant to establish the foundation for future research using a network science approach to study the complex system of honey bee colonies and their collective behavior.

The methodology of this work consists of two parts, described in detail in Chapter 3. The first part details the approach to infer and define spatial proximity networks using the honey bee tracking data. The second part analyzes properties, communities, and the development of the identified networks.

## 1.3 Outline

This thesis is organized as follows. Chapter 2 is a short introduction to social network analysis. I define network measures, terms, and algorithms used throughout this work and provide a brief summary of the current state of research concerning social insect networks, temporal networks and community detection in animal social networks. In Chapter 3, I describe my research approach in general and how the pipeline infers networks out of the given dataset, what steps are needed and what parameters it uses. Also, I explain and justify what decisions I took during the network analyses and community detection process. In Chapter 4, I report the results of the network analysis and the characteristics of the extracted communities. Finally, in chapter 5, I summarize the results, discuss limitations and conclude with directions for future work.

# Chapter 2

## Theoretical Background for Network Analysis of Insect Colonies

The following chapter gives a short introduction into social network analysis (SNA) and introduces social insect interaction networks, as a specific type of a biological network. It defines terms and concepts used throughout this work and explains the applied network metrics and algorithms. I also provide a summary of the most relevant studies that use a network analysis approach to investigate interaction networks of social insect colonies.

### 2.1 Social Network Analysis

A *social network* is a representation of a social structure comprising actors such as individuals, affiliations, as well as their social interactions. The network model conceptualizes social, economic, or political structures as lasting patterns of interactions between actors [5]. In mathematical terms, networks are graphs and thus consist of *nodes* (vertices, representing individuals), and *links* (edges, relationships or interactions). Social network analysis provides a set of methods, measures, and theories, borrowed from network and graph theory, to investigate social structures and their dynamics.

This work focuses on the special case of social insect networks, where individuals are nodes and links represent interaction events between individuals. Those networks are called *interaction networks* or association networks. According to Charbonneau, Blonder, and Dornhaus [6], the interactions used as a link can be of four different types when looking at social insect networks: spatial proximity, physical contact (usually with antennae, “antennation”), a food exchange event (trophallaxis), or specific communication signals. Trophallaxis is the directed mouth-to-mouth transfer of fluids and is used by social insects to exchange information and food.

Links can be directed or undirected and weighted or unweighted. The link weights represent the strength of the relationship; commonly the number or duration of interactions is used [7].

In the course of this work I use the term *frequency* to refer to the total number of occurrences (absolute frequency), as opposed to the use of this term in physics.

**Table 2.1: Definitions** Basic network science terms and definitions.

<b>Network size</b>	$N$ is the total number of nodes, representing the number of animals in a network.
<b>Number of links</b>	$L$ is the total number of links in a network, representing social interactions.
<b>Link weight</b>	$w_{ij}$ of a link $l_{ij}$ between node $i$ and $j$ is an indicator of how important that link is.
<b>Component</b>	is a subnet of nodes in a network so that there is a path between any two nodes that belong to the component.
<b>Degree</b>	$k_i$ of a node $i$ represents the number of other animals the $i$ -th animal interacts with.
<b>Average degree</b>	$\langle k \rangle$ is the number of interaction partners an animal has, on average, in the network.
<b>Shortest path length</b>	$d_{ij}$ is the shortest number of links between the nodes $i$ and $j$ , also called distance.
<b>Average shortest path length</b>	$\langle d \rangle$ is the average of all shortest paths between all pairs of nodes.
<b>Diameter</b>	$d_{\max}$ represents the length of the longest of all shortest paths, the longest possible path length in the network.

### 2.1.1 Network Measures and Metrics

The following definitions are mainly taken from Barabási [8] and Newman [9] and refer to undirected and unweighted static network measures, except for strength. Table 2.1 summarizes the basic variables and terms used in the course of this work.

**Density**  $D$  is the number of realized links divided by the number of theoretically possible links is defined as

$$D = \frac{2L}{N(N - 1)}. \quad (2.1)$$

**Strength**  $s_i$  of a node  $i$  is also called the weighted degree. It measures the total weight of links connected to that node. The average strength is denoted as  $\langle s \rangle$ . According to Barrat, Barthelemy, Pastor-Satorras, *et al.* [10] it is defined as

$$s_i = \sum_{j=1}^N w_{ij}. \quad (2.2)$$

**Global clustering coefficient (gcc)**  $c_\Delta$  is also called transitivity. The gcc is based on triplets. A connected triplet consists of three nodes connected by two links. A

triangle consists of three connected triplets. According to Wasserman and Faust [5] the gcc is defined as

$$c_{\Delta} = \frac{(\text{number of triangles}) \times 3}{(\text{number of connected triples})}. \quad (2.3)$$

**Local clustering coefficient (lcc)**  $c_i$  of a node  $i$  quantifies how close the node's neighborhood is to being a complete subgraph and is defined as

$$c_i = \frac{(\text{number of pairs of neighbors of } i \text{ that are connected})}{(\text{number of pairs of neighbors of } i)}. \quad (2.4)$$

**Centrality** When looking at the node level structure of a network, it is possible to identify nodes, which are important or central to different aspects of the network. This concept is called *centrality* and measures the influence of a node in a network.

**Degree Centrality**  $C_D^i$  is the normalized degree  $k_i$  of a node  $i$  in relation to the whole network. It is calculated as

$$C_D^i = \frac{k_i}{N - 1}. \quad (2.5)$$

**Closeness Centrality**  $C_C^i$  of a node  $i$  measures how close this node is to all other nodes in the network. It is the average length of shortest paths between node  $i$  and all other nodes in the network and is calculated as

$$C_C^i = \frac{N}{\sum_j d_{ij}}. \quad (2.6)$$

**Betweenness Centrality**  $C_B^i$  of a node  $i$  measures the extent to which a node lies on shortest paths between other nodes. Nodes that occur on many shortest paths between other nodes have a higher betweenness than those that do not.

## 2.1.2 Community Detection

To understand the large-scale structure of networks, one can look at the network's community structure. Communities are naturally occurring groups within a network, usually also called clusters, cohesive groups or modules and have no widely accepted, unique definition [11]. For my work, I adapt the definition according to Barabási [8, p. 322]: "In network science, we call a community a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities."

In contrast to a simple graph partition, the number and size of communities is not predetermined or set in advance.

In animal social networks, communities refer to groups of individuals that are associated more with each other than they are with the rest of the population. These communities reflect an intermediate level of social organization, which is located between the individual and population level [12].

There are a lot of different approaches and algorithms that address the detection of communities. Fortunato [13] gives an extensive overview of the various types of community detection algorithms. Explaining any of those would be beyond the scope of this work. For example, traditional methods include algorithms based on graph partitioning, hierarchical clustering, and spectral clustering. There are also divisive and agglomerative algorithms. The algorithms used in this work are described in the following sections and include the leading eigenvector [14] and walktrap [15] algorithm.

## Modularity

Modularity is a quantity that measures the quality of a partitioning. It can be used to compare one community partition to another and to decide which is the better one. Modularity optimization is also used for community detection algorithms. A high modularity of a network indicates more connections between nodes within a community and fewer connections between nodes of different communities. The basic idea is: If the fraction of links inside the community is higher than expected in the same community of a corresponding random graph with the same degree distribution, then it is a community in the sense of modularity. This difference is summed up and normalized.

Fewer links inside the community than expected result in a negative value, more links positive. If all nodes fall into one community, the modularity is zero.

## Leading Eigenvector and Walktrap

The *leading eigenvector* algorithm was proposed by Newman [14]. It uses the eigenvectors of matrices for finding community structures in networks. It is a top-down hierarchical approach that optimizes modularity. The algorithm starts with all nodes inside one community (modularity is 0). In each step, the network is split into two parts, so that the modularity of the new separation increases. The splitting is done by first calculating the leading eigenvector of the modularity matrix and then splitting the network in a way that maximizes the modularity improvement based on the leading eigenvector. The algorithm stops if the modularity does not further increase.

The *walktrap* algorithm by Pons and Latapy [15] is based on random walks. The authors use random walks as a tool to calculate the similarity between nodes of a network. The algorithm uses a bottom-up hierarchical approach, which means the

algorithm starts with each node in a single community. The basic idea of walktrap is that short distance random walks (the step size is a parameter) tend to stay in the same community because there are only a few links that lead outside a given community. The results of these random walks are used to merge separate communities. Again modularity can be used to cut the dendrogram at an optimal place.

### 2.1.3 Temporal Networks

When modeling *temporal* or *dynamic networks*, two main approaches exist: (1) time-aggregated (discrete), where the data is aggregated either in a disjoint, overlapping or cumulative snapshot; and (2) the time-ordered (continuous) approach, with interactions having a start and end timestamp [16]–[18].

The time-aggregated approach cumulates the data for each snapshot and reduces the available information per link. In contrast, the time-ordered approach retains the information about when interactions occurred and how long they lasted. It provides a detailed insight when timing and order of interactions are important. It can be used to model the topological flow of information through a network.

Choosing a suitable time interval for data aggregation is challenging [17], but many methods for analyzing time-aggregated networks already exists, whereas, for time-ordered networks, only a limited toolset is available. To model nodes and link weights in time-aggregated networks can be challenging, when since interactions, which occurred earlier or later in time, should be weighted accordingly.

## 2.2 Related Studies

Studies using a network analysis approach focusing on interaction networks to investigate the behavior of social insects, especially honey bees are relevant for my work. I mainly reviewed studies mentioned in the survey papers of Pinter-Wollman, Hobson, Smith, *et al.* [17], Krause, James, Franks, *et al.* [19, Chapter 15] and Charbonneau, Blonder, and Dornhaus [6].

The most relevant studies were classified by:

- **Type of analysis**

temporal or static analysis using automated or manual tracking over a long or short term

- **Studied species**

honey bees or other social insects

I reviewed the limitations of the studies in regards to time, space, and the number of tracked individuals. Table A.2 (Appendix A) summarizes the selected studies and the characteristics of: duration of study, observation period, sampling resolution, the number of colonies, the number of marked individuals, and space limitations. I also recorded whether the studies included age cohorts in their analysis and listed the software tools used for network analysis.

Within the scope of my literature review, I found a lot of studies in the field of static network analysis of ants [20]–[25], wasps [26] and bumblebees [27], but only a few related to honey bees [28]–[31]. I did not find any studies focused on temporal aspects of honey bee colonies, but I did find several studies focused on temporal aspects of ant colonies [4], [32], [33].

### 2.2.1 Static Network Analysis of Honey Bee Colonies

The most advanced work studying honey bees using a network science approach is by Baracchi and Cini [28]. Their study revealed a highly compartmentalized structure inside the honey bee colony: Bees organize by age groups, which occupy separate areas of the comb and perform different tasks. There is limited contact between these groups.

Generally, the theory that bees change tasks over the course of their lifetime, starting as nurses in the nest and ending as foragers outside, termed as temporal polyethism, is widely accepted and has been studied for a long time [34]–[36]. Johnson [35] observed two groups of within-nest bees: young bees responsible for the brood care and middle-aged bees specialized on nectar processing and nest maintenance. Seeley [34] observes four age subcastes among worker bees besides the queen cast: cell cleaning, brood nest, food storage, forager.

Lindauer [36] defined certain tasks a bee can perform at any given age. Also, a bee can perform several different tasks per day. The bee is flexible and responds to the given needs of the hive. Young bees mostly clean cells and old bees mainly forage,

instead middle-aged bees perform several tasks. [36]

Baracchi and Cini [28] use the frequency of interactions between bees as link weights in an undirected worker-worker interaction network. The body length of a bee defines the radius of spatial proximity. Baracchi and Cini use the node level measures strength (weighted degree), closeness and eigenvector centrality to investigate the networks. They also perform a cluster analysis using as similarity the local network measures. The main shortcomings of their work are sample size and observation frequency. They studied one colony with 4,000 individuals, marking only 211 bees from three predefined age cohorts, and observed only one side of the observation hive for ten hours by capturing with a low resolution of one frame per minute.

Scholl and Naug [30] investigated the mechanism behind the emergence of organizational immunity of honey bee colonies by using unweighted, undirected physical contact and trophallaxis networks. They observed one hour per day, with three days of observation spread over three weeks. In the field of network analysis, they investigated the interactions between three predefined age cohorts.

Naug [29] inspects the network structure of weighted, directed trophallaxis networks using four age cohorts. He evaluates the changes in transmission dynamics produced by experimental manipulation. The dataset is limited to one hour of observation and only first- and second-order trophallaxis interactions are considered. The food transfer from the forager to a worker bee is called first level interaction, the food transfer from that worker bee to other bees is called second-order.

### 2.2.2 Temporal Network Analysis of Insect Colonies

Mersch, Crespi, and Keller [4] apply similar methods to my work. They automatically tracked all individuals of six ant colonies over a period of 41 days using a resolution of two frames per second. For each observation day, the authors extracted time-aggregated weighted contact networks per colony, using antennation as the physical contact event. They applied the Infomap community detection algorithm to each daily network and revealed three distinct and robust groups. Each group represents a functional behavioral unit, with ants changing groups as they age. Except for community detection, they did not use any other network science methods to investigate the network properties.

Jeanson [33] also used automatic tracking. His work is focused on the investigation of the temporal stability of spatial proximity networks in four ant colonies over three weeks. Here, proximity is defined as  $4/3$  of an ant's body length. Per week and per colony they generated weighted time-aggregated networks, using the total duration of interaction as the link weights. They investigated the strength, betweenness and closeness centrality and found that the networks are stable over time, without the queen contributing to the network structure. Also they state that individuals with long lasting interactions seem to have a reduced tendency to move, while mobile ants interact homogeneously with their nestmates. The observed colonies ranged in size from 55 to 58 individuals.

## 2.2. Related Studies

In these studies, each of the observed ant colonies contained a maximum of 200 individuals. This number is relatively small compared to the size of honey bee colonies used in the static analysis approaches.

# Chapter 3

## Methodology

In this chapter, I outline the methodology I applied to reach my research goals. In the first section, I describe the data and the experimental setup. In the second section, I share my findings concerning data quality and the steps needed to obtain a reliable dataset. Then I describe my approach to infer networks. The last section explains the methods I used to analyze the properties, communities, and development of the resulting networks.

### 3.1 The Experimental Setup and Dataset

The dataset is derived from high-resolution video files that capture tagged honey bees of one colony in a single frame observation hive. The bees are uniquely tagged with circular 12-bit markers (Figure 1.1, section 1). Two cameras per side filmed the complete honeycomb. Figure 3.1 illustrates the camera setup. The *recording period* lasted nine weeks (63 days), from July 19, 2016 until September 19, 2016, with some interruptions due to maintenance work and technical failures. An overview about the complete recording period is given in Figure B.2.

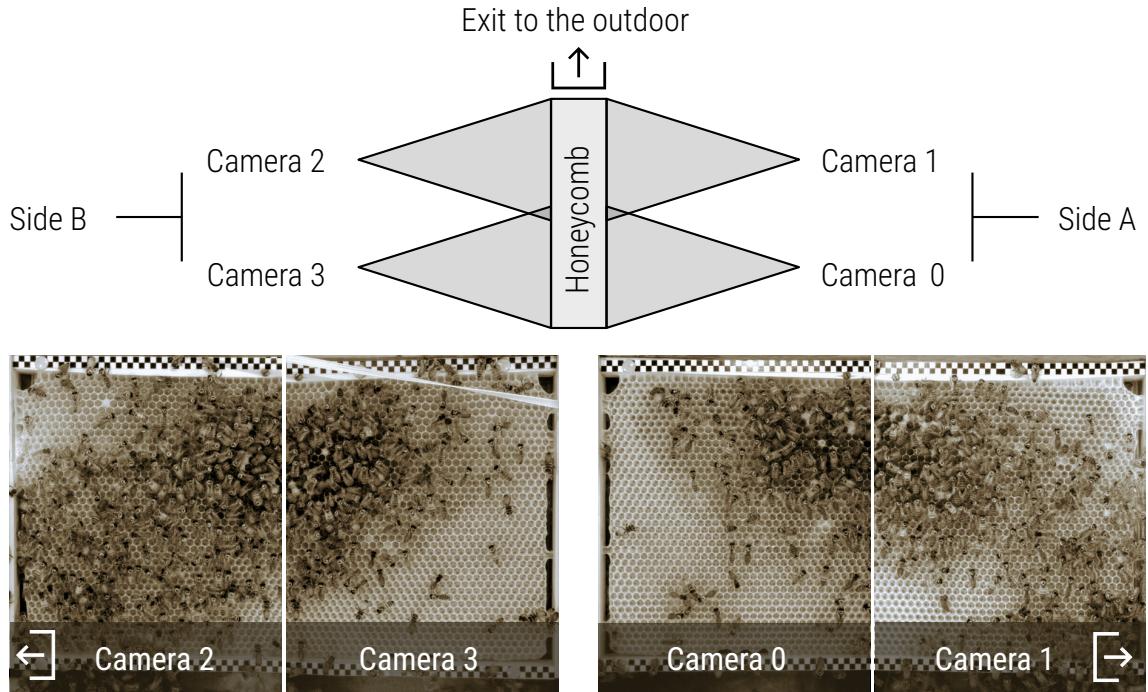
All four cameras, each with a resolution of  $4000 \times 3000$  pixel, recorded 3.5 frames per second. An image analysis pipeline [1] detects all bees in each frame. The resulting detection data is stored in a binary file format. A python library<sup>1</sup> provides a frame-level access to those binary files. The size of the dataset is 470 GB, about 7.5 GB of binary data per day.

The 67 day *tagging period* began on June 28, 2016 and lasted until September 2, 2016, resulting in 3,191 tagged bees. Bees were already tagged three weeks before the observation started. The young bees, which were raised in a separate incubator, were tagged and then added to the observation hive at noon each day. Figure B.1 shows the frequency of tagged bees per day. The hatching day for each bee is documented and therefore the age of each bee at a particular point in time can be calculated. The life expectancy of a honey bee during summer ranges from 30 to 60 days, according to Menzel and Eckoldt [37, p. 27]. The maximum number of present bees in the hive is about 1,600.

---

<sup>1</sup>The library is called bb-binary and is created by the Biorobotics Lab. It can be found on GitHub: [https://github.com/BioroboticsLab/bb\\_binary](https://github.com/BioroboticsLab/bb_binary); Last accessed: February 16, 2016; 4:28 p.m.

### 3.2. Data Quality and Data Cleansing



**Figure 3.1: Observation setup** Each side of the honeycomb is filmed by two cameras. The two cameras per side overlap, so bees inside this area are detected from both cameras.

