FACULTEIT
BIO-INGENIEURSWETENSCHAPPEN

MACHINE LEARNING AND POLLINATION NETWORKS: THE RIGHT FLOWER FOR EVERY BEE

Aantal woorden: 31 707

Sarah Vanbesien
Studentennummer: 01300049

Promotor: prof. dr. Bernard De Baets
Copromotor: prof. dr. ir. Guy Smagghe
Tutor(s): dr. ir. Michiel Stock, ir. Niels Piot

Masterproef voorgelegd voor het behalen van de graad: master in de richting Bio-ingenieurswetenschappen: Milieutechnologie.

Academiejaar: 2017 - 2018

UNIVERSITEIT
GENT

Ghent, June 8, 2018

The promoter,	The copromoter,
prof. dr. Bernard De Baets	prof. dr. ir. Guy Smagghe

The tutor,	The tutor,	The author,
dr. ir. Michiel Stock	ir. Niels Piot	Sarah Vanbesien

DANKWOORD

’Oh, een thesis over bloemetjes en bijtjes?’ Het wekt initieël toch net iets senuelere gedachten op dan nodig. Je zou denken dat er wanneer je dan de minder sexy kant van je onderzoek begint uit te leggen veel mensen afhaken, en wel, dat klopt. Gelukkig zit ik in een vakgroep vol mensen die formules en modelleren even interessant vinden als ik.

Allereerst mijn tutor Michiel, naar jou gaat zonder twijfel mijn grootste dank uit! Iedere dag stond je klaar om mijn ontelbare vragen te beantwoorden. Ik kon je altijd bereiken als ik weer eens in de knoop zat met wat code of als ik een *second opinion* nodig had over mijn geschreven stukken. Bedankt voor je geduld en voor alle moeite die je in mijn thesis hebt gestoken! Verder wil ik mijn tweede tutor Niels bedanken, die altijd bijsprong als ik informatie over bijtjes of planten zocht. Bedankt dat je altijd geïnteresseerd was in het verloop van mijn thesis, ook al ligt de focus vrij ver van het onderzoek waar je zelf mee bezig bent. Ik wens jullie beide onnoemelijk veel succes met jullie doctoraat/post doc en alles wat erbij komt kijken!

Naast mijn tutors wil ik prof. De Baets en prof. Smagghe bedanken. Ik vind het zalig dat jullie vakgroepen gezamenlijke onderwerpen aanbieden, zodat onderzoeken als deze kunnen tot stand komen. Voor mij echt een ideale combinatie!!!

Dan is er nog één deur in de gang waar ik heel dikwijls heb aangeklopt. Joris, ongelofelijk merci om uw programmeer skills zo met mij te delen! We konden heel serieus overleggen, maar ook de grootste zever tegen elkaar verkopen in de resto of in de tuin! Ik wens u zoveel succes met uw band en kom met alle plezier kijken naar uw optreden op de Gentse feesten! You rock!

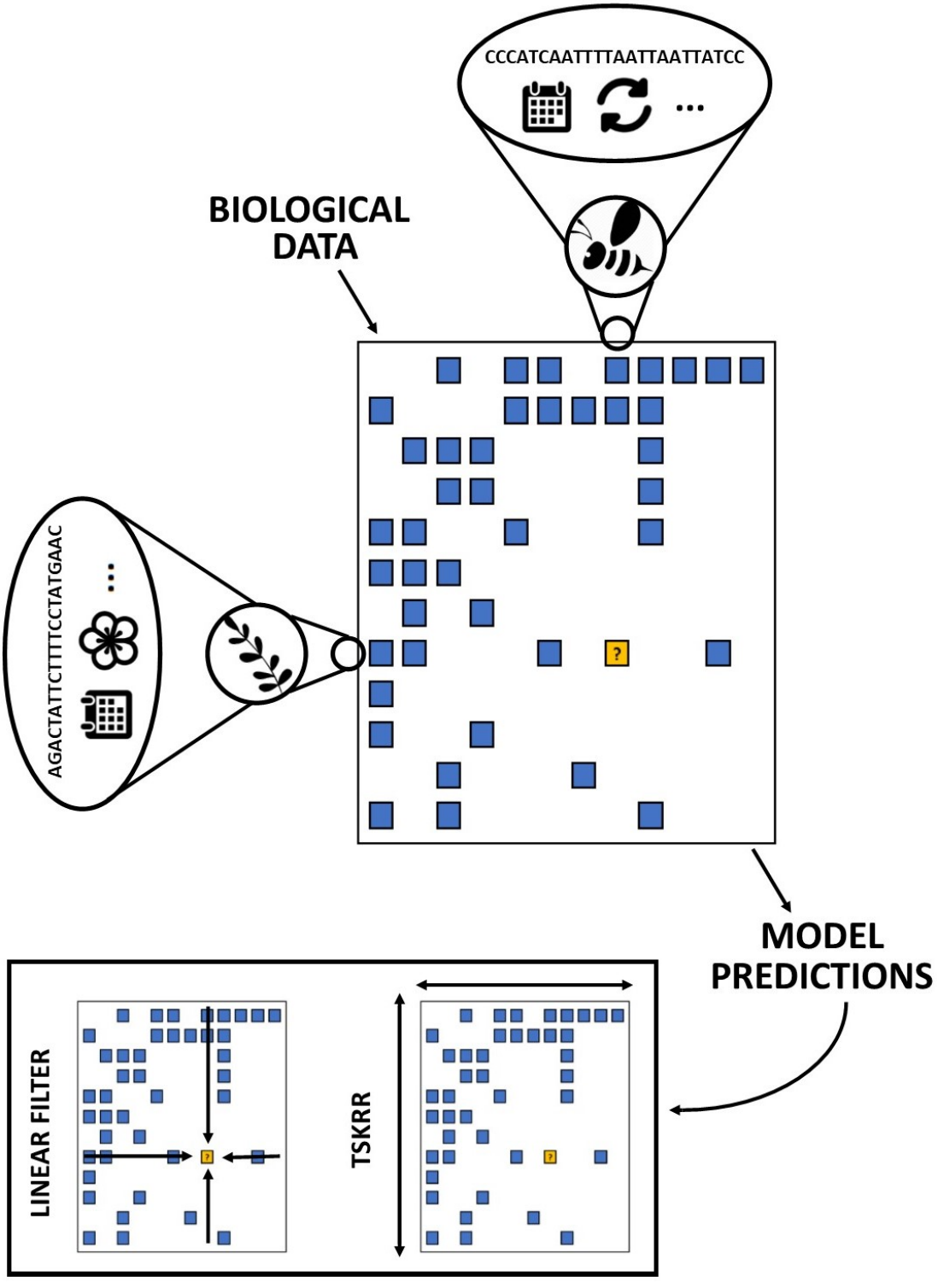
Mam en pap, ook jullie kunnen niet ontbreken! Merci om altijd naar mijn waterval aan Gent-verhalen te luisteren, om mij zo te soigneren tijdens de blok en altijd mee te leven. Bedankt voor alle kansen!

Maar nu, last but not least, het summum van alle vermeldingen, mijn grootste steun en toeverlaat: de VLK! Mijn mede-praesidiumleden hebben me meer weggetrokken van m’n thesis dan omgekeerd, maar zonder hen was mijn jaar nooit hetzelfde geweest. "Uw studententijd is de schoonste tijd van uw leven", "maar uw thesisjaar

telt eigenlijk niet meer", twee statements die ik nog nooit zoveel gehoord heb als nu. Met de eerste ga ik meer dan akkoord, en gelukkig mag ik met de tweede eens goed lachen als iemand ze zegt. Al mijn jaren *aan het boerekot* waren even fantastisch en ik kijk er met zoveel vreugde op terug! Thanks for everything guys!

CONTENTS

GRAPHICAL ABSTRACT



SUMMARY

This thesis handles about the combination of pollination networks and certain machine learning methods. The reason for this is that ecological datasets of interactions are mainly built by aggregating positive observations, giving no clear evidence that the negatively classified (i.e. not-observed) interactions are assuredly non-happening. There may be some missing values in the composed network. Therefore prediction models can be of use, to further finetune existing pollination (and by extension all ecological) networks. The work is centered around one central dataset (of the *Florabeilles* project in France).

In the case of biological interactions, the data is pairwise. Such networks can then be represented by a graph (with all species representing a node) or an interaction matrix (with all species making up the rows and columns). Also, models should be adapted or redesigned to cope with this twofold input. The input of a model is now called a *dyad*, and consists by definition of two instances (here being a plant and a pollinator species). Therefore too, four distinct prediction settings (A,B,C,D) need to be defined, as a *new dyad* is now no longer unambiguously defined. The model can be trained with the information of both species, only one species (plant or pollinator) or none of the species for which a prediction will be made. This leads to four possible settings, where of course the last one mentioned is the hardest one to make predictions for. When possible, performances for all four settings need to be reported, as one single value can give an over- or underestimation of the real performance.

Machine learning is a field of computer science which lets a model learn patterns from data. It provides self-learning algorithms, which can be used in cases where no known or no unambiguously defined rules apply. Two models are presented in this work. The first one is a kind of collaborative filtering, namely a so-called linear filter. This approach takes a lot of known pollination interactions as examples and makes predictions based on the frequency of interaction. In fact it generates a score between zero and one of how likely it is for an interaction to happen. This model does not use any extra information about the species, but only focusses on the binary interaction matrix itself. Still, good results are obtained as an AUC of about 84.3% could be reached. This value was obtained after optimization of the four parameters of the model, although many parameter combinations proved to lead to comparable

values.
Next to this theoretical performance estimate, also a practical validation is done. The predictions of the linear filter model were compared to other real life databases from neighbouring countries. Indeed some not-observed interactions of the dataset, but positively predicted by the filter, seemed to be happening in real life. This proofs that the prioritization of interactions with a model can be of actual use in ecological datasets, to detect missing values.