## 3.2 Data Quality and Data Cleansing

In this section, I describe the data scheme and investigate the quality of the tracking data and completeness of bee tracks. I propose a way to filter invalid detections to gain a cleaned and valid dataset, which can be used to infer networks. Frequently used terms are listed in Table 3.1.

### 3.2.1 Data Scheme

The data is organized in *frame containers*. Each frame container corresponds to one video file from a particular camera and contains approximately 1,024 frames. Each *frame* contains a list of bees, which were detected by the image analysis pipeline. A bee *detection* includes following attributes:

<b>x pos</b>	<i>x</i> coordinate of bee with respect to the image in pixel
<b>y pos</b>	<i>y</i> coordinate of bee with respect to the image in pixel
<b>decoded ID</b>	decoded 12-bit ID
<b>cam ID</b>	ID of the camera: 0, 1, 2, 3
<b>timestamp</b>	unix timestamp with milliseconds

**Table 3.1: Terms related to the dataset**

<b>Frame container</b>	Contains all frames, which belong to a specific video file of a certain camera.
<b>Frame</b>	A frame is one picture of one camera and includes all bee detections.
<b>Detection</b>	Detection of a bee at a certain point in time.
<b>Decoded ID</b>	Identifier of a bee consisting of twelve probability values, representing twelve bits.
<b>ID</b>	Decimal representation of a decoded ID.
<b>Bee time series</b>	Binary sequence, indicating the absence and presence of a certain bee in a particular time interval.
<b>Pair time series</b>	Binary sequence, indicating whether or not two bees are close to each other, in a particular time interval.

The data can be accessed by iterating on the frame level, using a start and end timestamp for specifying time interval. The complete data scheme can be found on GitHub<sup>2</sup>.

### 3.2.2 ID Probabilities, Confidence Level, and Quality

Twelve bits can encode the identity of 4,096 bees. Each bit of the decoded ID represents a probability between 0 and 255, normalized to a value between 0 and 1. Therefore, a bit indicates the confidence of the image analysis pipeline for that specific bit. I define the confidence  $c$  for a bit  $b$ , analogously to Leon Sixt [38, p. 14], as

$$c(b) = 2 \cdot |b - 0.5| \quad (3.1)$$

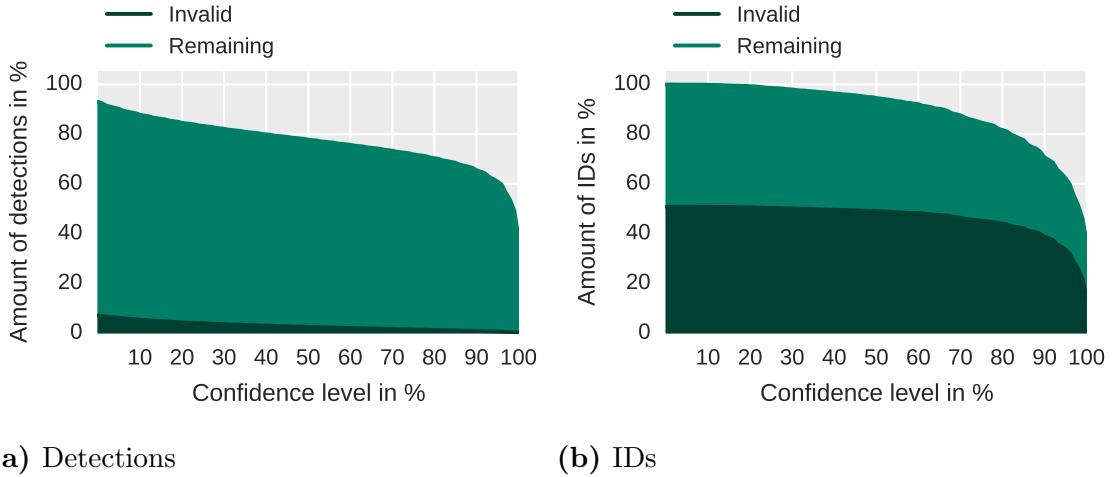
The confidence of a decoded ID is, accordingly, the minimum of all twelve bits' confidences. Detections with a confidence below a certain level are removed from the dataset. Consequently, a high level of confidence reduces the amount of data available for further processing.

I use the age of the bees to check the quality of the remaining data. If the pipeline detected a code that has not been assigned a bee will have a negative age value. A bee has a negative age, if the pipeline detected a code, that was not used yet. I examined the number of remaining bee detections and IDs, depending on the chosen confidence by calculating the age of each bee detection and ID. A bee detection with a negative age is counted as an *invalid detection* and an ID with a negative age is counted as an *invalid ID*.

As expected, with increasing confidence levels, the number of remaining detections

---

<sup>2</sup>[https://github.com/BioroboticsLab/bb\\_binary/blob/master/bb\\_binary/bb\\_binary\\_schema.capnp](https://github.com/BioroboticsLab/bb_binary/blob/master/bb_binary/bb_binary_schema.capnp); Last accessed: February 16, 2016; 4:46 p.m.



**Figure 3.2: Quality of detections and IDs** *Light green* represents the number of remaining data and *dark green* indicates the fraction of invalid data. (a) Proportion of remaining and invalid detections. (b) IDs, that are detected at least once, in relation to invalid IDs. [Dataset: ten minutes; four cameras; July 26, 2016; 4 p.m.]

and IDs decreases (Figure 3.2), as does the fraction of invalid detections and IDs. With a confidence level of 100%, the fraction of invalid detections reaches 2.5%. However, the fraction of invalid IDs detected during a time interval of ten minutes remains at the high value of 30.2%. Consequently, selecting a high level of confidence is not sufficient. To obtain a more reliable dataset, invalid detections need to be filtered out.

### 3.2.3 Detection Frequency Filter

A good indicator if a bee detection represents a real bee on the comb is the detection frequency of its ID. Individuals with a very low detection rate may be detection errors. To check this hypothesis, I investigate the correlation between the detection frequency of bees and their age. Figure 3.3 shows that bees with a negative age are observed less often than bees with a positive age.

During my analysis, I noticed the existence of a group of bees with a negative age and a high detection frequency. I inspected the corresponding photos and confirmed that those bee detections correspond to living individuals and are not artifacts. This results likely from a mistake in the reported hatching date for each bee. Consequently, these bees were excluded from the analysis. Also I excluded bees ( $n = 10$ ), whose age is unknown.

For each analysis day, the number of detections per ID is calculated, excluding the mentioned IDs. To obtain a reliable dataset, I filtered invalid detections, by discarding all detections with an ID frequency below the 99th percentile of the negative IDs.

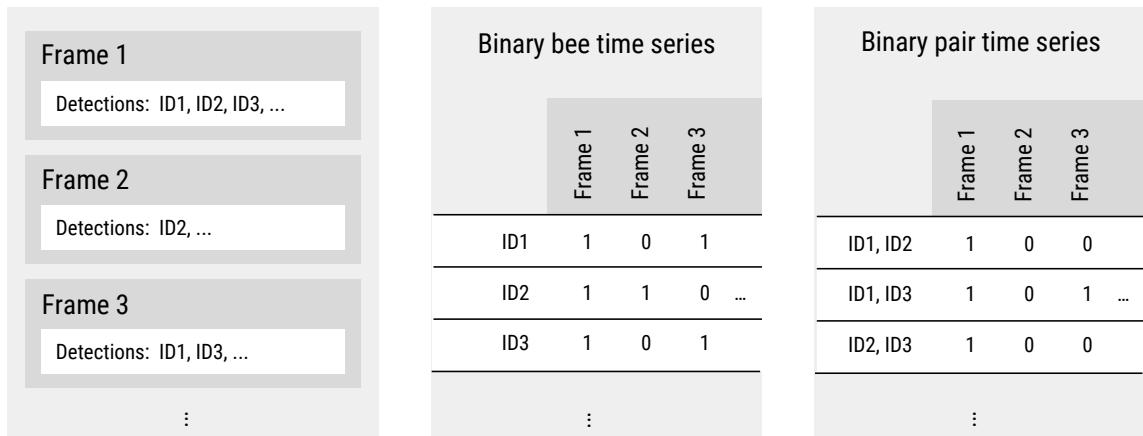
### 3.2. Data Quality and Data Cleansing



**Figure 3.3: Detection frequency of IDs** *Orange* corresponds to bees with a negative age and *green* displays bees with a positive age. The *gray* line represents the 99th percentile of bees with a negative age. [Dataset: August 20, 2016; 24 hours, number of total frames: 302,400]

#### 3.2.4 Time Series of Bees and Bee Pairs

I investigated the quality of the initial data regarding its completeness of bee tracks. A bee track represents the movement of an individual over time. I transformed the initial dataset into binary *bee time series*, depicted in Figure 3.4 left and middle. A bee time series, similar to a track, represents the absence and presence of a bee over a specified sequence of frames. For further processing I use the bee time series to extract *pair time series* of bees that are spatially close (Figure 3.4, right). A one indicates that a pair of bees is detected and both bees are spatially close in a certain frame.



**Figure 3.4: Structure of dataset** *Left:* original dataset - containing a sequence of frames with bee detections; *Middle:* binary bee time series - zero and one indicate absence and presence of a bee; *Right:* binary pairs time series - zero and one indicate the absence and presence of two bees in the same frame.

By analyzing the resulting pair time series, I noticed that detection sequences were often interrupted by short intervals without valid detections. As stated before, the higher the level of confidence, the more data is discarded. This data reduction leads to more zeroes (gaps) in both time series. Gaps in the pair time series frequently

correspond to gaps in one or both bee time series and are thus the result of missing detections of the required confidence and do not represent any meaningful behavior of the bees. Bees are not able to approach each other and move apart within a second because they simply do not move that fast. Therefore, I concluded, that those gaps originate from detection errors and consequently need to be treated in an appropriate way during further data processing.

## 3.3 Inferring Spatial Proximity Networks

The network inference was driven by a combination of an exploratory data analysis and an iterative pipeline development process. It serves as a prerequisite for the network analysis part of this thesis. To generate functional and non-functional requirements of the pipeline, I conducted an analysis of the tracking data and a literature review, presented in Section 2.2. The analysis led to a general understanding of the data, its structure, characteristics and an estimation of its quality. The purpose of the literature review was to get an overview of the common methods and approaches regarding network analysis in the field behavioral insect studies.

Both results are then used to select a network type and to define its nodes and links. Furthermore, I inferred specific pipeline parameters and decided for the procedure of network extraction. The pipeline was developed, tested and refined in an iterative process. Accordingly, the results of the evaluation lead to new or changing functional requirements. The evaluation is conducted by reviewing the pipeline parameters' effects on network properties and checking the validity and quality of the networks by investigating the age of bees in the resulting network.

### 3.3.1 Defining the Network and its Parameters

As this work constitutes the first step towards network analysis using this tracking data I chose to infer a time-aggregated spatial proximity network. Accordingly, the interactions are undirected but weighted. Methods for analyzing static networks are widely established. Static tools and algorithms are already implemented and used by a large community. The results of the network analysis are easy to understand and interpret. Additionally, static networks are the precondition for applying traditional community detection algorithms. My choice also establishes the comparability with Mersch, Crespi, and Keller [4].

Each node in the network represents a bee, identified by an ID. The network consists only of bees that interact with other bees at least once, during the specified time interval. Two bees are associated (spatially close to each other) if their distance is smaller than a *maximum distance*. Using only this criterion leads to many interactions, resulting in a very dense network, because an interaction could only last for 0.33 seconds. Therefore, an additional parameter, the *minimum contact duration*, is introduced. It specifies the minimum time two bees have to spend close to each other to be called associated. Links are assigned two attributes. The first one is the

frequency of contacts, meaning how often they share a close position. The second parameter refers to the total duration of all contacts.

The network pipeline takes two types of parameters. The first set of parameters defines the resulting network and the exact type of spatial proximity. The second set relates to the given data. Both parameter types are described below.

#### Pipeline parameters for network

<b>Maximum distance</b>	Level of closeness between individual bees. (in pixel)
<b>Minimum contact duration</b>	The number of frames two individuals need to spend close to each other to count it as an interaction (in frames)
<b>Start timestamp</b>	Starting point of the network aggregation. (as UTC string)
<b>Window size</b>	Size of time window for aggregating the network. (in minutes)

#### Pipeline parameters for data

<b>Confidence</b>	Level of confidence, as described in Section 3.2.2. (in percent)
<b>Valid IDs</b>	List of valid IDs within a specified time interval, as described in Section 3.2.3. (in CSV file format)
<b>Gap Size</b>	Gaps in time series of bee pairs are assumed to be the result of missing detections. Gaps of this size are filled up. (in frames)
<b>Number of CPUs</b>	Number of used CPUs for parallelization.
<b>Year</b>	Calculate bee IDs and stitching of camera images according to the observation period. (2015 or 2016)

### 3.3.2 Implementing the Network Pipeline

This section describes the pipeline for generating spatial proximity networks out of honey bee tracking data. The input for the network pipeline are the data and a set of parameters previously described. It outputs a network in graphML file format. I parallelized the pipeline on the frame level. That means each process gets a fraction of the data and extracts interactions for a time interval of five minutes. The main process accumulates the resulting interactions and creates a final network.

The pipeline consist of the following steps:

#### 1. Prefilter detections

All detections below the chosen level of confidence are filtered out.

#### 2. Simple stitching

For each camera positioned on the right hand side of the hive (camera 1 and 3), the  $x$ -coordinates of each detection is offsetted to the right ( $2 \times$  maximum distance), to combine all detections per side into one coordinate system.

**3. Synchronize cameras**

The cameras do not capture all images synchronously and sometimes take more or less than 3.5 images per second. Therefore, the cameras per side need to be synchronized. In the normal case the difference between consecutive frames should be about 0.332 seconds, due to technical problems this value can be lower (0.003) and higher (2.932) at certain times. The frames of cameras 3 and 2 and cameras 1 and 0 are matched, frames without a match are discarded.

**4. Discard detections with certain IDs**

All bee detections with an ID contained in a valid ID list are kept, all other detections are discarded.

**5. Extract close pairs**

For each side of the hive, all close bee pairs according to the maximum distance parameter are calculated using an KDE-tree.

**6. Generate time series of bee pairs**

The data structure (frames and pairs of bees) is transformed to time series of bee pairs.

**7. Correct pair time series.**

The time series of bees are corrected by filling in the gaps of less than a certain length (gap size parameter).

**8. Extract interactions**

The links and link attributes (frequency and duration) are extracted from the time series of bees using the minimum contact duration parameter. A sequence of ones with a length of at least the minimum contact duration are evaluated as an interaction. The frequency of those series and the total duration (number of ones) are the attributes.

I used python to implement the network pipeline because the library to access the tracking data is only available in python. The data analysis is done with pandas<sup>3</sup> and numpy<sup>4</sup> using the web-based environment Jupyter Notebook<sup>5</sup>. The networks are saved in graphML format and are created using the python library networkX<sup>6</sup> version 1.11. iGraph<sup>7</sup> was used for community detection and for network analysis.

### **3.3.3 Choosing Parameter Values for Network Analysis**

For network analysis, I chose three days: August 20, August 22, and August 24, 2016. These days were selected because bees from a wide range of age groups were present and older bees, which are likely to be foragers, were more represented in the hive during these days. Additionally, no data is missing due to camera failures. The following values are chosen according to biological constraints and similar to other

---

<sup>3</sup><http://pandas.pydata.org/>

<sup>4</sup><http://www.numpy.org/>

<sup>5</sup><http://jupyter.org/>

<sup>6</sup><https://networkx.github.io/> Last accessed: 2017-02-17, 08:07PM

<sup>7</sup><http://igraph.org/python/>

**Table 3.2: Parameters chosen for network analysis** The maximum distance corresponds to the length of a bee body and the minimum contact duration is about one second. The networks are aggregated for ten hours.

Parameter	Value	Unit
Maximum distance	212	px
Minimum contact duration	3	frames
Window size	600	minutes
Confidence	95	percent
Gap size	2	frames

studies, for better comparability.

I chose the length of a bee body, according to Baracchi and Cini [28], as the maximum distance between two bees (Figure 3.5). The average bee length of 212px ( $\pm 16$ px) was determined by manually measuring the length of all bees ( $n = 337$ ) in images from the four cameras using the tool ImageJ<sup>8</sup>. The minimum contact duration is set to three frames (one second). This value corresponds to Mersch, Crespi, and Keller [4], as they also exclude interactions below one second. The networks are aggregated for ten hours during daylight; this corresponds to the biological rhythm of bees.

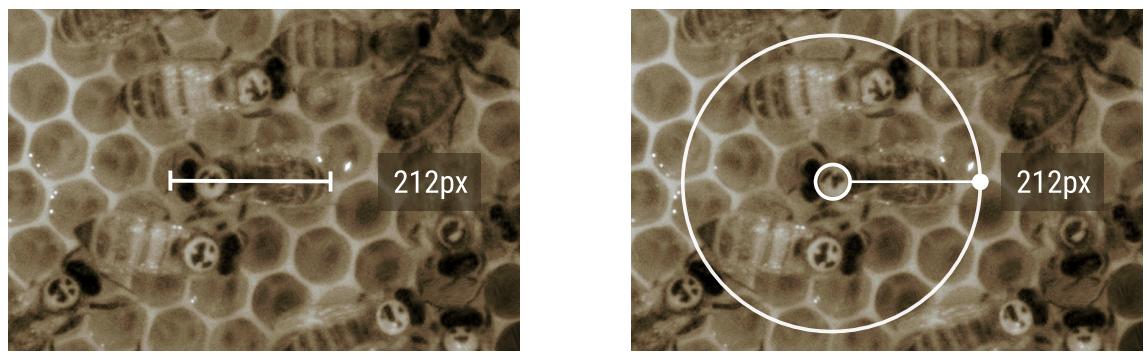
The confidence level is set to 95%, which will keep about 60% of the data. The gap size is set to two frames. This value corresponds to the median gap length in the time series of bee pairs.

### 3.3.4 Summary

The goal, as mentioned in 1.2, was to answer the question whether it is possible to infer temporal networks with the provided honey bee tracking data and to work out

---

<sup>8</sup><http://imagej.net/Welcome>; Last accessed: February 22, 2016; 12:34 p.m.



**Figure 3.5: Maximum distance of bees:** A length of a bee body is chosen as the maximum distance between two bees.

challenges and limitations regarding the provided dataset. Furthermore, it was a goal to identify the parameters necessary for the pipeline.