The second method is called two-step kernel ridge regression and takes, in contrast with the first one, a lot more data with it. It receives information about all comprised species, in a genetic, morphological, environmental and temporal way. Every species is characterized by a vector of information, which is of course modified afterwards to serve as proper input for the model. The collected information can be split up in two sides, being phylogenetic information (i.e. DNA sequences of typical genes) and traits (e.g. height, flower colour, symmetry... of plants and size, flying period, abundance... of pollinators). These can be used separately (leading to a trait-based and a phylogeny-based model) or all aggregated in one model (called the combined model). The used information for a specific model is always stored in two kernel matrices (defining similarity): one for the pollinators and one for the plant species. Then two successive kernel ridge regressions are executed, using the labels of the binary interaction matrix.
There are two regularisation parameters in the model. With an initial - arbitrary chosen - parameter combination, the combined model achieves the best results. When the performance estimations are made using nested cross-validation, both the trait-based model and the combined model seem to score quasi equally. The concept of nested cross-validation is to separate the final test set from all parts (folds) used for parameter optimization. In this way the estimate is more fair. The 'honest' AUC's of the combined model for setting A, B, C and D are respectively 87.3%, 81,3%, 86.6% and 81.1%. The fact that setting C scores a lot higher than setting B shows that is easier for the model to generalize to new plants. This is mainly due to larger amount of plants (almost double of the pollinators) that has been included in the training.

Both models proved to be able to recreate the most likely interactions of a dataset. These predictions can be used to detect probable missing values, which can afterwards be prioritized in further field research. Not all predicted values necessarily have to happen in reality, but the prioritization of them can be a time-saving factor in ecological research.

SAMENVATTING

Deze thesis behandelt de combinatie van pollinatienetwerken en bepaalde machine-learning methoden. De aanleiding hiervoor is dat ecologische datasets voornamelijk worden opgesteld door positieve waarnemingen te aggregeren, zonder een duidelijk bewijs te leveren dat negatief geclassificeerde (dus niet-geobserveerde) interacties gegarandeerd niet plaatsvinden. Bepaalde interacties kunnen ontbreken in het opgestelde netwerk. Daardoor kunnen predicitiemodellen nuttig zijn om bestaande pollinatie (en bij uitbreiding alle ecologische) netwerken verder te verfijnen. Dit werk is gecentreerd rond één dataset (van het *Florabeilles* project in Frankrijk).

In het geval van biologische interacties is de data paarsgewijs. Zulke netwerken kunnen dan voorgesteld worden door een graaf (waar ieder species een node voorstelt) of een interactiematrix (waar de species de rijen en kolommen van uitmaken). Daarnaast moeten modellen ook worden aangepast om met deze dubbele input om te gaan. De input van een model wordt nu een *dyad* genoemd en bestaat per definitie uit twee items (hier een plant en een bestuiver). Hierdoor dienen ook vier verschillende predictiesettings (A, B, C, D) te worden gedefinieerd, omdat een *nieuwe dyad* nu niet langer eenduidig gedefinieerd is. Het model kan namelijk getraind zijn met de informatie van beide species, slechts één species (plant of bestuiver) of geen van beide species waarvoor de predictie gemaakt wordt. Dit leidt tot vier mogelijke settings, waarbij de laatstgenoemde duidelijk de moeilijkste is om voorspellingen voor te doen. Indien mogelijk moeten steeds de performanties van alle vier de settings worden gerapporteerd, omdat één enkele waarde een vertekend beeld kan geven (een over- of onderschatting van de werkelijke prestatie).

Machine learning is een tak in de informatica waarbij men een model patronen laat herkennen in data. Het gaat om zelflerende algoritmen, die kunnen worden ingezet voor toepassingen waarbij geen gekende of geen ondubbelzinnig gedefinieerde regels gelden. In dit werk worden twee modellen toegelicht. De eerste is een soort collaboratieve filtering, namelijk een lineaire filter. Deze methode vergt veel bekende bestuivingsinteracties als voorbeeld en maakt nadien predicties op basis van de interactiefrequenties. In feite genereert het een score tussen nul en één die beschrijft hoe waarschijnlijk het is dat de interactie plaatsvindt. Dit model gebruikt geen extra informatie over de species, maar richt zich enkel op de binaire interactiematrix. Toch

worden goede resultaten bekomen, met een AUC van ongeveer 84.3%. Deze waarde is bekomen na optimalisatie van de vier parameters van het model, al bleken veel parameter-combinaties vergelijkbare waarden op te leveren.
Naast deze theoretische performantie, is ook een praktische validatie uitgevoerd. De voorspellingen met de lineaire filter werden vergeleken met andere real-life databases van naburige landen. Sommige niet-geobserveerde interacties van de dataset, maar positief voorspeld door de filter, bleken inderdaad in de natuur voor te komen. Dit bewijst dat focussen op positief voorspelde interacties daadwerkelijk nuttig kan zijn voor het detecteren van ontbrekende interacties in ecologische datasets.