#### Pipeline Parameters

This analysis results in two types of pipeline parameters. The first category specifies the resulting network, concerning the definition of spatial proximity, duration of interaction and size of the aggregated time window. The second type represents parameters resulting out of the characteristics of the dataset.

##### 1. Pipeline parameters for network

maximum distance, minimum contact duration, start timestamp, window size

##### 2. Pipeline parameters for data

confidence, list of valid IDs, gap size, number of CPUs, year

#### Limitations

It is possible to infer networks, but a complex preprocessing of the dataset is essential with two major steps:

##### 1. Reduction of data

Reduce the amount of data to obtain a reliable dataset, by filtering out detections with a low confidence value or by IDs with a low detection frequency.

##### 2. Combine camera data

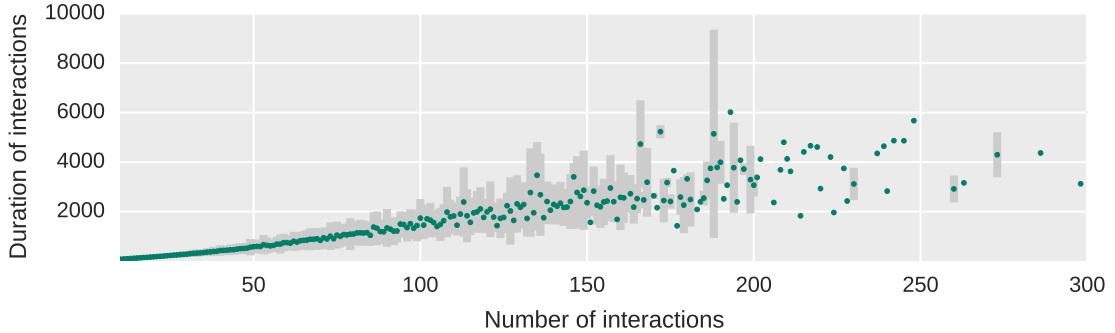
This step consists of the time synchronization of each of the two cameras and the joining of the data per frame.

A tradeoff exists between the remaining amount of data that can be used for network inference and the data's. A high confidence value reduces the amount of data and produces gaps. The gap size parameter tries to fix this problem.

It is also possible to infer time-aggregated networks, but with restrictions. When limiting the window size for network aggregation to the biological rhythms of day and night, only a small amount of useful analysis days remain due to a large number of interruptions. Any other window size entails the inclusion of the duration of biological processes related to honey bees, I would need to know beforehand. Alternatively, I would need to apply a method to infer an appropriate window size out of the given data, this is out of scope.

## 3.4 Methods for Analyzing Spatial Proximity Networks

This section outlines the methods I used to investigate the networks on a global, intermediate and local level. I present the choice of network measures used for the



**Figure 3.6: Frequency and total duration of interaction** The two link weight values show a strong positive correlation. The data of the three snapshots is aggregated.

global analysis and explain the decision to use a community detection algorithm. I illustrate the methods to examine the segregation of communities by age and spatial distribution. Furthermore, I describe the approach used to study the development of communities.

### 3.4.1 Investigating the Topology and Network Characteristics

I summarized the network analysis methods of the reviewed studies (Chapter 2.2) to gain an overview of established procedures in the field of social insect networks. I grouped the methods by global measures, node level measures and other network analysis methods. Table A.3 summarizes the network analysis methods applied by the reviewed studies. The network measures I chose for the global and node level analysis are listed in Table 3.3 and were defined in Chapter 2.1.1.

Each link in the network is attributed with the frequency of interactions and total duration of interactions between the two individuals. Figure 3.6 shows a strong positive correlation between those two values. For further analysis I decided to use the frequency of interactions as the weight for links, analogously to [4], [28].

The degree of a bee represents the number of other bees this focal animal interacts with. Bees with a high number of interaction partners have a high degree. This measure was chosen because it reveals a lot about the general topology of a network. The strength of a bee is the sum of its link weights. A high strength refers either to a large number of interaction partners with a low link weight or a low number of interaction partners with high link weights. Especially for aggregated networks this measure accumulates valuable information regarding the interaction activity of bees. The local clustering coefficient (lcc) of a bee indicates how close its interaction partners are to form a complete subgraph. A large lcc shows that most of its interaction partners interact with each other. A low lcc indicates the absence of those interactions. It is a good indicator for the embeddedness of single bees.

The betweenness of a bee measures the number of shortest paths between two other bees that go through that bee. A bee with a high betweenness would be important for the information flow of the network. Removing this bee from the hive would

**Table 3.3: Measures used for analysis** Each measure is explained in Chapter 2.1.1

Global level measures	Node level measures
Number of nodes $N$ and links $E$	Degree $k$
Average degree $\langle k \rangle$	Strength $s$
Average strength $\langle s \rangle$	Local clustering coefficient $c$
Density $D$	Closeness centrality $C_C$
Diameter $d_{max}$	Betweenness centrality $C_B$
Number of components	
Global clustering coefficient $c_\Delta$	
Average shortest path length $\langle d \rangle$	
Link weights $w$	

lead to the breakdown of information or food flow and would negatively affect the robustness of the network. The closeness of a bee measures how fast this bee can reach all others in the network. A high closeness would indicate a very short path to every other bee. A bee with high closeness can spread information to all other bees very quickly.

I examined all node level measures concerning the age of bees and their detection frequency. The global network measures are compared to an Erdős-Rényi random network, by averaging over 100 runs.

### 3.4.2 Detecting Communities

For finding an appropriate community detection algorithm, I checked the reviewed studies for applicable methods and scanned papers that compare various community detection algorithms. I identified a subset of algorithms as suitable and checked them.

The reviewed studies only include one examples of community detection and one example of cluster analysis. Mersch, Crespi, and Keller [4] used the infomap [39], [40] algorithm. According to the authors, this algorithm only works with sparse networks and is therefore not applicable in my case of densely connected spatial proximity networks. Baracchi and Cini [28] use a hierarchical clustering to infer groups of bees within the network that are similar in strength, eigenvector, and betweenness centrality. In contrast to the resulting groups of community detection, groups identified by hierarchical clustering do not automatically refer to dense subgraphs of the network.

Abundant literature on the comparative analysis of community detection algorithms exists, e.g. [41], [42]. Some studies seem to be promising for choosing an appropriate algorithm, but assume either a power law degree distribution or evaluate networks with a low density, which is not applicable here. Thererfore, I tested community detection algorithms, implemented in python, to find an algorithm that works well for animal social networks. The three most common python libraries for network

### 3.4. Methods for Analyzing Spatial Proximity Networks

analysis were reviewed: NetworkX<sup>9</sup>, igraph<sup>10</sup>, and graph-tool<sup>11</sup>. The algorithm needs to fulfill the following criteria: support for large ( $(N = 1000)$ ) and dense ( $D > 50\%$ ) networks, support for weighted links, as well as a fast runtime.

**Table 3.4: Comparing community detection algorithms** Comparison of algorithms implemented in python. Criteria are the support of weighted links, runtime and number of communities. A runtime indicated by “–” means no termination after 15 minutes.

	Fastgreedy <sup>1</sup>	Leading eigenvector <sup>1</sup>	Louvain <sup>2</sup>	Multilevel <sup>1</sup>	Walktrap <sup>1</sup>	Infomap <sup>1</sup>	Label propagation <sup>1</sup>	Edge betweenness <sup>1</sup>	K-clique communities <sup>2</sup>	Optimal modularity <sup>1</sup>	Spinglass <sup>1</sup>	Statistical inference <sup>3</sup>
Link weights	×	×	×	×	×	×	×	–	×	×	×	–
Runtime in sec	3.6	6.3	11.7	0.7	19.4	13.2	0.2	–	–	–	–	–
Communities	3	2	2	3	2	1	1	–	–	–	–	–
Size	473 434 15	488 434 33	469 453 (1)	462 427	490 431	922 922	922	–	–	–	–	–

<sup>1</sup> igraph, <sup>2</sup> NetworkX, <sup>3</sup> graph-tool

Table 3.4 gives an overview about the algorithms reviewed. Five algorithms did not terminate after 15 minutes and were therefore excluded from further investigations. Infomap and label propagation tend to partition all nodes into a single community, this is known, especially in dense graphs [13], [41]. The Louvain algorithm implemented in networkX is the same as multilevel implemented in igraph, but takes longer and produces almost the same communities and was also excluded. Walktrap was tested for different step size parameters, as suggested in [15], the communities remained almost the same, only a few nodes switched communities.

I examined the number and size of detected communities for the algorithms fast-greedy, leading eigenvector, multilevel, and walktrap for all three snapshots. Table 3.5 gives an overview of the results. All algorithms found at least two communities. Except for leading eigenvector, most tend to find three communities.

I decided to use two algorithms for community detection: leading eigenvector and walktrap. Farine and Whitehead [7] explain that leading eigenvector is often used

<sup>9</sup><https://networkx.github.io/>; Last accessed: March 16, 2016; 6:36 p.m.

<sup>10</sup><http://igraph.org/python/>; Last accessed: March 16, 2016; 6:38 p.m.

<sup>11</sup><https://graph-tool.skewed.de/>; Last accessed: March 16, 2016; 6:39 p.m.

### 3.4. Methods for Analyzing Spatial Proximity Networks

**Table 3.5: Number of community members per algorithm and snapshot** Four algorithms were tested and compared regarding the number of detected communities and the size of the communities.

	Fastgreedy	Leading eigenvector	Multilevel	Walktrap
Snapshot 1	473	488	462	490
	434	434	427	431
	15		33	(1)
Snapshot 2	504	503	481	372
	467	475	439	311
	7		58	294
				(1)
Snapshot 3	534	537	505	310
	388	385	415	390
	(2)		231	

with animal social networks and works well. Walktrap is also chosen for also examining the third community (Table 3.5).

### **Age and Spatial Distribution of Communities**

To answer the question whether communities reflect different age groups, I examined the average age and the general age distribution of the communities. I also investigated, if the age division persists in each snapshot. A two-sample Kolmogorov-Smirnov test was used to determine the statistical difference of the age distribution between communities.

To investigate whether communities reflect groups of bees working in different areas of the comb, I used heat maps to determine the core regions per group. I stored the positions of all bees present within the ten hour time windows in an SQLite database for faster access to the data and to eliminate the time-consuming parsing.

#### **3.4.3 Development of Community Members**

According to Aynaud, Fleury, Guillaume, *et al.* [43] and Bródka, Saganowski, and Kazienko [44], there are three main approaches for community detection in temporal networks (sometimes referred to as community tracking): (1) using a static community detection algorithm on several snapshots and then solving a matching problem, (2) using algorithms that are directly suited for temporal networks and (3) using incremental or online algorithms when processing data streams. For each of the three approaches, several methods already exist. As community tracking is not the main focus of this work, I chose to apply the most natural method: detecting static communities for each snapshot and then matching those communities using set theory.

Two communities at successive times are matched if they share enough nodes. The *match value* between two communities  $C$  and  $D$  according to Hopcroft et al. [45] is defined as

$$\text{match}(C, D) = \min \left( \frac{|C \cap D|}{|C|}, \frac{|C \cap D|}{|D|} \right). \quad (3.2)$$

This value is between 0 and 1. A high match value occurs, if two communities share many nodes and are of a similar size. Communities with high match values represent the same community at different points in time. The author suggests applying a threshold to more precisely define what “share enough nodes” means. Otherwise, communities with only 0.1% of nodes overlapping could be matched.

To investigate the total number of bees that remain in the network over the three snapshots, I inspected the match value of bees in consecutive snapshots. I also calculated all match values between communities in consecutive snapshots and the number of intersecting bees. I visualize the dynamic movement of bees between

### 3.4. Methods for Analyzing Spatial Proximity Networks

groups of different snapshots with a flowchart diagram using the JavaScript library D3.js<sup>12</sup>.

---

<sup>12</sup><https://d3js.org/>; Last accessed: April 19, 2016; 3:18 p.m.

# Chapter 4

## Results of Network Analysis

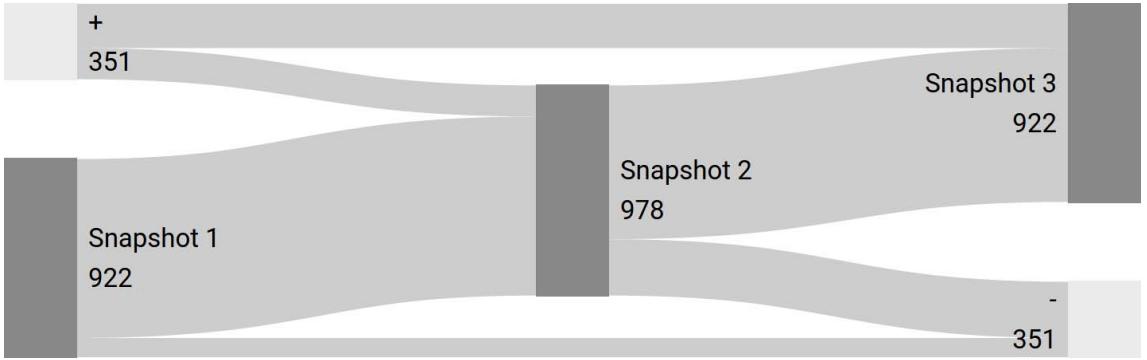
This chapter summarizes the results of the analysis of the temporal, spatial proximity network of honey bees, consisting of three consecutive time-aggregated snapshots. Section 4.1 describes static aspects of the network and Section 4.2 focuses on the temporal network aspects. The last section of this chapter summarizes the main results and discusses the findings.

### 4.1 Static Perspectives of Honey Bee Networks

The networks are examined on three levels. First, I examine the network's global structure and derive properties of the overall colony (global level). Next, I study the characteristics of individual bees (local level) and the relation of the characteristics to detection frequency and age. Finally, I investigate the intermediate level of the colonies social organization by detecting communities and inspecting their practical meaning. I analyzed a temporal network, consisting of three time-aggregated snapshots; these are referred to below as snapshot 1 ( $N = 922$ ), snapshot 2 ( $N = 978$ ) and snapshot 3 ( $N = 922$ ). The snapshots are aggregated for ten hours (108,000 frames) starting at 8 a.m. and lasting until 6 p.m. See Table 4.1 for details about the bees added each day. Figure 4.1 shows the proportion of intersecting bees between consecutive snapshots. This figure illustrates the stability of the network size.

**Table 4.1: Sampling period** Overview of the chosen aggregated daily snapshots including the number of added bees and the time they were added to the hive.

	8/20/2016	8/21/2016	8/22/2016	8/23/2016	8/24/2016
Snapshot ID	1	—	2	—	3
Number of bees added	0	0	110	60	0
Time added	—	—	2 p.m.	6 p.m.	—



**Figure 4.1: Transitions of bees per snapshot** Each column represents a time step, the dark gray rectangles represent the snapshot for each step, and the height of the rectangles corresponds to the amount of members, as referenced by the number. The light gray boxes represent the number of bees that are added to the colony and bees that disappear. This figure shows the amount of bees for each snapshot and the proportion of intersection.

### 4.1.1 Properties of the Bee Colony

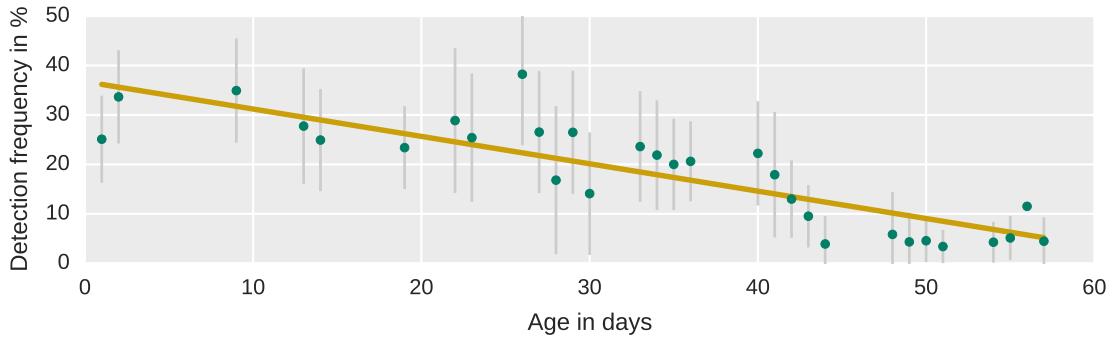
Each snapshot consists of one component. The density  $D$  is over 50% for all snapshots (69%; 54%; 61%). The diameter  $\langle d_{\max} \rangle$  is 3 and the average shortest path length  $\langle d \rangle$  is between 1 and 2. The global clustering coefficient (gcc)  $c_\Delta$  of all snapshots is higher than compared to an Erdős-Rényi random graph, averaged over 100 runs using the same number of nodes and links. On average, each bee is connected to at least 50% of the colony (68%; 52%; 61%). During the ten-hour observation period, a bee interacts over 4,000 times (5,680; 3,978; 4,206) on average. Table 4.2 summarizes the basic network properties for each snapshot and lists the values of its corresponding random graph.

For further analysis, I select snapshot 3 because no young bees were added to the colony during that day and, unlike snapshot 1, bees below the age of five days were part of the colony (Figure C.10). Figure 4.2b shows the age distribution of snapshot 3. This distribution corresponds to the artificial tagging of the bees. Consequently, bees of certain age groups are simply not present. The detection frequency of an individual bee is negatively correlated with its age (Figure 4.2a). The link weight distribution is shown in Figure 4.2c. Most links have a low weight; only a few links have a high weight. The logarithmized frequency distribution of the link weights appears to be an exponential decay function with an exponent of 0.015. The fitted plot is shown in Figure C.2.

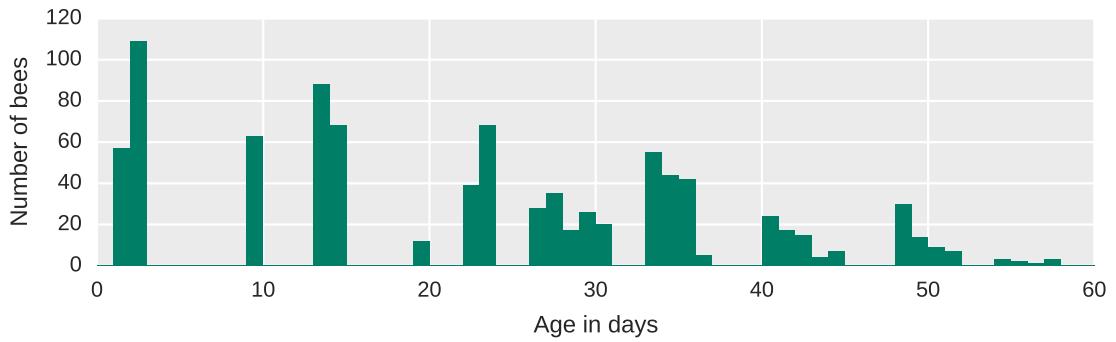
### 4.1.2 Characteristics of Bees

I inspected the properties of each honey bee in snapshot 3. Properties included in the analysis were its degree  $k$ , strength  $s$ , local clustering coefficient (lcc)  $c$ , betweenness centrality  $C_B$  and closeness centrality  $C_C$ .