De tweede methode wordt tweestaps kernel ridge regressie genoemd en gebruikt, in tegenstelling tot de eerste, veel meer gegevens. Het verwerkt informatie van alle species van het netwerk, zowel genetisch, morfologisch, ecologisch als temporeel. Elk species wordt gekenmerkt door een vector van informatie, die uiteraard nadien wordt aangepast om als geschikte input voor het model te fungeren. De verzamelde data over de species kan opgesplitst worden in twee luiken, zijnde fylogenetische informatie (nl. DNA-sequenties van typerende genen) en *traits* (bijv. hoogte, bloemkleur, symmetrie... van planten en grootte, vliegperiode, abundantie... van bestuivers). Deze kunnen afzonderlijk worden gebruikt (leidend tot een *trait*-gebaseerd en een fylogenie-gebaseerd model) of geaggregeerd in één model (het gecombineerd model genoemd). De informatie gebruikt voor een bepaald model is steeds opgeslagen in twee kernel matrices (die gelijkenissen definiën: één voor de pollinatoren en één voor de planten. Nadien zijn twee opeenvolgende kernel ridge regressies uitgevoerd, gebruik makend van de labels van de binaire interactiematrix.
Met een initiële - arbitrair gekozen - parametercombinatie, behaalt het gecombineerd model de beste resultaten. Wanneer de prestatie-inschatting wordt gemaakt met behulp van geneste kruisvalidatie, blijken het *trait*-gebaseerde model en het gecombineerde model quasi even goed te scoren. Het concept van geneste kruisvalidatie is om de definitieve testset gescheiden te houden van alle delen die gebruikt worden voor parameteroptimalisatie. Op deze manier is de inschatting objectiever. De AUC's van het gecombineerd model voor settings A, B, C en D zijn respectievelijk 87.3%, 81,3%, 86.6% en 81.1%. Het feit dat setting C een stuk hoger scoort dan setting B toont aan dat het model makkelijker generaliseert voor nieuwe planten. Dit komt hoofdzakelijk door het feit dat er dubbel zoveel planten als pollinatoren zijn meegenomen in de training.

Beide modellen hebben aangetoond in staat te zijn de meest waarschijnlijke interacties van een dataset te reconstrueren. Deze voorspellingen kunnen worden gebruikt om mogelijks ontbrekende waarden te detecteren, waar later op kan worden gefocust in verder veldonderzoek. Niet alle voorspelde interacties hoeven in de realiteit

te gebeuren, maar de prioriteitstelling ervan kan een tijdsbesparende factor zijn in ecologisch onderzoek.

INTRODUCTION AND OUTLINE

Plant-pollinator interactions are one of the many interaction types we know in nature. Pollination is of major importance for both species biodiversity and humans. A great biodiversity of plants and animals depend mutually on each other for pollination and food, and also people depend on this mutualism for their feeding habits [30]. Pollination is still a hot topic in literature (as e.g. managed honeybee stocks are getting more controversy [39], or as climate change starts to influence species phenologies and abundances [49]).
A lot of field research has been done to document pollination behaviour and to set up datasets of which pollinators are interacting with which plants. Years and years of observation are performed, and with this data ecological datasets are constructed by aggregating positive results. All detected pollinations are classified as positive, but one is never sure that the remaining plant-pollinator combinations are surely non-happening in this neighbourhood. Indeed, there are some so-called forbidden links between species (by the lack of compatibility), but other interactions might just have been missed during the timespan of assembling the dataset.
For this reason, researchers are always looking for new methods to review pollinations networks. DNA-barcoding methods become more popular, but here we focused on machine learning methods. The concept will be elucidated and two techniques will be discussed.

The main goal of this work is to provide well-functioning algorithms that can predict whether a specific plant and pollinator will interact or not. The purpose is however not to fully predict a purely hypothetical interaction network (although this would be a proper use of the model), but more to finetune existing networks. As stated earlier, datasets are often a collection of positive observations. In this way, no evidence is created that negative interactions are undoubtedly impossible or not occurring. Prediction algorithms can prioritize interactions for further research or highlight species that deserve special attention. In this way, possible missing interactions (so-called false negatives) in the dataset can be traced and completed. Machine learning will never replace field research, as a lot of training examples are necessary for the model's build up, but in this way mutual benefits can take place.

The work consists of 5 chapters. The first chapter starts with an introduction about ecological networks. They are approached from a biological point of view, but also a more mathematical side of interaction networks is highlighted. The second chapter is about machine learning. Two different algorithms are elucidated: collaborative filtering and kernel methods. Afterwards a section is dedicated to the proper ways of estimating performance of these built models. The last theoretical section of chapter 2 describes the concept of optimal transport, which will later be used in a practical way. Chapter 3 describes the main dataset of the work. Everything is centered around this dataset, so a proper introduction is fundamental. Different properties of the network are defined, and all comprised species are considered in multiple ways. Chapter 4 presents and interprets the results with the two different machine learning approaches. A parameter optimization is done, a theoretical evaluation and a practical validation. Results are compared and an explanation of the different patterns is given. At last chapter 5 does not focus on the predictive part, but more on the interacting part. Here the theory of optimal transport is connected to the main dataset. Again properties of the originating network are computed.

CHAPTER 1

ECOLOGICAL NETWORKS

1.1 Introduction and representation

In nature, many ecological and biological networks are present, ranging from large-scale trophic food webs to interactions at molecular level like protein-ligand interactions. In this work, the focus lies on species interaction networks.
There are five different types of biotic interactions, distinguished by the effect on the interacting species.