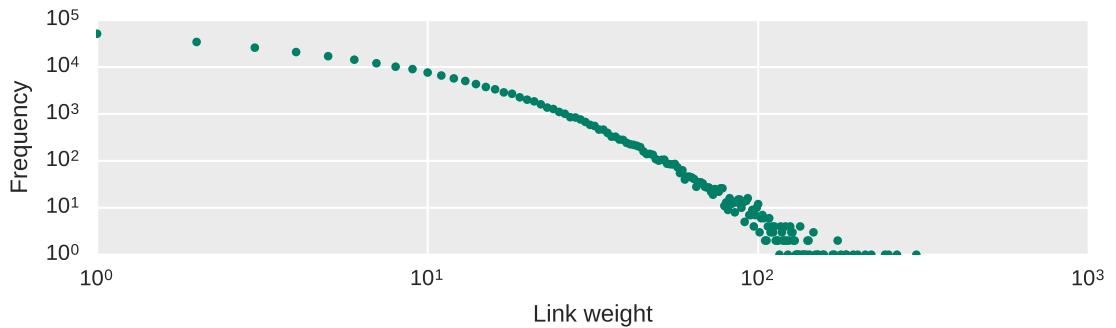
#### 4.1. Static Perspectives of Honey Bee Networks



(a) Correlation of detection frequency and age



(b) Age distribution



(c) Link weight distribution (log-log plot)

**Figure 4.2: Age distribution, correlation with detection frequency and link weight distribution of snapshot 3** (a) Detection frequency and the age of a honeybee seem to be negatively correlated. (b) The age of bees ranges from 1 to 60 days, but some age groups are missing. (c) The link weight as a log-log plot.

**Table 4.2: Global network properties**  $N$  number of nodes,  $L$  number of links,  $D$  diameter,  $\langle d_{\max} \rangle$  average path length,  $\langle d \rangle$  diameter,  $c_\Delta$  global clustering coefficient,  $\langle k \rangle$  average degree and  $\langle s \rangle$  represents the average strength, as introduced in Section 2.1.1.

	$N$	$L$	$D$	$\langle d_{\max} \rangle$	$\langle d \rangle$	$c_\Delta$	$\langle k \rangle$	$\langle s \rangle$
Snapshot 1	922	291,179	0.69	3	1.32	0.79	631.62	5,680.17
Random 1	922	291,179	0.69	2	1.31	0.69	631.62	-
Snapshot 2	978	256,066	0.54	3	1.46	0.72	523.65	3,977.94
Random 2	978	256,066	0.54	2	1.46	0.54	523.65	-
Snapshot 3	922	259,421	0.61	3	1.39	0.75	562.74	4,205.99
Random 3	922	259,421	0.61	2	1.39	0.61	562.74	-

### Low Hierarchical Structure

The degree is normally distributed (panel (a) in Figure 4.3). Therefore most bees have the same high number of interaction partners. The absence of hubs, a small number of highly connected bees, indicates a low hierarchical structure of the network. Strength and lcc are also normally distributed (panel (d) and (g) in Figure 4.3). That also shows the absence of extreme values and confirms that bees are similar to each other regarding those properties. Closeness and betweenness centrality (panel (j) and (m) in Figure 4.3) also follow a normal distribution. This distribution leads to the assumption that no central or important bees exist. However, this could be a consequence of the definition of interaction (spatial proximity). All bees are similarly close to all other bees in the network, and every bee can reach any other bee with a few steps. That also corresponds to the low average path length, and the small diameter of the network described in Section 4.1.1. The absence of bees with a high betweenness suggests that the colonies functionality is robust concerning the disappearance of single individuals.

### Local Network Measures and Detection Frequency

Degree, strength, closeness and betweenness (panel (b), (e), (k), and (n) in Figure 4.3) are positively correlated with the detection frequency. A low value corresponds to a low detection frequency. In contrast, the lcc (panel (h) in Figure 4.3) and detection frequency are negatively correlated.

### Local Network Measures and Age of Bees

The histograms of degree, strength, betweenness, and closeness show a normal distribution with a tendency for bimodality. The lcc distribution is instead right skewed, with one peak at 0.75.

There is no sharp border between the two modes in the degree distribution plot (a),

but a value around 0.4 can be estimated. The strength histogram (d) seems to have a border at 1,000. For closeness (j) and betweenness (m), a border can be seen at 0.6 and 0.0001. All distributions indicate a small group (100 bees) and a second larger group containing the rest of the colony. The correlation between all measures is depicted in the scatter plot in Figure C.9.

The first small group interacts on average with 20% of the colony and has a very low strength (number of total interactions below 250). The closeness value is compared to the second group smaller but still over 0.5. The betweenness has a small range and is close to 0 for the first smaller group. The second group interacts with about 80% of the colony with an average strength of 5,000. A high strength can result from lots of neighbors with low link weights or a few neighbors with high link weights. As only a few links with high weights exist (Figure 4.2c), the second option can be excluded. The second group is characterized by a very high closeness (0.75) and a still very low betweenness but higher than the first group (0.0005).

All age-correlation plots show a separate group of bees older than 45 days, corresponding to the first smaller group of bees described above. This older group is characterized by a low degree, a low strength, and low closeness and betweenness. In contrast, a high lcc, compared to the younger group is noticeable. The younger group has a high degree and strength, as well as a high betweenness and closeness compared to the first group, but a lower lcc. The high lcc of the older group indicates high connectivity within the younger group and less connectivity between bees of the older group.

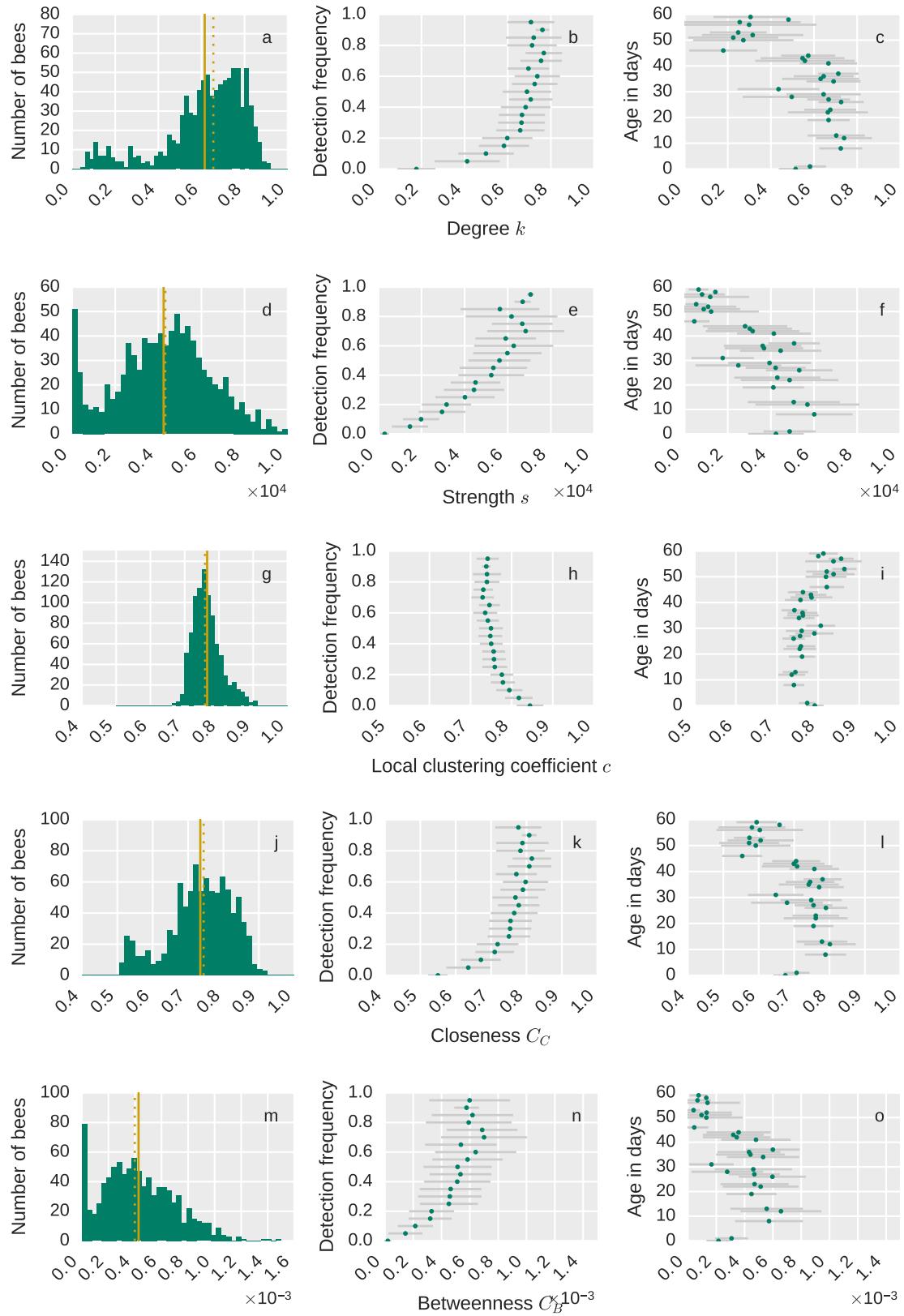
### 4.1.3 Functional Groups within the Colony

The leading eigenvector (LE) community detection algorithms revealed two communities with a similar size (modularity score of 0.25). The walktrap algorithm (WT) discovered three communities instead, also evenly distributed (modularity score of 0.23). Table 4.3 lists the precise number of members per community and algorithm for snapshot 3.

For both algorithms the communities correspond to different age groups. For LE, the average age of the young community is 13.2 days and the average age of the old community is 28.7 days. For WT, the average age of the young community is 6.6 days and 29.3 days for the older community. The average age of the third middle-aged community of WT is 25.1 days. The age distribution for each algorithm is represented in Figure 4.4a and 4.4b. The two sample Kolmogorov-Smirnov test confirmed that the age distributions per community are significantly different. The corresponding  $p$ -values are listed in Table 4.4.

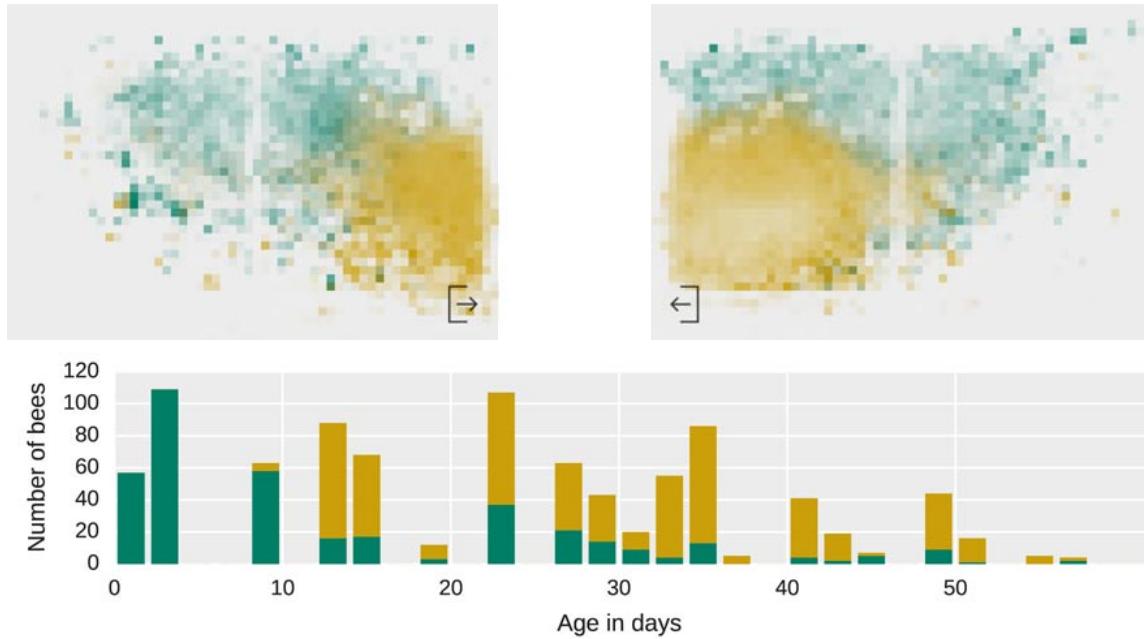
Each community occupies a different region of the comb. Figure 4.4 shows that the young communities spend the most time in the comb center and the old communities closer to the hive exit. The middle-aged community is positioned between the young and old community and in the periphery of the comb.

#### 4.1. Static Perspectives of Honey Bee Networks

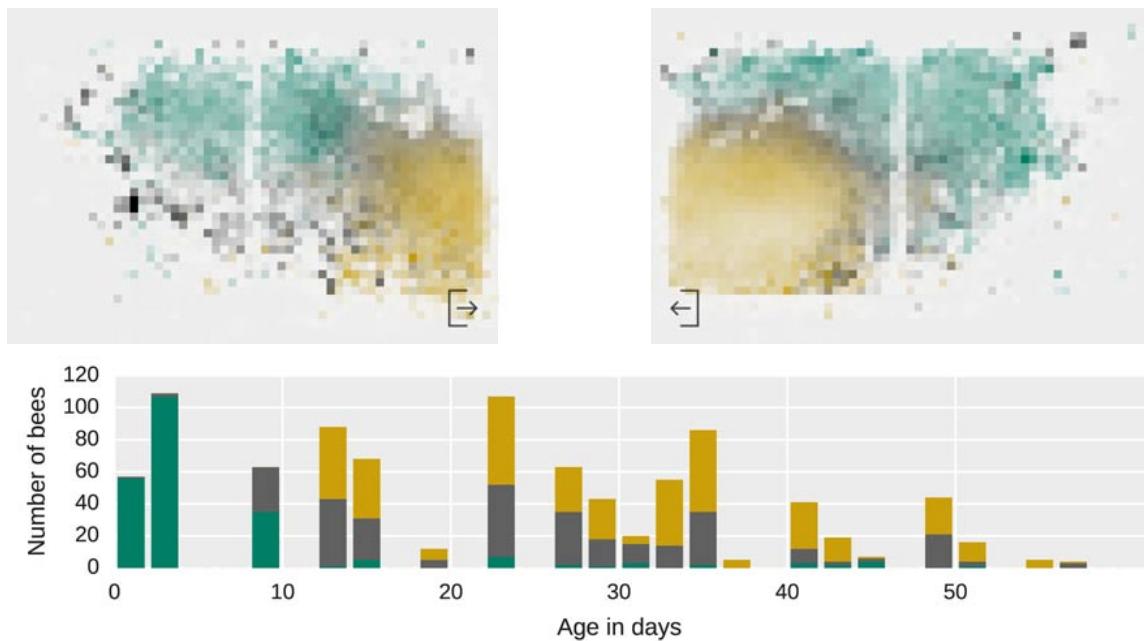


**Figure 4.3: Local measures of snapshot 3**

#### 4.1. Static Perspectives of Honey Bee Networks



(a) Leading eigenvector (LE) communities



(b) Walktrap (WT) communities

**Figure 4.4: Age and spatial distribution of communities** Green represents the young community occupying the center area of the comb and orange the old community, which is situated closer to the hive access. For WT, the gray middle-aged community is positioned between the other two and in the periphery of the comb.

**Table 4.3: Communities per algorithm** Communities marked with \* contain the queen. Age and standard deviation (SD) are measured in days. The queen and nine bees with a negative age are excluded from this analysis.

Community ID		Members	Proportion	Age	SD
LE	CY	*381	41.78%	13.15	$\pm 13.50$
	CO	531	58.22%	28.70	$\pm 11.67$
WT	CY	*229	25.11%	6.55	$\pm 10.36$
	CM	298	32.68%	25.08	$\pm 11.97$
	CO	385	42.21%	29.29	$\pm 11.44$

**Table 4.4: Kolmogorov-Smirnov test**  $p$ -values for leading eigenvector (LE) and walk-trap (WT)

Communities	LE p-value	WT p-value
CY, CO	$5.10 \times 10^{-66}$	$5.51 \times 10^{-67}$
CY, CM		$1.10 \times 10^{-95}$
CM, CO		$1.98 \times 10^{-05}$

## 4.2 Temporal Perspectives of Honey Bee Networks

I investigate the stability of local and global properties, as well as the stability of age and spatial distribution of functional groups of bees. Furthermore, the dynamics of individual bees' group membership over time are examined.

For all three snapshots, the same link weights distribution can be seen in Figure C.1. The analysis of snapshot 1 and 2 showed that the same characteristic distribution of degree, strength, lcc, betweenness, and closeness for snapshot 1 (Figure C.7) and snapshot 2 (Figure C.8) exists. They also follow a normal distribution. The correlation between the local measure and detection frequency and age remains. All of this shows that the characteristics described in Section 4.1.2 apply for all three snapshots and are therefore stable for the investigated time interval. A low hierarchical structure and the correlation with age and detection frequency seem to be global properties of the colony.

### 4.2.1 Stability of Functional Groups

Table 4.5 lists the exact number of bees per community for each algorithm and snapshot. For each snapshot, LE detected two communities with about the same number of bees. The first communities CY(1,2,3) contain the queen and on average younger bees than the second communities CO(1,2,3).

In comparison, WT identified three communities in snapshot 2 and 3 but only two

## 4.2. Temporal Perspectives of Honey Bee Networks

**Table 4.5: Overview about communities per snapshot** Communities marked with \* contain the queen. Age and standard deviation (SD) are measured in days. For each network the queen and bees with a negative age are excluded: snapshot 1: 12 bees, snapshot 2: 119 bees, snapshot 3: 10 bees.