- Competition (–,–) means that the interaction is detrimental for both involved species, e.g. the lion (*Panthera leo*) and the spotted hyena (*Crocuta crocuta*) both feeding on the same resource.
- Mutualism (+,+) is an interaction type beneficial for both species, e.g. the shelter-defence interaction between the acacia ant (*Pseudomyrmex ferruginea*) and the bullhorn acacia (*Acacia cornigera*).
- Predation and parasitism (+,–) both belong to the interaction class where one species benefits from the interaction, while the other one gets harmed. A predation example is the Iberian lynx (*Lynx pardinus*) hunting and consuming the common rabbit (*Oryctolagus cuniculus*); an example of species with a parasitism relationship are the *Anopheles gambiae* mosquito (and on its turn humans) being host organism for *Plasmodium falciparum*, the unicellular protozoan causing malaria.
- Commensalism (+,0) holds an advantage for one of the interacting species, without helping or harming the second one, e.g. spearfish (*Remora brachyptera*) attaching themselves with a sucker to blacktip reef sharks (*Carcharhinus melanopterus*).
- Amensalism (–,0) at last occurs when one interacting species stays indifferent, while the other is negatively affected, e.g. *Penicillium expansum* inhibiting growth and life of many types of bacteria [40].

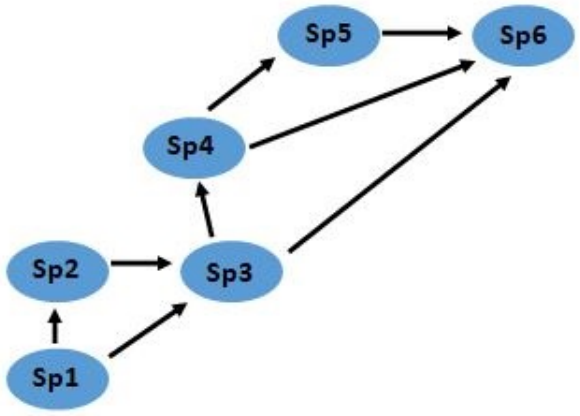
The two most common (and in fact equivalent) ways to represent an ecological network are a graph and an interaction matrix. In a graph, the nodes represent the species and the edges are the lines connecting the nodes. When the relation between two species is asymmetric (e.g. in a food web where a predator eats a prey), the network is represented as a directed graph or digraph. Here, the edges are arrows from one node to the other, originating in the prey node and terminating (with an arrowhead) at the predator node [1]. A food web is also an example of a unipartite (or homogeneous [59]) network, meaning there is only one set of species interacting with one another. Each species eats and can be eaten. The interaction matrix, further denoted as Y, is square: it contains all species in all rows and the *same* set of species in all columns. An interaction value then indicates whether the species in the row eats the species in the column (1) or not (0). In bipartite networks on the other hand, there are two different sets of species (e.g. pollinators and plants), meaning that the interaction matrix does not have to be square. The rows represent one set, the columns the other, and the value of $Y_{i,j}$ quantifies whether there is an interaction between species i and j (1) or not (0) [56]. All this is illustrated in Table 1.1. We assume that each species (and hence node of the graph) is described by a feature vector x, containing any relevant information on phylogeny, morphology or other characteristics.

In the binary context of $Y_{i,j}$ being 0 or 1, the positive class (1) denotes interacting species and the negative class (0) non-interacting species. This is not to be confused with the biological meaning of positive and negative effects on the interacting species, as stated above with five possible relationships [58].
The other possibility is a real-valued interaction matrix. In this case, $Y_{i,j}$ describes how relevant every interaction is, for example by the number of pollinator visits to a plant. In this way one can distinguish between strong and rather weak interactions in a network. The data to construct these quantitative matrices is collected by field observations, so the outcome needs to be handled with care. Researchers are never sure the observed abundances are also the real abundances (i.e. the real ratio of species presences or the real ratio of species interactions). Making matrices binary can cause a potential loss of information but avoids the problem of uncertain abundances [64].

The advantage of using the matrix approach and associated linear algebra is that one may then deal with complex systems but now in arbitrary dimensions [1].

	Species1	Species2	Species3	Species4	Species5	Species6
Species1	0	1	1	0	0	0
Species2	0	0	1	0	0	0
Species3	0	0	0	1	0	1
Species4	0	0	0	0	1	1
Species5	0	0	0	0	0	1
Species6	0	0	0	0	0	0

	Plant1	Plant2	Plant3	Plant4	Plant5	Plant6	Plant7	Plant8	Plant9
Pollinator1	1	0	0	0	0	0	1	0	0
Pollinator2	0	1	0	0	0	0	1	0	0
Pollinator3	0	1	0	0	0	1	1	1	0
Pollinator4	1	1	0	1	0	0	1	0	0
Pollinator5	1	1	1	1	1	0	1	1	1
Pollinator6	0	1	0	0	1	0	0	0	0



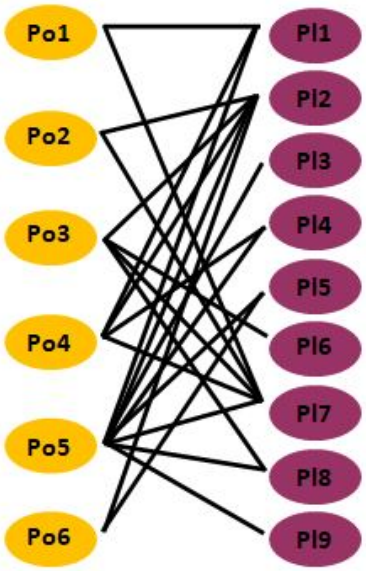


Table 1.1: Visualisation of the interaction matrix and the corresponding graph for a unipartite directed food web (left) and a bipartite undirected pollination network (right).

1.2 Properties of networks

Species are usually heterogeneously distributed in an ecological network, in a such way that most species only have a few interactions, while a few species are much more connected than expected by chance [23]. *Nestedness* is a meaningful measure for bipartite networks to quantify the latter, representing the extent to which the interactions of specialists and generalists[1] overlap. The nestedness for the row set of species and the column set of species is respectively given by $\eta^{(R)}$ and $\eta^{(C)}$. Bastolla et al. (2009) defines the nestedness of these margins as

$$\eta^{(R)} = \sum_{i<j} \frac{n_{i,j}^{(R)}}{\min(n_i^{(R)}, n_j^{(R)})} \qquad \text{and} \qquad \eta^{(C)} = \sum_{i<j} \frac{n_{i,j}^{(C)}}{\min(n_i^{(C)}, n_j^{(C)})}, \tag{1.1}$$

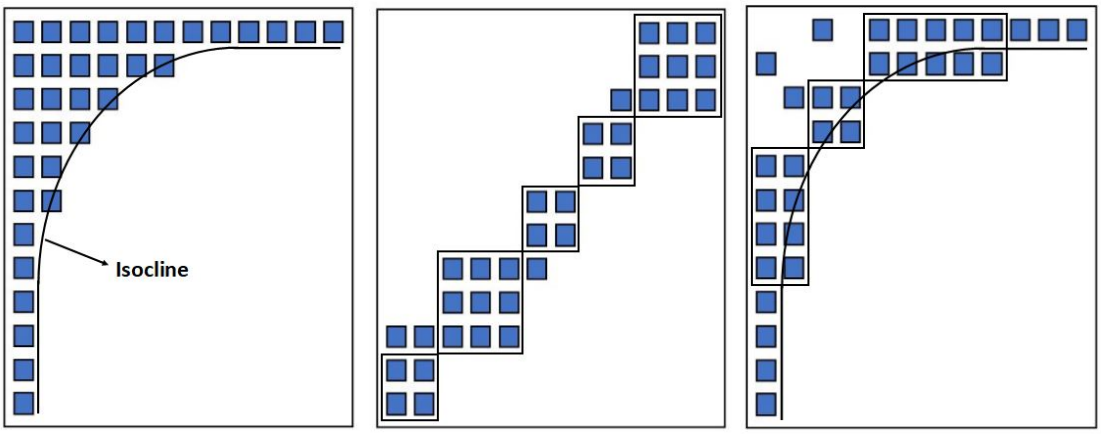
where $n_i^{(R)}$ and $n_j^{(R)}$ are the number of interactions of row species i and j respectively, where $n_{i,j}^{(R)}$ denotes the number of shared interactions between these two row species, where $n_i^{(C)}$ and $n_j^{(C)}$ are the number of interactions of column species i and j respectively and where $n_{i,j}^{(C)}$ denotes the number of shared interactions between these two column species [7].

The nestedness of the whole network is the average of those of the two margins, defined as $\eta = \frac{\eta^{(R)}+\eta^{(C)}}{2}$. All three metrics lie between zero and one, where $\eta = 1$ defines complete nestedness [46]. Graphically, complete nestedness is indicated with an isocline in the interaction matrix Y. An example of a nested network can be seen in Figure 1.1a [23]. It has been shown that a nested structure minimizes competition and increases the number of coexisting species [7]. All mutualistic networks (like e.g. plant-pollinator networks) tend to show a nested behaviour. In these we can hence say that a new species will be more likely to interact with a generalistic species of the other set than with a very specialized one [53]. This idea will be crucial in a later described linear filter model (Sections 2.2.1 and 4.1).

A related variable is *modularity*. This does not define how much specialists and generalists are connected, but represents the tendency for subsets of species to be strongly connected, while they are weakly connected to the rest of the network. It quantifies how compartmentalized a network is. In this context, one defines a module as a densely connected subset of species, not overlapping with other subsets. Mathematically, with Q the modularity:

$$Q = \frac{1}{2k} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(Y_{i,j} - \frac{n_i n_j}{2k} \right) S_{i,j}\,. \tag{1.2}$$

[1] It can here be assumed that specialists are species only interacting with one or a select group of partners while generalists are interacting with a wide variety of other species. A more scientific definition of both is given in the following section.



(a) Nested matrix. (b) Modular matrix. (c) Nested & modular matrix.

Figure 1.1: Example networks to demonstrate the properties of nestedness and modularity.