	ID	Members	Proportion	Age	SD
<b>Leading eigenvector (LE)</b>					
Snapshot 1	CY1	*430	47.25%	17.12	$\pm 10.97$
	CO1	480	52.75%	27.24	$\pm 10.96$
Snapshot 2	CY2	*392	45.63%	20.24	$\pm 12.01$
	CO2	467	54.37%	28.10	$\pm 10.88$
Snapshot 3	CY3	*381	41.78%	13.15	$\pm 13.50$
	CO3	531	58.22%	28.70	$\pm 11.67$
<b>Walktrap (WT)</b>					
Snapshot 1	CY1	*427	46.92%	17.07	$\pm 10.92$
	CO1	482	52.97%	27.23	$\pm 11.00$
Snapshot 2	CY2	*263	30.62%	18.23	$\pm 11.46$
	CM2	305	35.51%	25.20	$\pm 11.47$
	CO2	291	33.88%	29.47	$\pm 10.06$
Snapshot 3	CY3	*229	25.11%	6.55	$\pm 10.36$
	CM3	298	32.68%	25.08	$\pm 11.97$
	CO3	385	42.21%	29.29	$\pm 11.44$

communities in snapshot 1. The first communities CY(1,2,3) consist of the queen and on average younger bees than the second CM(2,3) and third communities CO(1,2,3). The bees in CM2 and CM3 are on average younger than the bees in CO2 and CO3. Figure C.5 and C.6 depicts the age distribution for each community and snapshot.

A two-sample Kolmogorov–Smirnov test showed that the age distributions are significantly different ( $p < 0.001$ ) for both algorithms. The spatial segregation of the communities is very similar in all three snapshots. For further reference see the heat maps in C.4 and C.3. The detected communities seem to differ in their respective age and occupy different areas of the comb, but remain stable over this inspected time interval.

### 4.2.2 Dynamic of Individual Bees

Figure 4.5a (LE) and Figure 4.5b (WT) show the flow of bees between consecutive snapshots and communities. For LE communities, the majority of bees stay in their age group, and a small fraction of bees switch to older communities. Only a few bees change to younger communities.

The new middle-aged communities (CM2) of WT are formed equally by members of

**Table 4.6: Kolmogorov-Smirnov test**  $p$ -values for leading eigenvector (LE) and walktrap (WT) for each snapshot and its communities.

		LE p-value	WT p-value
Snapshot 1	CY1, CO1	$2.18 \times 10^{-33}$	$1.52 \times 10^{-32}$
Snapshot 2	CY2, CO2	$2.99 \times 10^{-20}$	$2.3 \times 10^{-32}$
	CY2, CM2		$4.72 \times 10^{-10}$
	CM2, CO2		$1.00 \times 10^{-04}$
Snapshot 3	CY3, CO3	$5.10 \times 10^{-66}$	$5.51 \times 10^{-67}$
	CY3, CM3		$1.10 \times 10^{-95}$
	CM3, CO3		$1.98 \times 10^{-05}$

the young (CY1) and old (CO1) communities. The switching behavior of individuals between communities is similar to LE. Individual bees change communities as they age.

## 4.3 Discussion of Results

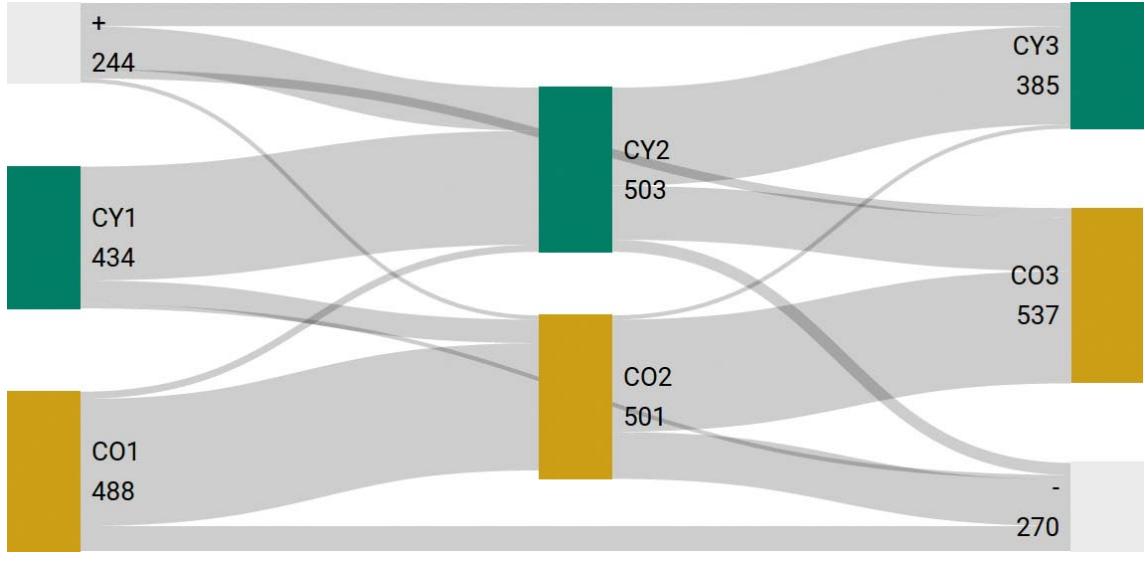
In the following chapter, I summarize and discuss my results considering the current state of research. This part is structured according to the research goals, listed in Section 1.2. First I discuss the topology of the spatial proximity networks of honey bees and its characteristic properties. Secondly, I compare the observed communities and their development over time with existing theories regarding temporal polyethism.

### 4.3.1 Network Topology and Properties of Honey Bee Colonies

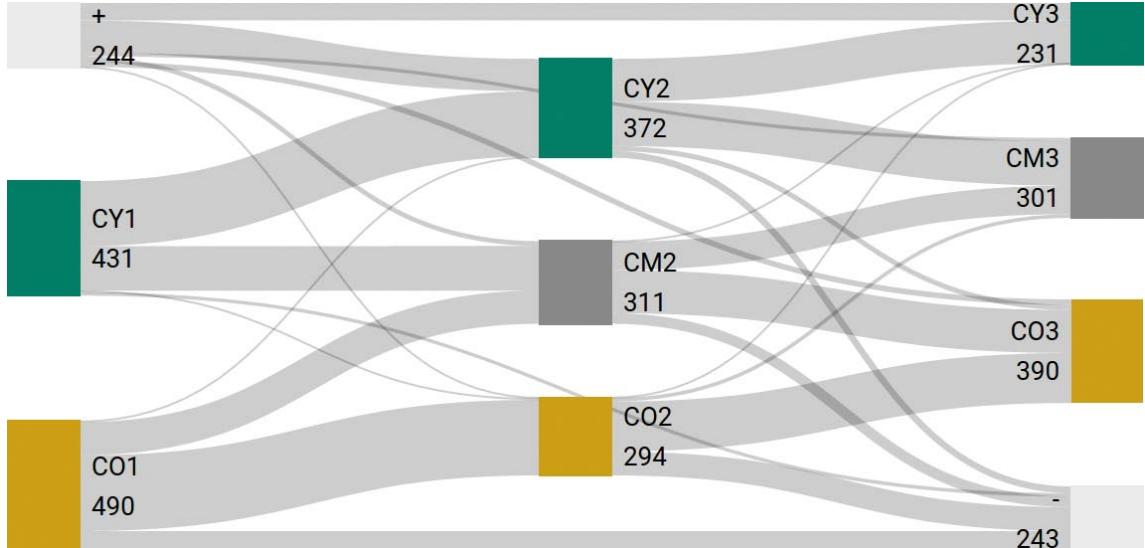
The honey bee spatial proximity networks are characterized by a high density (69%, 54%, 61%), which means the bees encounter many nestmates during the ten hours of data aggregation. This results either from high activity or the fact that the comb is simply very full. The latter increases the probability that two bees are close to each other.

Comparing this result to the ant contact networks of Mersch et al. [4] ( $D = 72\% \pm 5.3$ ), the values are similar. In contrast, when compared to Baracchi and Cini [28] ( $D = 0.15$ ) the density is higher, probably due to their lower observation resolution of one frame per minute.

The small diameter ( $d_{\max} = 3$ ) of my investigated networks and the low average shortest path of 1.4 in combination with a high global clustering coefficient (0.79, 0.72, 0.75) are characteristic for a class of networks known as small-world networks [46]. This type of networks allows for rapid and efficient communication between individuals.



(a) Leading eigenvector (LE) communities



(b) Walktrap (WT) communities

**Figure 4.5: Dynamics of bees** Each column represents a time step, the colored rectangles represent the communities for each step, and the height of the rectangles corresponds to the amount of its community members, as referenced by the number. *Green* indicates the community containing young bees and the queen, *gray* represents the community containing middle-aged bees (only for WT), and *orange* the community containing old bees. This figure shows that the major part of the bees either stays in the same aged community or switches to an older group. The *light gray* boxes represent the number of bees that are added to the colony and bees that disappear.

### 4.3. Discussion of Results

Charbonneau, Blonder, and Dornhaus [6] state that it is assumed that many biological networks, including insect colonies, approximate scale-free networks. For some of them, the scale-free property has been shown, but for social insect networks this question remains open. Investigated social insect colonies are often small and therefore the methods for the recognition of scale-free phenomena are limited. They do not specify the type of social insect networks, whether the inference of interactions is based on spatial proximity, physical contacts, or food transfer events.

The network I explored is large compared to past studies (Section 2.2). The degree distribution of the investigated spatial proximity network of honey bees does not follow a power-law; the absence of hubs and a non-hierarchical structure characterizes this network. This result corresponds to the decentralized structure of a honey bee colony, and the absence of a central authority described by Seeley [47].

I noticed bimodal degree, strength, closeness and betweenness distributions and a right skewed lcc distribution, corresponding mainly to bees older than 45 days. While inspecting this group of bees, I found that this group has a very low detection rate and is not part of any other following snapshot. Probably this group of bees dies during that day. Bees who are present in the hive earlier that day and are then absent for the rest of the day have very low network measure values. The total number of old bees is relatively small compared to other age groups. Consequently, low network measure values strongly affect the mean of that old group and should be excluded in future studies.

Generally, I observed a correlation between the detection frequency of a bee, its age, and its corresponding network measure value. Older bees are detected less often than younger bees and therefore differ in their network measures. The age-based task division of bees in a colony observed by Seeley [48] might be an explanation; namely, old bees are foragers, the middle-aged bees conduct several tasks inside the hive but mainly they store resources, and young bees are primarily nursing. Baracchi and Cini [28] also assumed that the time which bees spend outside the hive probably affects their connectedness within the interaction network.

#### 4.3.2 Characterization of Functional Groups and its Dynamics

According to the definition of communities in Section 2.1.2, I found two to three communities, depending on the algorithm applied.

The algorithms (LE and WT) detected communities, despite a high network density and without thresholding links of low values, as opposed to Mersch, Crespi, and Keller [4]. The authors reduced the network's density artificially to 25% to apply the infomap algorithm.

I also examined the spatial fidelity of the revealed communities and their age composition, similar to Baracchi and Cini [28]. I found that younger bees are located close to the brood (upper center of the comb); older bees are situated closer to the hive exit, and middle-aged bees are placed between the two groups and around the brood, where the cells for honey storage are located.

### 4.3. Discussion of Results

I inspected three snapshots over a period of five days and found that the detected communities are stable over time. Age-division and spatial fidelity can be observed in all the snapshots. Bees from younger communities move to older communities as they age. Only a few bees changed from older to younger communities.

It is surprising that my results align with Baracchi and Cini [28], because they did not use a community detection algorithm. The authors conducted a hierarchical clustering based on the network measures strength, eigenvector and betweenness centrality of individual bees. Moreover, their colony contained bees of three predetermined age cohorts, instead of representing all age groups ranging from 0 to 60 days, as in my study.

The communities I detected are similar to the groups of bees formed by temporal polyethism. The old bees positioned closer to the hive exit may be the foragers, the middle-aged group spatially close to the storage cells may be the food storage bees and the group of young bees may be the cell cleaning and brood care bees because they are located close to the brood. My findings are very close to the ones of Baracchi and Cini [28].

The two approaches discovered the same functional groups of the bee colony, on the one hand by node level network measures (hierarchical clustering) and on the other hand by a higher than expected density of nodes (community detection). That acknowledges the existence of the age-based division of labor in honey bee colonies as well as the higher communication frequency within groups than between groups. Nevertheless, the low modularity score indicates that the segregation of groups is not that obvious and strict; therefore a lot of interaction between groups exists.

Mersch, Crespi, and Keller [4] revealed that the behavioral maturation of ants is a slow and noisy process. Instead of investigating the transition of individuals day wise, they grouped 41 days in four periods. For each period they assigned each ant to a community if it was found in this community 70% of the time. It seems that honey bee transitions are in contrast to ants faster and smoother.

# **Chapter 5**

## **Conclusions**

The purpose of this thesis was to investigate worker-worker interaction networks of a honey bee colony. To achieve this, I implemented a pipeline for the extraction of time-aggregated networks using high-resolution honey bee tracking data. The topology, community structures and development of community members were examined from the resulting weighted, undirected, spatial proximity networks of three consecutive time steps.

As opposed to most real-world networks, the examined honey bee interaction networks are not scale-free. They are characterized by a non-hierarchical topology and a decentralized structure. The small-world characteristic of honey bee networks allows for efficient communication within the bee colony. The observed communities within the honey bee colony, are formed of age-based functional groups with a spatial fidelity towards different areas of the comb. There are different types of cells (e.g. brood, honey, and pollen) and honeybees occupying different areas have distinct tasks.

The global network structure of the honey bee colony is stable over time, but its local structure is highly dynamic. Individual bees change communities as they age. My results are in agreement with the established state of the research: the absence of a central authority and the decentralized organization of honey bee colonies shaped by temporal polyethism.

My network analysis results verify the network pipeline, my definition of spatial proximity networks and the initially chosen parameters. The network pipeline provides an excellent foundation for further investigations.

### **5.1 Limitations**

The following section outlines limitations concerning the accuracy and quality of the resulting networks. Especially, consequences of the networks' high density are pointed out.

Despite a complex preprocessing procedure, several shortcomings concerning the quality of the extracted networks still exist. I filtered out erroneous detections before network generation; however a few individuals remain in the extracted networks that should not exist according to the tagging and hatching documentation. Besides this,

bees which are dying at some point during the aggregation period, are part of the network. This fact needs to be considered depending on the research context.

The prefiltering of detections, as well as the synchronization of four cameras, reduces the amount of data that remains for the extraction of interactions. The gap size parameter was an attempt to compensate for this shortcoming of the data but does not perfectly solve the problem. I believe that some observed interactions are shorter than they are in reality, which distorts the networks.

Spatial proximity is an indicator for interaction but it does not capture actual interactions. The definition of spatial proximity by a maximal distance and a minimum contact duration is very loose, especially on a honeycomb with limited space. It leads to many links and a high density of the network resulting in high noise. This noise creates a blurred image of real interactions between bees. My choice of aggregating the networks for ten hours fosters this noisiness, resulting in a global state of the colony, rather than capturing finer granular dynamics.

In this context, the network property strength is the only measure, that profits from the aggregation. All other measures are less meaningful. Due to the high density and size of the network, the methods I can apply for community detection are limited. The selection of an algorithm for detecting communities is restricted to algorithms finding only non-overlapping structures.

## 5.2 Recommendations

This sections list recommendations for improving the applied methods. I am focusing on concepts to reduce the noise within the network.

**More Dynamic and Temporal Analysis** Lowering the window size of the aggregated network and investigating different granularities could allow more dynamic analysis of the networks. Instead of using time-aggregated networks, one could shift towards the use of time-ordered networks by using time-stamped interactions.

**Focusing on Important Interactions** The space on the honeycomb is limited and crowded. For reducing the number of link to only meaningful interactions, I see three main approaches. For the time being, it is an option to fine tune the pipeline parameters by lowering the size of the maximum distance and by raising the number of frames for minimum contact duration. Instead of keeping the definition of spatial proximity, I would recommend extracting contact events (e.g. by including an angle, so bees facing each other) or trophallaxis events for defining the links, especially when using those networks to investigating more specific biological research questions.

**Using the Potential of Weighted Links** A simple global threshold for excluding links below a certain value could be used. Instead of applying a global threshold

to reduce the density, a network reduction algorithm could be implemented to extract the backbone structure of the network. Serrano, Boguná, and Vespignani [49] propose a disparity filter algorithm, which seems promising but needs further investigation. The disparity measure characterizes the level of local heterogeneity of links [50]. For all network measures utilized in this work, weighted versions exist. The weighted measures (e.g., closeness and betweenness) implemented in igraph and networkX favor link weights over the number of links and simply apply Dijkstra [51] for calculating the shortest paths. Opsahl, Agneessens, and Skvoretz [52] propose weighted network measures by providing a generalized degree and shortest path algorithm. The tuning parameter, Opsahl et al. introduce, has to be chosen. This parameter defines whether to emphasize the number of links or the weights of links and must be selected according to a predefined research question.

**Normalizing by the Detection Frequency of Individuals** Depending on the topics of further research a normalization of the networks regarding the detection rate of individuals could be useful. I propose two options: either normalize the link weight by applying the simple ratio index<sup>1</sup> (SRI) [23] or normalize the particular node level measure by taking the detection frequency of that focal individual into account.

**Random Geometric Graph** Instead of comparing the honeybee network to an Erdős-Rényi graph a new model could be implemented. As the starting point, a random geometric graph [53] can be used. In each frame, the nodes could be placed not completely randomly, preferably by modeling the behavior of a bee as a random walker. The direction of movement could be chosen randomly, but the distance of a step might be selected according to the average speed of bees.

## 5.3 Outlook

To fine tune the pipeline parameters for the network, one should systematically investigate the parameters effects on network properties. I started to analyze this, but only for a few combinations of values and for window sizes up to one hour. Similarly, the robustness of the detected communities regarding the pipeline parameters is worthy of further study. In addition, the provided dataset facilitates the investigation of seasonal change in honey bee colonies using network analysis methods. Long-term dynamics offer a high potential for further studies. It would be interesting to compare my network analysis results of domesticated honey bees to the social networks of wild honey bees, to discover differences regarding individual behavior and global colony organization.

---

<sup>1</sup>Dividing the link weight by the proportion of times two individuals were seen together out of the total number of times those individuals were observed.

## 5.4 Closing Remarks

My work was an important first step to gain trust in the honey bee tracking data generated by the BeesBook system. It identifies limitations, pinpoints scope for improving the system and lays the foundations for further network analysis.

Studying non-human animal data in an explorative way that fosters the framing of novel biological hypotheses demands a profound domain knowledge or the constant support of experts of the studied species. The process of manual data collection by observing the animals face to face creates valuable information which is beneficial for data analysis and understanding the context of research. An automatic observation process veils this part and therefore increases the abstraction level and encourages alienation between the researchers and observed animals.