The derivation of the formula is based on ratios, and more precisely the portions of interactions occurring between nodes of a number of modules (subsets) opposed to how many interactions that would be randomly expected. Therefore a modularity of zero represents complete random links. k stands for the total sum of all values in the interaction matrix: $\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{i,j}$. n_i represents again the number of interactions species i establishes. The S is a $(n \times m)$ matrix with a 1 if species i and j belong to the same module and 0 otherwise [41]. As modules are composed of species having many interactions among themselves while having very few with species of other modules, a blocked structure can be recognised [23]. This structure can be seen in Figure 1.1b. Nestedness and modularity will be used in Chapter 5.

Interaction matrices are a handy tool for network description, but still there is a remark to make. Actually, an interaction between two species is no pure yes-no-event, as the occurrence of the interaction may be rare or may depend on several local and behavioural circumstances. Therefore a lot of variation can exist between ecological networks, so only presenting them with fixed graphs or interaction matrices may not be sufficient. To solve this, the question 'Do these two species interact' is replaced by 'How likely is it that these two species interact?', in which the interactions are treated as probabilities. Now some metrics borrowed from the information theory can be used to characterize the network, because this is done by modelling each interaction as a Bernoulli experiment and by calculating the expected value of each metric. Hence, Y is a matrix where each element is $P_{i,j}$, being the probability that species i establishes an interaction with species j. It is assumed that all interactions are independent and can be represented as a series of Bernoulli trials[2], so that $0 \leq P_{i,j} \leq 1$ [46]. The information theory is a branch of the mathematical theory of probability and statistics. Its formulas are quite abstract but are applicable in all systems with a probabilistic

[2] A Bernoulli trial is the realization of a probabilistic event that gives 1 with probability p and 0 otherwise.

or statistic basis. Then a, for example, ecological meaning can be assigned to the derived metric values [36].

Next to nestedness and modularity, now some metrics based on this information theory are discussed. *Entropy* is frequently used in thermodynamics but also plays a role in the information theory. This actually tells how surprising a certain outcome is. When the probabilities of all possible outcomes of an event are equal, the entropy is maximal because the effective outcome will be most surprising. To illustrate this with a small example: when one throws a dice, all six possible outcomes are of equal probability. The random variable X is defined as the outcome of the experiment, and $p(X = x)$ as the probability that the outcome X takes value x (for example $p(X = 2) = \frac{1}{6}$). Because every $p(X = x)$ is identical, we say that its probability distribution $p_X(x)$ is uniform. Intuitively in this scenario, the outcome of an experiment is the hardest to predict. It will be most surprising, so the entropy is high.
When on the other hand the distribution of all outcomes is very favoured towards one specific value, the outcome is easier to predict and hence, the entropy is low. To stay with the same example: when the dice is adulterated and almost only lands on one of its six sides, the outcome will not be so surprising. The probability distribution $p_X(x)$ is no longer uniform but has one exceptionally high value. It is not hard to predict the outcome. The entropy is low.
Entropy is mathematically defined as

$$H(X) = \sum_{x \in A_X} p_X(x) \log_2 \frac{1}{p_X(x)}, \tag{1.3}$$

with X a random variable, p_X its distribution and A_X the set of all possible values of X. When only the set of pollinators or only the set of plants is filled in for X, this results in the marginal entropies.

The *joint entropy* of two variables X and Y is given by

$$H(X, Y) = \sum_{x \in A_X, y \in A_Y} p_{X,Y}(x, y) \log_2 \frac{1}{p_{X,Y}(x, y)} \tag{1.4}$$

and the *conditional entropy* of the variable X for a given $Y = y$ by

$$H(X|Y = y) = \sum_{x \in A_X} p_{X|Y}(x|y) \log_2 \frac{1}{p_{X|Y}(x|y)}. \tag{1.5}$$

To make this clear, if in a (for example plant-pollinator) interaction matrix a plant only interacts with one specific pollinator, the conditional entropy of this plant is zero. There is no uncertainty or surprise when determining its interaction partner.
To use terminology that will be helpful afterwards: Conditional entropies can be in-

terpreted as the expected number of binary questions that have to be asked to determine the particular species of an interaction, when the interaction partner is known [35].

Joint entropy, marginal entropy and conditional entropy can all be linked to each other by following equation:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)\,. \tag{1.6}$$

Finally the *mutual information* and the *variance of information* are defined as

$$MI(X;Y) = H(X) - H(X|Y)\,, \tag{1.7}$$

$$VI(X;Y) = H(X,Y) - MI(X;Y) = H(X|Y) + H(Y|X)\,. \tag{1.8}$$

The mutual information is the average reduction of uncertainty of X when the value of Y is known. It is the average amount of information that Y contains about X. It can be verified that $MI(X;Y)$ is equal to $MI(Y;X)$. The variance of information on the other hand is the sum of the average uncertainty that stays on X when Y is known and the average uncertainty that stays on Y when X is known. It is now clear that $H(X,Y) = MI(X;Y) + VI(X;Y)$.