The automatic tracking of a vast number of animals over an extended observation period with a high sampling resolution leads to an enormous amount of data. Applying network analysis methods to novel datasets, which were not collected for a specific study, opens the space to investigate the data in an explorative way and to discover the unexpected. However, it also carries the risk of either simply describing network structures of various species or leading to the restating of well-known facts.

Framing biological research questions that benefit from network science methods or the development of new techniques in the field of network analysis with the help of this unique dataset should be the overall goal.

# Bibliography

- [1] F. Wario *et al.*, “Automatic methods for long-term tracking and the detection and decoding of communication dances in honeybees”, *Frontiers in Ecology and Evolution*, vol. 3, p. 103, 2015.
- [2] J. D. Crall *et al.*, “BEEtag: A low-cost, image-based tracking system for the study of animal behavior and locomotion”, *PLOS ONE*, vol. 10, no. 9, 2015.
- [3] M. Fiala, “Comparing artag and artoolkit plus fiducial marker systems”, in *2005 IEEE International Workshop on Haptic Audio Visual Environments and their Applications*, IEEE, 2005, pp. 6–13.
- [4] D. P. Mersch *et al.*, “Tracking individuals shows spatial fidelity is a key regulator of ant social organization”, *Science*, vol. 340, no. 6136, pp. 1090–1093, 2013.
- [5] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge University Press, 1994, vol. 8.
- [6] D. Charbonneau *et al.*, “Social insects: A model system for network dynamics”, in *Temporal Networks*, Springer, 2013, pp. 217–244.
- [7] D. R. Farine and H. Whitehead, “Constructing, conducting and interpreting animal social network analysis”, *Journal of Animal Ecology*, vol. 84, no. 5, pp. 1144–1163, 2015.
- [8] A.-L. Barabási, *Network science*. Cambridge University Press, 2016.
- [9] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, 2010.
- [10] A. Barrat *et al.*, “The architecture of complex weighted networks”, *Proceedings of the National Academy of Sciences*, vol. 101, no. 11, pp. 3747–3752, 2004.
- [11] G. Palla *et al.*, “Uncovering the overlapping community structure of complex networks in nature and society”, *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.
- [12] D. P. Croft *et al.*, *Exploring animal social networks*. Princeton University Press, 2008.
- [13] S. Fortunato, “Community detection in graphs”, *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.
- [14] M. E. Newman, “Finding community structure in networks using the eigenvectors of matrices”, *Physical Review E*, vol. 74, no. 3, p. 036104, 2006.
- [15] P. Pons and M. Latapy, “Computing communities in large networks using random walks”, in *International Symposium on Computer and Information Sciences*, Springer, 2005, pp. 284–293.

- [16] J. Moody *et al.*, “Dynamic network visualization 1”, *American Journal of Sociology*, vol. 110, no. 4, pp. 1206–1241, 2005.
- [17] N. Pinter-Wollman *et al.*, “The dynamics of animal social networks: Analytical, conceptual, and theoretical advances”, *Behavioral Ecology*, vol. 25, no. 2, p. 242, 2014.
- [18] B. Blonder *et al.*, “Temporal dynamics and network analysis”, *Methods in Ecology and Evolution*, vol. 3, no. 6, pp. 958–972, 2012.
- [19] J. Krause *et al.*, *Animal social networks*. Oxford University Press, 2014.
- [20] E. Greenwald *et al.*, “Ant trophallactic networks: Simultaneous measurement of interaction patterns and food dissemination”, *Scientific Reports*, vol. 5, 2015.
- [21] N. Pinter-Wollman *et al.*, “The effect of individual variation on the structure and function of interaction networks in harvester ants”, *Journal of the Royal Society Interface*, vol. 8, no. 64, pp. 1562–1573, 2011.
- [22] L. E. Quevillon *et al.*, “Social, spatial, and temporal organization in a complex insect society”, *Scientific Reports*, vol. 5, 2015.
- [23] V. A. Formica *et al.*, “Fitness consequences of social network position in a wild population of forked fungus beetles (*bolitotherus cornutus*)”, *Journal of Evolutionary Biology*, vol. 25, no. 1, pp. 130–137, 2012.
- [24] J. S. Waters and J. H. Fewell, “Information processing in social insect networks”, *PLOS ONE*, vol. 7, no. 7, e40337, 2012.
- [25] A. B. Sendova-Franks *et al.*, “Emergency networking: Famine relief in ant colonies”, *Animal Behaviour*, vol. 79, no. 2, pp. 473–485, 2010.
- [26] D. Naug, “Structure and resilience of the social network in an insect colony as a function of colony size”, *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1023–1028, 2009.
- [27] M. C. Otterstatter and J. D. Thomson, “Contact networks and transmission of an intestinal pathogen in bumble bee (*bombus impatiens*) colonies”, *Oecologia*, vol. 154, no. 2, pp. 411–421, 2007.
- [28] D. Baracchi and A. Cini, “A socio-spatial combined approach confirms a highly compartmentalised structure in honeybees”, *Ethology*, vol. 120, no. 12, pp. 1167–1176, 2014.
- [29] D. Naug, “Structure of the social network and its influence on transmission dynamics in a honeybee colony”, *Behavioral Ecology and Sociobiology*, vol. 62, no. 11, pp. 1719–1725, 2008.
- [30] J. Scholl and D. Naug, “Olfactory discrimination of age-specific hydrocarbons generates behavioral segregation in a honeybee colony”, *Behavioral Ecology and Sociobiology*, vol. 65, no. 10, p. 1967, 2011.
- [31] D. Naug and B. Smith, “Experimentally induced change in infectious period affects transmission dynamics in a social group”, *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 274, no. 1606, pp. 61–65, 2007.

- [32] B. Blonder and A. Dornhaus, “Time-ordered networks reveal limitations to information flow in ant colonies”, *PLOS ONE*, vol. 6, no. 5, e20298, 2011.
- [33] R. Jeanson, “Long-term dynamics in proximity networks in ants”, *Animal Behaviour*, vol. 83, no. 4, pp. 915–923, 2012.
- [34] T. D. Seeley, “Adaptive significance of the age polyethism schedule in honeybee colonies”, *Behavioral Ecology and Sociobiology*, vol. 11, no. 4, pp. 287–293, 1982.
- [35] B. R. Johnson, “Within-nest temporal polyethism in the honey bee”, *Behavioral Ecology and Sociobiology*, vol. 62, no. 5, pp. 777–784, 2008.
- [36] M. Lindauer, “Ein Beitrag zur Frage der Arbeitsteilung im Bienenstaat”, *Journal of Comparative Physiology A*, vol. 34, no. 4, pp. 299–345, 1952.
- [37] R. Menzel and M. Eckoldt, *Die Intelligenz der Bienen: Wie sie denken, planen, fühlen und was wir daraus lernen können*. Knaus Albrecht, 2016.
- [38] L. Sixt, “Rendergan: Generating realistic labeled data - with an application on decoding bee tags”, Aug. 2016, unpublished thesis.
- [39] M. Rosvall *et al.*, “The map equation”, *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, 2009.
- [40] M. Rosvall and C. T. Bergstrom, “An information-theoretic framework for resolving community structure in complex networks”, *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7327–7331, 2007.
- [41] Z. Yang *et al.*, “A comparative analysis of community detection algorithms on artificial networks”, *Scientific Reports*, vol. 6, 2016.
- [42] S. Harenberg *et al.*, “Community detection in large-scale networks: A survey and empirical evaluation”, *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 6, pp. 426–439, 2014.
- [43] T. Aynaud *et al.*, “Communities in evolving networks: Definitions, detection, and analysis techniques”, in *Dynamics On and Of Complex Networks, Volume 2*, Springer, 2013, pp. 159–200.
- [44] P. Bródka *et al.*, “Community evolution”, *Encyclopedia of Social Network Analysis and Mining*, pp. 220–232, 2014.
- [45] J. Hopcroft *et al.*, “Tracking evolving communities in large linked networks”, *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5249–5253, 2004.
- [46] Q. K. Telesford *et al.*, “The ubiquity of small-world networks”, *Brain connectivity*, vol. 1, no. 5, pp. 367–375, 2011.
- [47] T. D. Seeley, “The honey bee colony as a superorganism”, *American Scientist*, vol. 77, no. 6, pp. 546–553, 1989.
- [48] ——, “Social foraging in honey bees: How nectar foragers assess their colony’s nutritional status”, *Behavioral Ecology and Sociobiology*, vol. 24, no. 3, pp. 181–199, 1989.

- [49] M. Á. Serrano *et al.*, “Extracting the multiscale backbone of complex weighted networks”, *Proceedings of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.
- [50] M. Barthelemy *et al.*, “Spatial structure of the internet traffic”, *Physica A: Statistical Mechanics and its Applications*, vol. 319, pp. 633–642, 2003.
- [51] E. W. Dijkstra, “A note on two problems in connexion with graphs”, *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.
- [52] T. Opsahl *et al.*, “Node centrality in weighted networks: Generalizing degree and shortest paths”, *Social networks*, vol. 32, no. 3, pp. 245–251, 2010.
- [53] J. Dall and M. Christensen, “Random geometric graphs”, *Physical Review E*, vol. 66, p. 016121, 1 Jul. 2002.

# List of Figures

1.1	Tagged bees inside the observation hive.	1
3.1	Observation setup	13
3.2	Quality of detections and IDs	15
3.3	Detection frequency of IDs	16
3.4	Structure of dataset	16
3.5	Maximum distance of bees	20
3.6	Frequency and total duration of interaction	22
4.1	Transitions of bees per snapshot	29
4.2	Age distribution, correlation with detection frequency and link weight distribution of snapshot 3	30
4.3	Local measures of snapshot 3	33
4.4	Age and spatial distribution of communities	34
4.5	Dynamics of bees	38
B.1	Tagging frequency	55
B.2	Recording season with maintainance and failures	56
C.1	Link weight distribution for all snapshots	57
C.2	Link weight fit	57
C.3	Communities per network for LE	58
C.4	Communities per network for WT	59
C.5	Age distribution for LE	60
C.6	Age distribution per snapshot for WT	61
C.7	Snapshot 1: Local measures in relation to age and detection frequency	62
C.8	Snapshot 2: Local measures in relation to age and detection frequency	63
C.9	Snapshot 3: Scatter plot for node level measure	64
C.10	Age distribution per snapshot	65

# List of Tables

2.1	Definitions . . . . .	5
3.1	Terms related to the dataset . . . . .	14
3.2	Parameters chosen for network analysis . . . . .	20
3.3	Measures used for analysis . . . . .	23
3.4	Comparing community detection algorithms . . . . .	24
3.5	Number of community members per algorithm and snapshot . . . . .	25
4.1	Sampling period . . . . .	28
4.2	Global network properties . . . . .	31
4.3	Communities per algorithm . . . . .	35
4.4	Kolmogorov-Smirnov test . . . . .	35
4.5	Overview about communities . . . . .	36
4.6	Kolmogorov-Smirnov test . . . . .	37
A.1	Type of network . . . . .	52
A.2	Properties of studies . . . . .	53
A.3	Network measures . . . . .	54

# **Appendix A**

## **Literatur Review**

**Table A.1: Type of networks**

Temporal Analysis			Static Analysis																																
B. Blonder et al. [18]	R. Jeanson [33]	D. P. Mersch et al. [4]	D. Naug and B. Smith [31]			M. C. Otterstatter and J. D. Thomson [27]			D. Naug [29]			D. Naug [26]			A. B. Sendova-Franks et al. [25]			N. Pinter-Wollman et al. [21]			J. Scholl and D. Naug, [30]			J. S. Waters and J. H. Fewell [24]			D. Baracchi and A. Cini [28]			E. Greenwald et al. [20]			L. E. Quevillon et al. [22]		
Type of network	ta/to	ta	ta	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s								
<b>Type of interaction</b>																																			
Spatial proximity (body(B) length)	4/3xBL			2/3xBL >0.2s												1xBL																			
Physical contact	A-B	A		B-B		ex (3)										A	A																		
Food exchange (throphallaxis)				> 5s		> 5s			x			x				x			x			x			> 1s										
Directed network	x - (1)	-	-	x	-	x	-	x	-	x	-	x	-	-	-	-	-	-	x	-	x	-	x	-											
<b>Weighted network</b>																																			
Duration of interaction	-	x	-	-	x	x	-	-	-	-	-	-	-	-	-	-	-	-	v (2)	-															
Number of interactions	-	-	x	-	x	-	x	-	x	-	x	-	x	-	x	-	-	x	-																

(1) both

(2) volume corresponds to duration

(3) except dominance interactions

A = antenna

B = body

BL = body-length

ta = time-aggregated

to = time-ordered

s = static

**Table A.2: Properties of studies**

	Temporal Analysis			Static Analysis											
	B. Blonder et al. [18]	R. Jeanson [33]	D. P. Mersch et al. [4]	D. Naug and B. Smith [31]	M. C. Otterstatter and J. D. Thomson [27]	D. Naug [29]	D. Naug [26]	A. B. Sendova-Franks et al. [25]	N. Pinter-Wollman et al. [21]	J. Scholl and D. Naug, [30]	J. S. Waters and J. H. Fewell [24]	D. Baracchi and A. Cini [28]	E. Greenwald et al. [20]	L. E. Quevillon et al. [22]	
<b>Tracking</b>															
Automatic	x	x			x			x			x			x	
Manual	x			x	x	x	x	x	x	x	x	x	x	x	
<b>Species</b>															
	A	A	A	HB	BB	HB	W	A	A	HB	A	HB	A	A	
<b>(1) Time</b>															
Total duration of study	3w	3w	41d	1d	40d		24d (6)	1d	1d	3w	1d	1d	1d	8d (5)	
Observation period	2x 30m	3x7x 24h	41x 24h	1h 12h (4)	40x 1h	24x 45x5m	30m	5m	3x 1h	2h	10h	30m	8x 30m		
Sampling resolution***	v/e	1 f/s	2 f/s	v/e	30 f/s	v/e	v/e	1 f/m	30 f/s	v/e	15 f/s	1 f/m	v/e	v/e	
<b>(2) Space*</b>															
1-frame hive											x (3)				
2-frame hive				x (2)		x(4)				x					
<b>(3) Size</b>															
Number of colonies	4	4	6	1	7	1	9	4	2	1	2	1	2(1)	2	
Colonie size**	6-90	55-58	122-192	4000	5-7	1000	8-40	42-95	131, 72	1500	89	4000	50-100	75	
Marked individuals	x	x	x		x		x	x	x	x	x	211	x	x	
Marked cohorts				6		4			3						
Age		x		x	x	x			x		x				
<b>Analysis Tools in R</b>															
igraph	x									x			x		x
t-net		x													
timeordered	x														
<b>Other Tools: netdraw, cytoscape, UCINET, FANMOD</b>															

(1) two species

(2) only video for one side

"(entrance designed so foragers should unload here)"

(3) only one side observed

(4) 6 day and 6 night"

(5) night

(6) each sampling day consisted of three sessions of 2 h each between 0630 and 1830 hours. in each session 15x5-min all-occurrence samplings were carried out resulting.

A = Ant

BB = Bumble Bee

HB = Honey bee

W = Wasp

\* only for honey bees

\*\* mean or range if &gt; 2

\*\*\* indicates all marked/tracked, if number, just subset

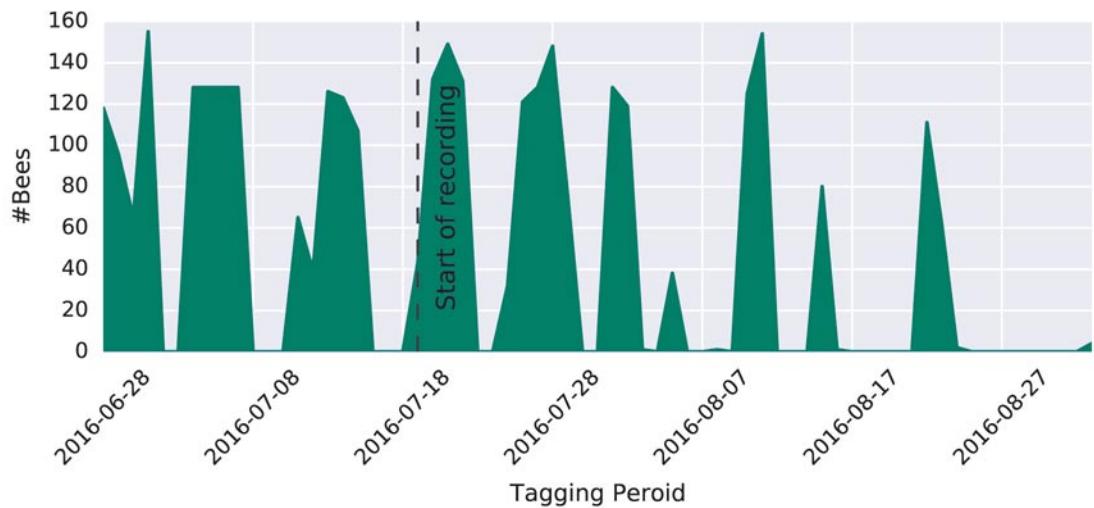
\*\*\*\* v=video, e=event, if no resolution given or manual video analysis was used

**Table A.3:** Network measures

	Temporal Analysis	Static Analysis			
B. Blonder et al. [18]					
R. Jeanson [33]					
D. P. Mersch et al. [4]					
D. Naug and B. Smith [31]					
M. C. Otterstatter and J. D. Thomson [27]					
D. Naug [29]					
D. Naug [26]					
A. B. Sendova-Franks et al. [25]					
N. Pinter-Wollman et al. [21]					
J. Scholl and D. Naug, [30]					
J. S. Waters and J. H. Fewell [24]					
D. Baracchi and A. Cini [28]					
E. Greenwald et al. [20]					
L. E. Quevillon et al. [22]					
<b>Global level measures</b>					
Average degree	x				
Maximal degree					x
Average strength		x	x		
Average shortest path length				x	x
Density	x	x			x
Diameter					x
<b>Node level measures</b>					
Degree	x		x	x	x
Strength	x	x	x	x	x
Betweenness centrality	x	x			x
Closeness centrality	x				x
Eigenvector centrality					x
Clustering coefficient		x	x		
<b>Other method</b>					
Burst constraint					x
Disparity	x				
Cluster or Community detection		x			x
Fitting of distributions	x			x	x
Compare to random			x	x	
Information flow	x		x		x
Interaction between age groups				x	
Ego network					x
Robustness		x			

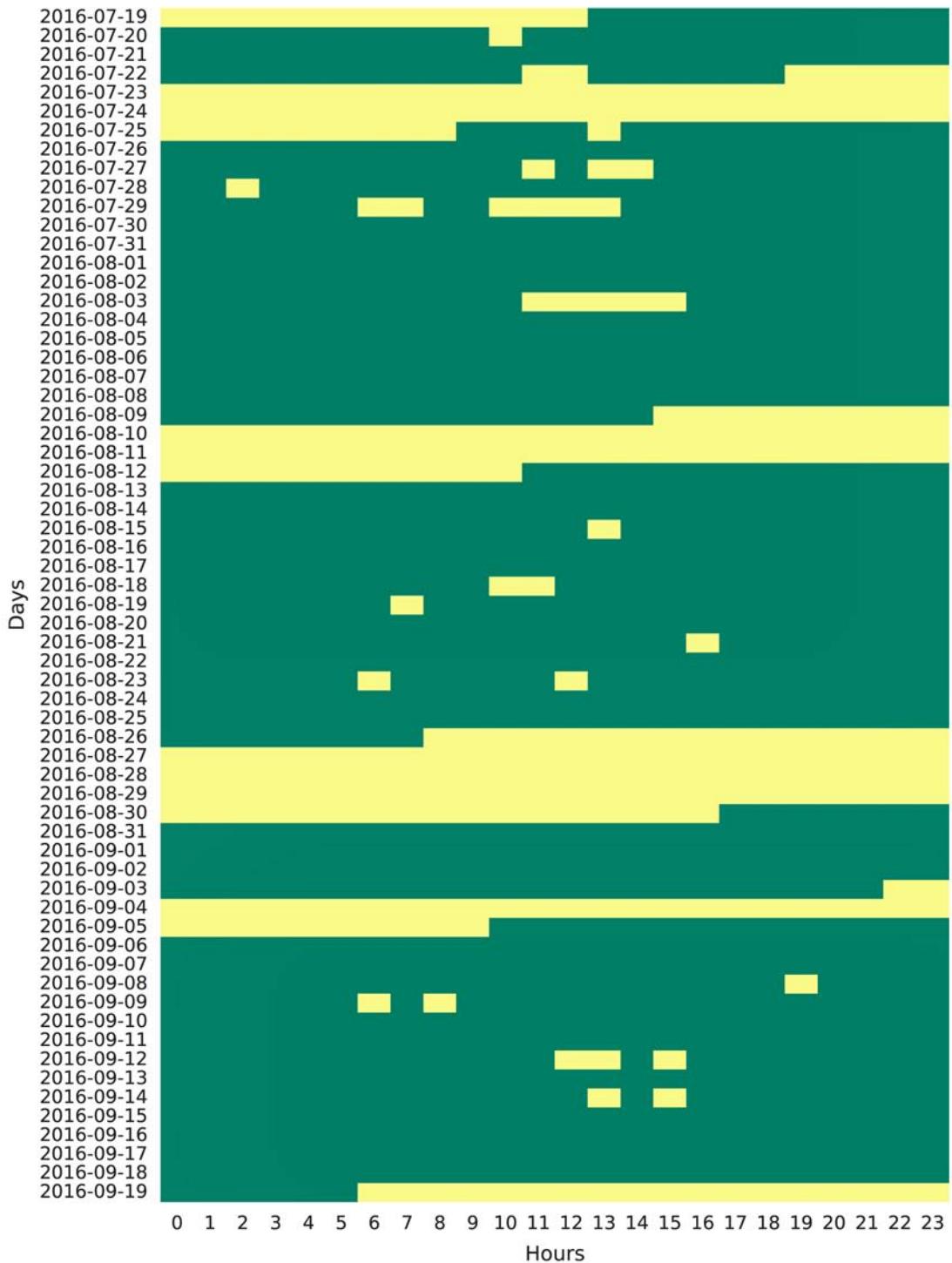
## Appendix B

### Additional Information about the Dataset



**Figure B.1: Tagging frequency** The bees were primarily tagged during the week. On average 48 bees were tagged each day, considering only tagging days, the average is about 91.

## B. Additional Information about the Dataset



**Figure B.2: Recording season with maintainance and failures** *Green* indicates recording went without any big interruption; *Yellow* indicates maintainance work or technical failures of one or all cameras. This is calculated using the expected number of files produced by each camera per hour.

# Appendix C

## Network Analysis

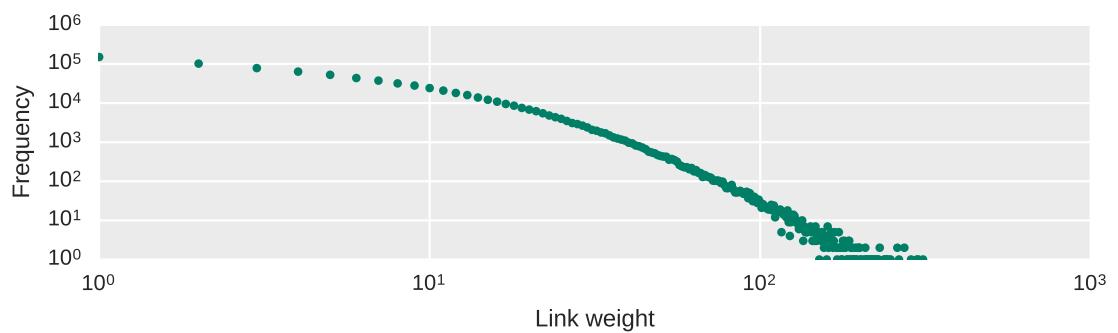


Figure C.1: Link weight distribution for all snapshots

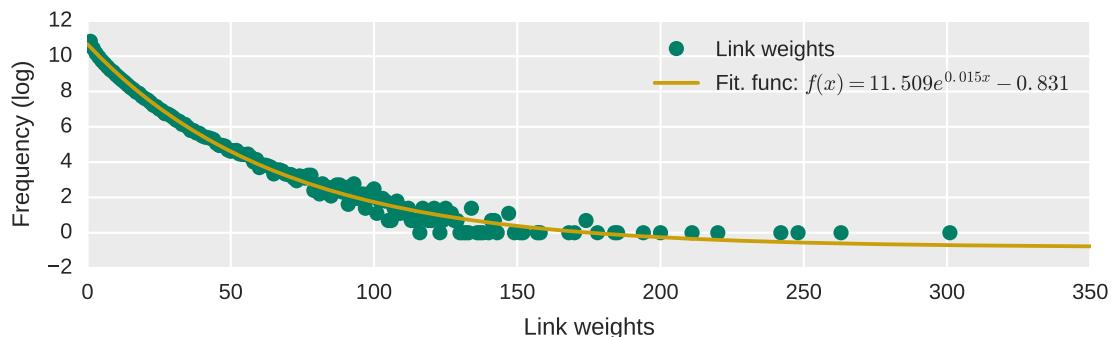
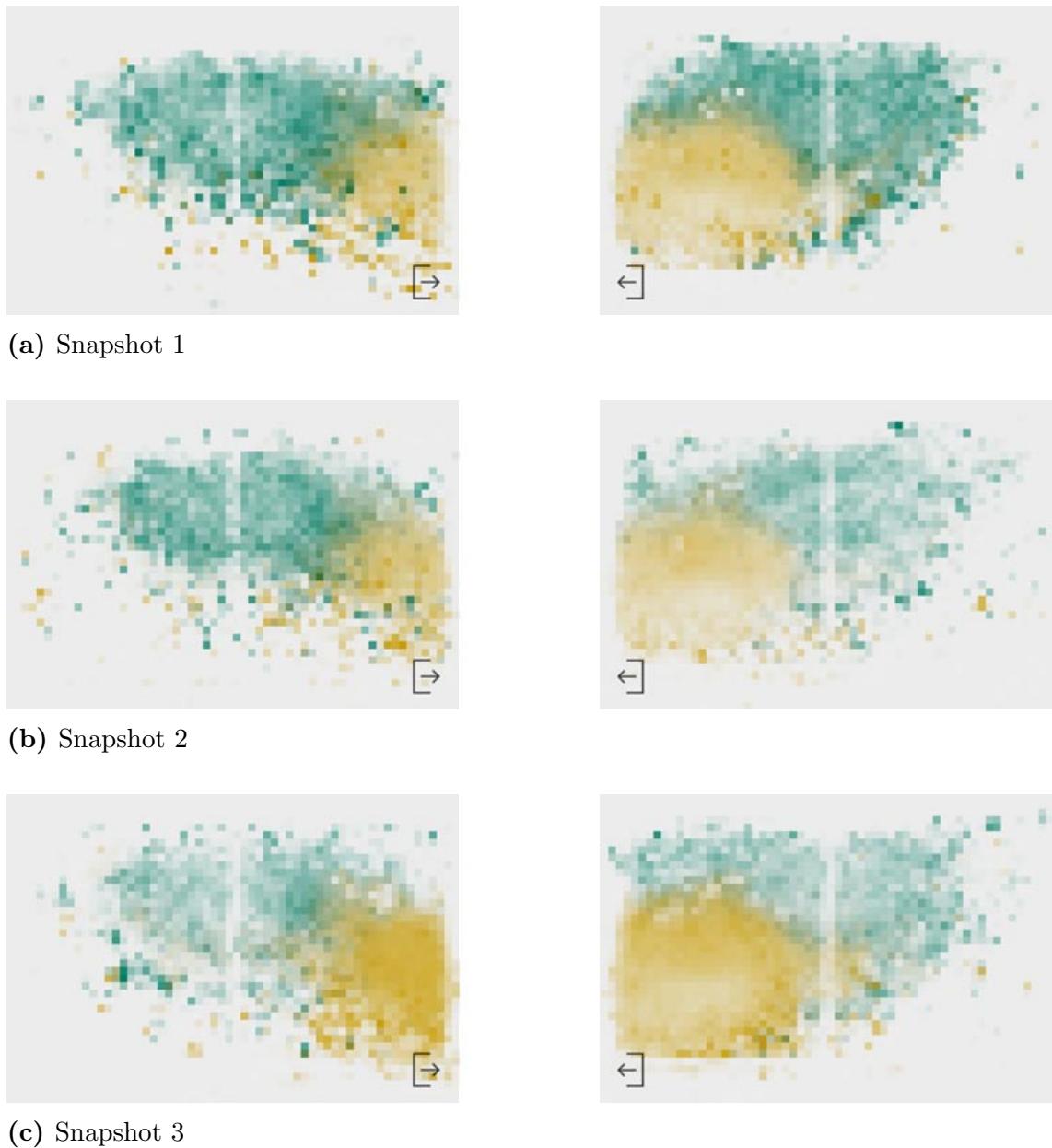
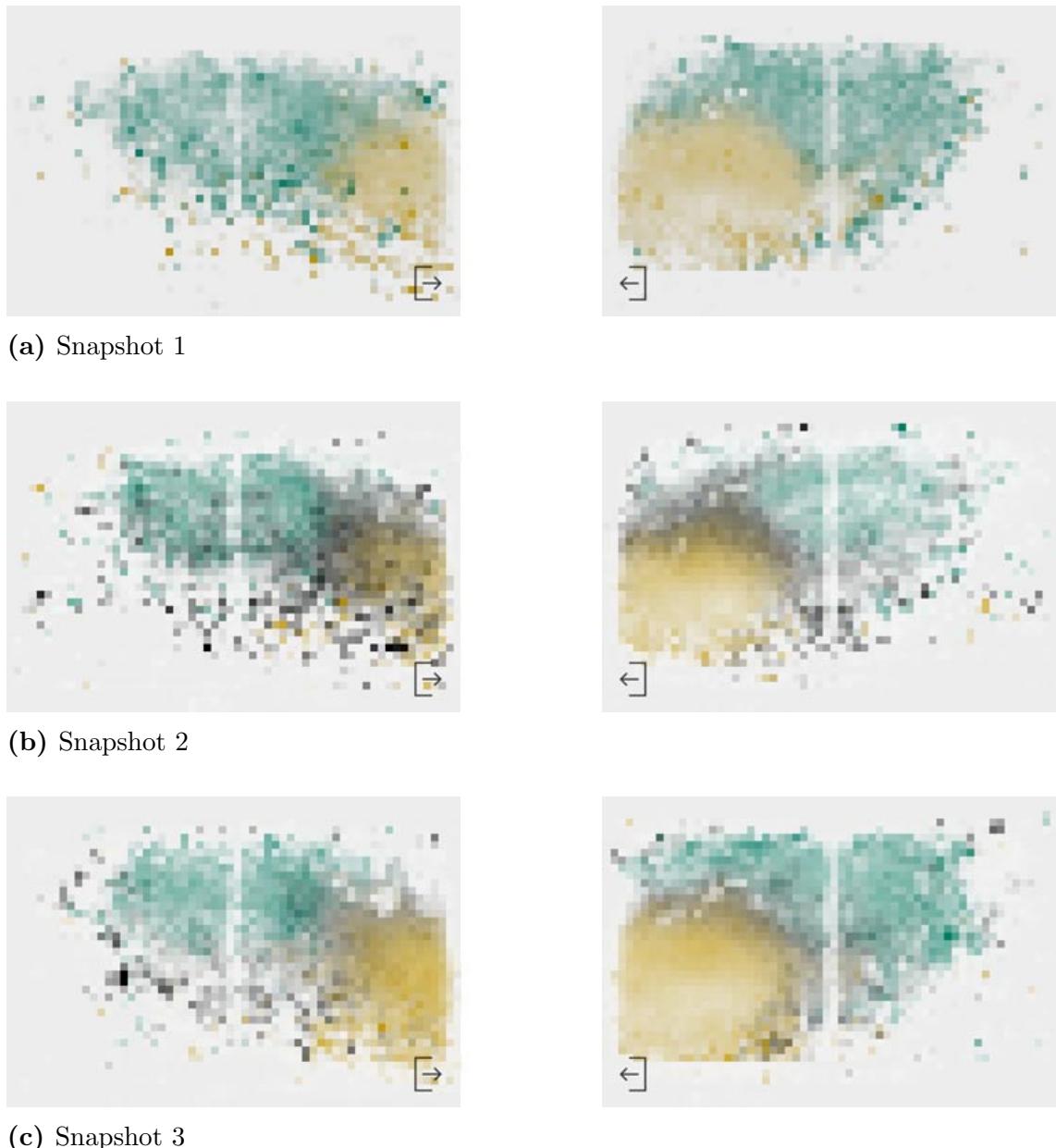


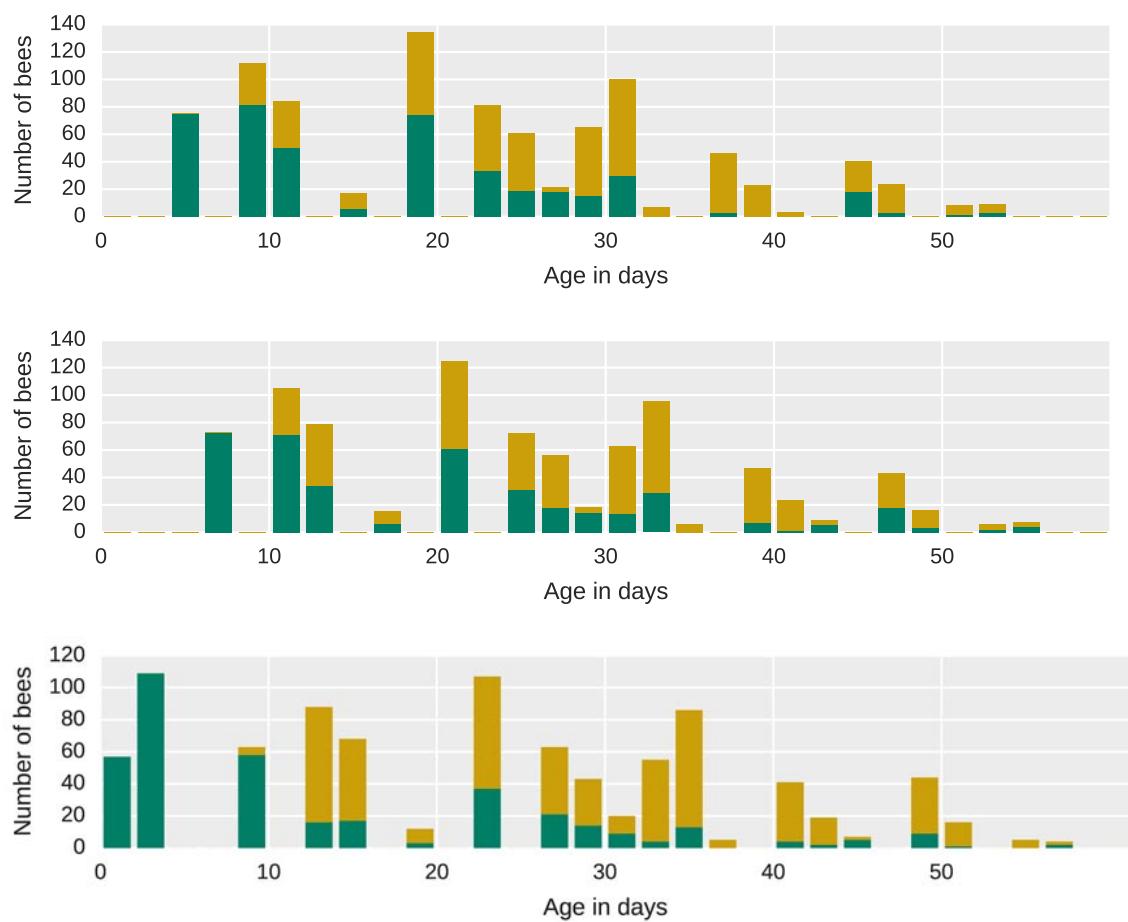
Figure C.2: Link weight fit



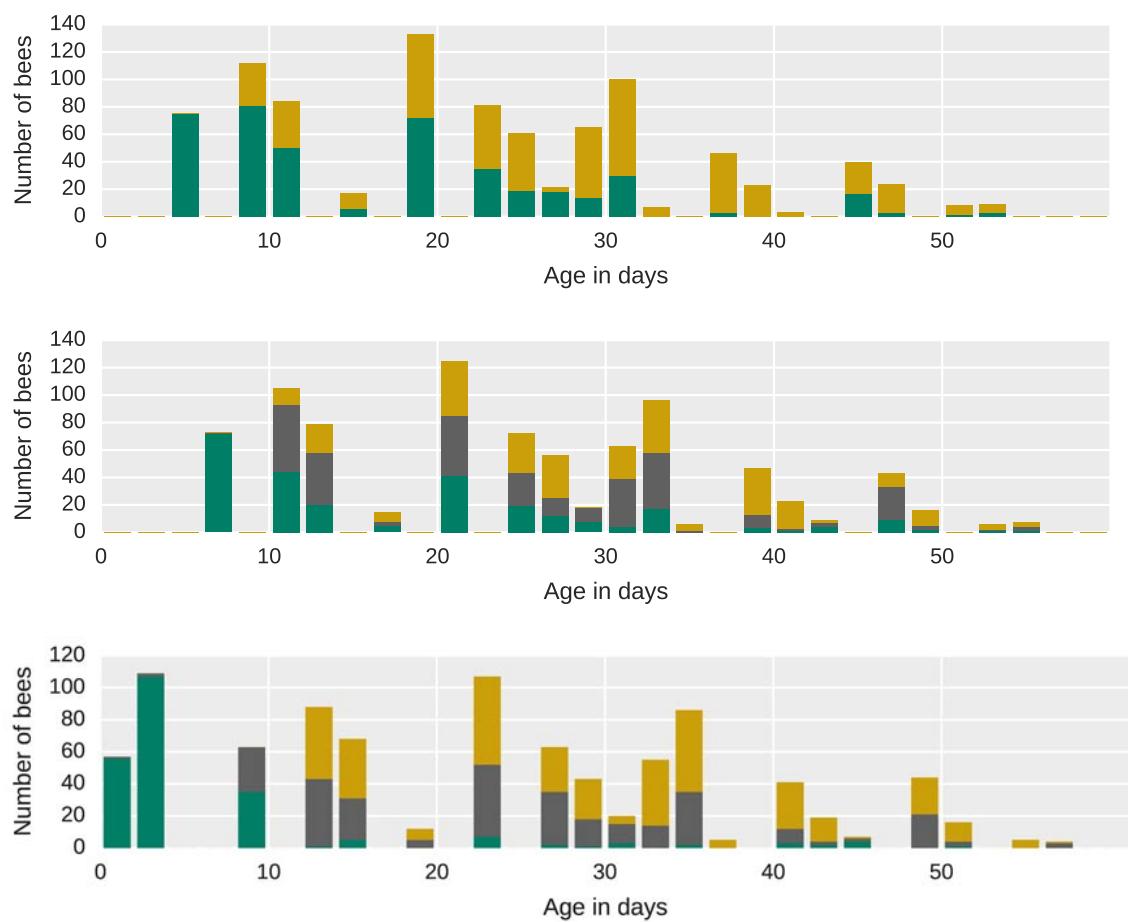
**Figure C.3: Communities per network for LE** The *green* colour represents the younger community, containing the queen. The *orange* color represents the older community. The hive exit on side A is on the bottom right and on side B on the bottom left. The data is aggregated for the complete timeframe of ten hours.



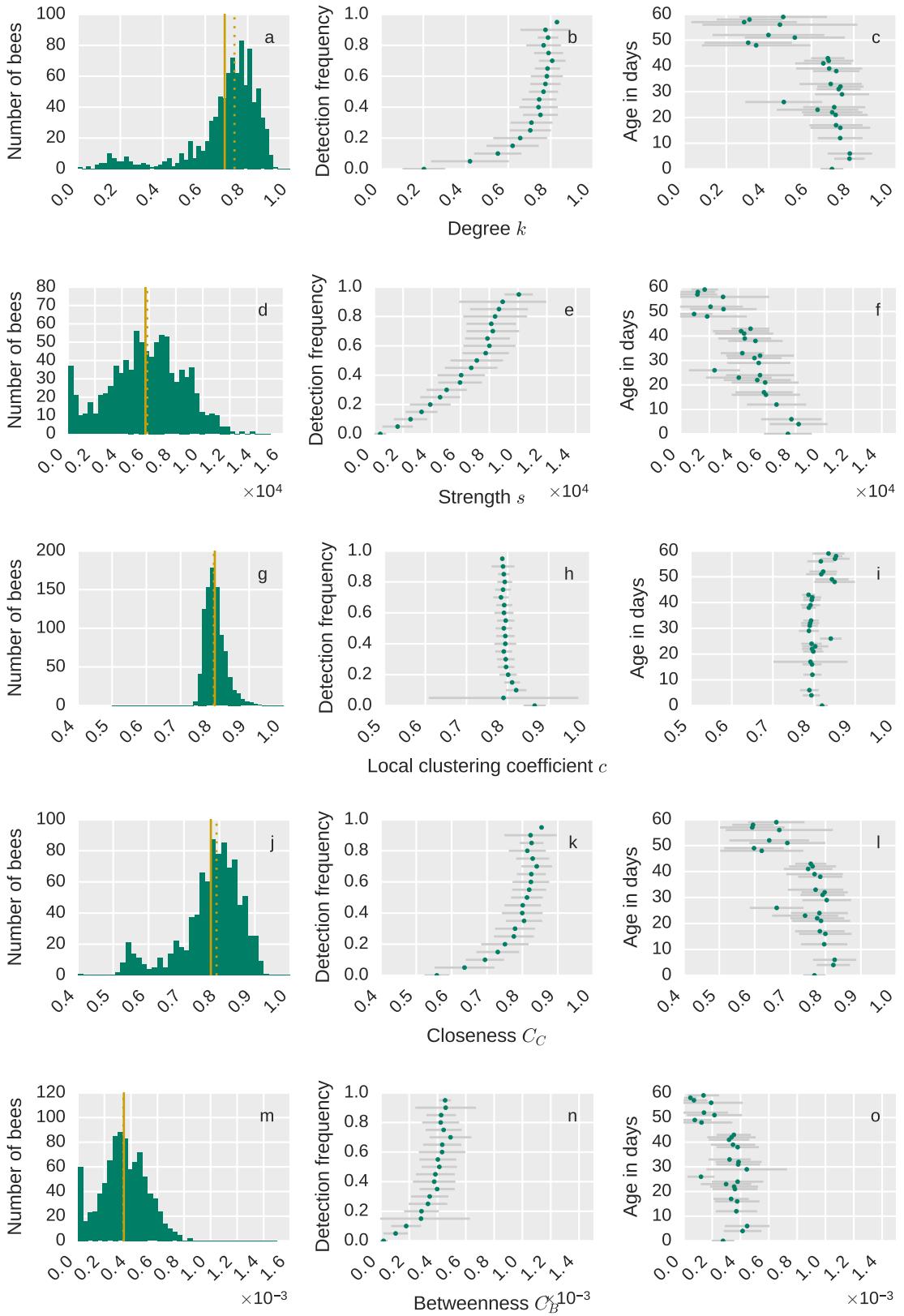
**Figure C.4: Communities per network for WT** The *green* colour represents the younger community, containing the queen. The *orange* color represents the older community. The *gray* represents the middle-age community. The hive exit on side A is on the bottom right and on side B on the bottom left. The data is aggregated for the complete timeframe of ten hours.



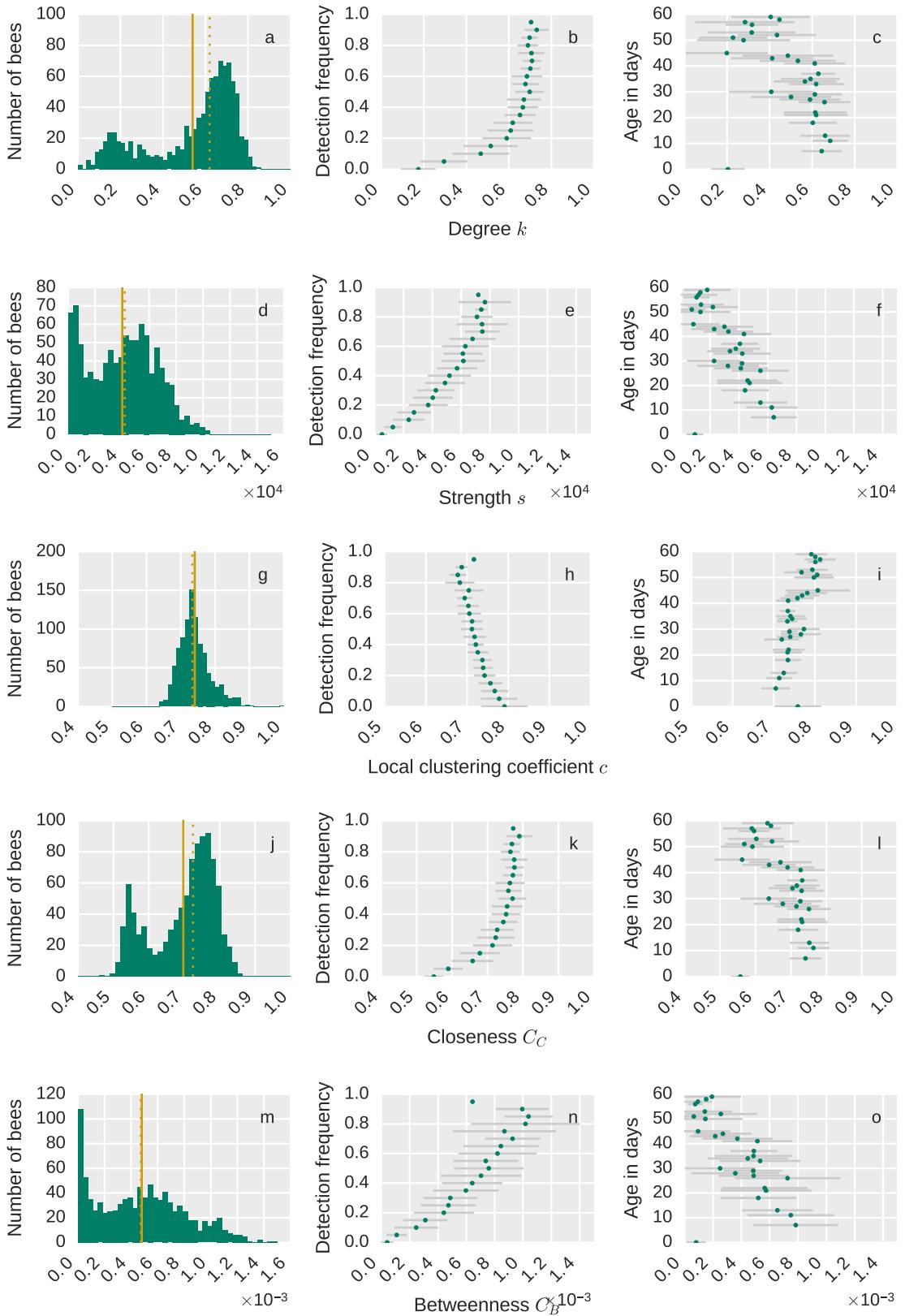
**Figure C.5: Age distribution for LE**



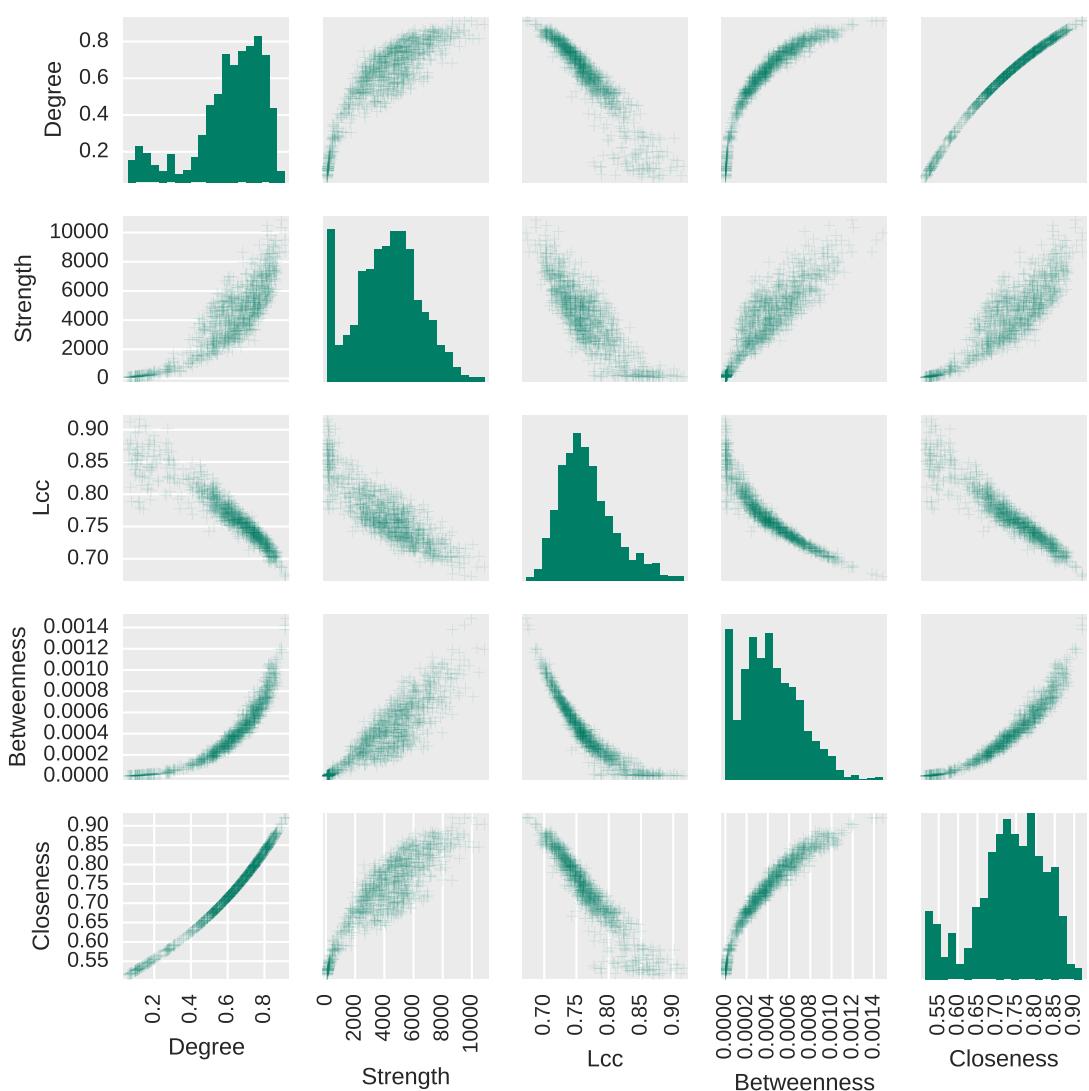
**Figure C.6: Age distribution per snapshot for WT**



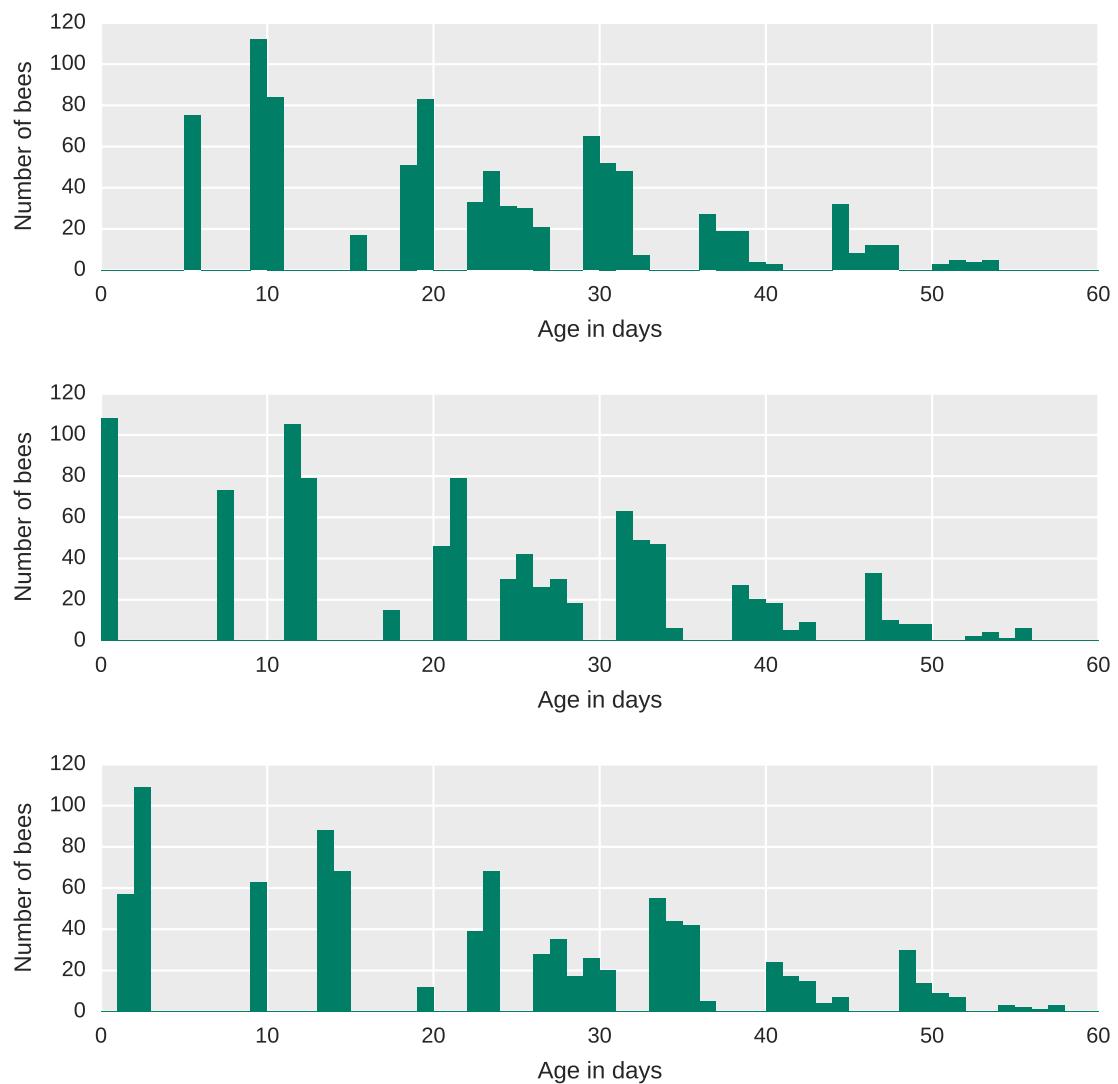
**Figure C.7: Snapshot 1: Local measures in relation to age and detection frequency**



**Figure C.8: Snapshot 2: Local measures in relation to age and detection frequency**



**Figure C.9: Snapshot 3: Scatter plot for node level measure**



**Figure C.10: Age distribution per snapshot**