The previously defined metrics can be linked to a uniform distribution, where the entropy is maximal (as seen earlier). When their probability distribution is uniform, the entropy of respectively X and Y can be computed as

$$H(U_X) = \log_2 |A_X| \qquad \text{and} \qquad H(U_Y) = \log_2 |A_Y|\,. \tag{1.9}$$

The difference in entropy between distributions and their uniform distributions is now

$$\Delta U = \big(H(U_X) - H(X)\big) + \big(H(U_Y) - H(Y)\big)\,. \tag{1.10}$$

By combining these metrics, one can determine the entropy of a uniformly distributed matrix as a formula with clear mathematical and afterwards ecological meaning:

$$H(U_{XY}) = \Delta U + 2MI(X;Y) + VI(X;Y)\,. \tag{1.11}$$

The relative contribution of the three components in this equation gives an indirect quality measurement of the information. The first term ΔU describes how strong the matrix deviates from the uniform distribution. When this term is large, this is an indication that one or more species (or species interactions) are overabundant in the

network. The second term is the mutual information. The larger this term, the better the information transfer from X to Y (i.e. knowing X reveals much info about Y) and vice versa. For the species this means that their number of possible interaction partners is limited. A matrix in which $MI(X;Y)$ approximates $H(U_{XY})$ has a high predictive power. It is in that case not hard to predict the interaction a species will establish, as the species are quite specialized (see the second example below). Finally, the remaining uncertainty is $VI(X;Y)$. The variance of information is uncertainty that cannot be explained by the other terms. The lower this term, the better, as this would increase the information that is obtained about X when Y becomes known and vice versa. The three extremes of contribution of the different terms are shown with respective toy networks:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

One interaction dominates the network.
This means that ΔU is maximal: both marginal distributions deviate completely from uniform distributions. The mutual information and the variance of information are both zero.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Perfect speciation of the network.
Here $MI(X;Y)$ is maximal: knowing X reduces the uncertainty on Y completely as when X is known, Y is known immediately. The variance of information is zero, and so is ΔU because both marginal distributions are uniform.

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Heterogenous network without speciation.
Now $VI(X;Y)$ is maximal because the uncertainty about X remains unchanged when Y is known. The mutual information is zero, and so is ΔU because both marginal distributions are uniform [64].

To end this section, a trade-off in ecological networks is highlighted. This can best be done by slightly transforming the previous equation to

$$H(U_X) = \Delta U_X + MI(X;Y) + H(X|Y)\,, \tag{1.12}$$

where

- $H(U_X)$ is the entropy of the network if all interactions would be uniformly distributed over the species, i.e. their freedom of choice is maximal.
- ΔU_X expresses again how much the observed data differs from a uniform distribution.

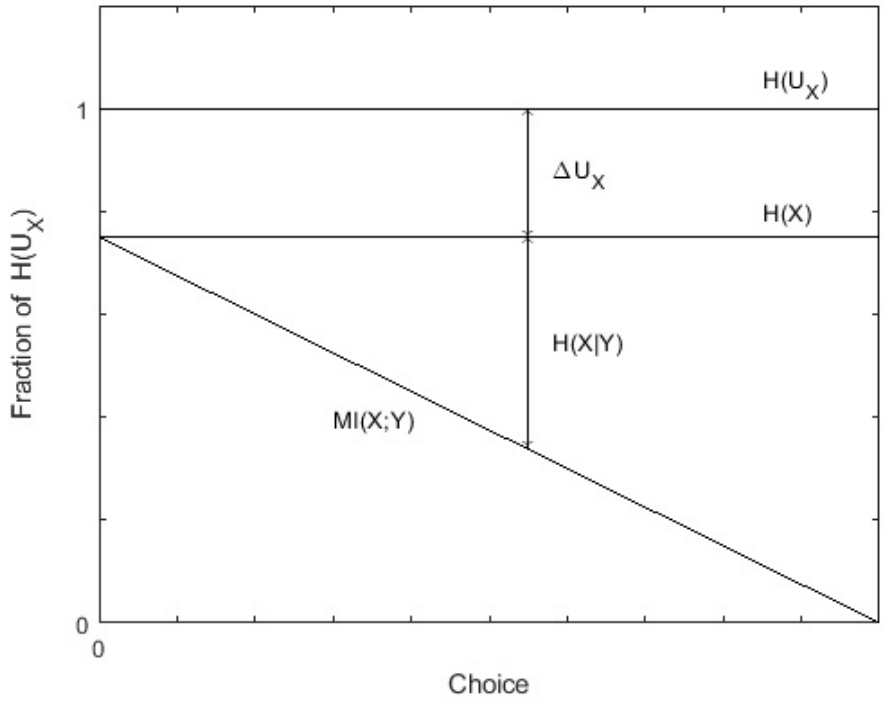
- $MI(X;Y)$ quantifies the organisation of the network, i.e. the limitation on possible interactions between X and Y. The ecological interpretation of this metric is that a restricted number of interactions can lead to higher efficiencies.

- $H(X|Y)$ is part of the uncertainty that remains if the whole structure of the interaction network is known. A large conditional entropy means that the species have a large variety of possible interaction partners out of which they can choose (so large uncertainty). This metric can be seen as a measure of network stability. Too strong restrictions on the interactions and on freedom of choice of the species decreases the stability.

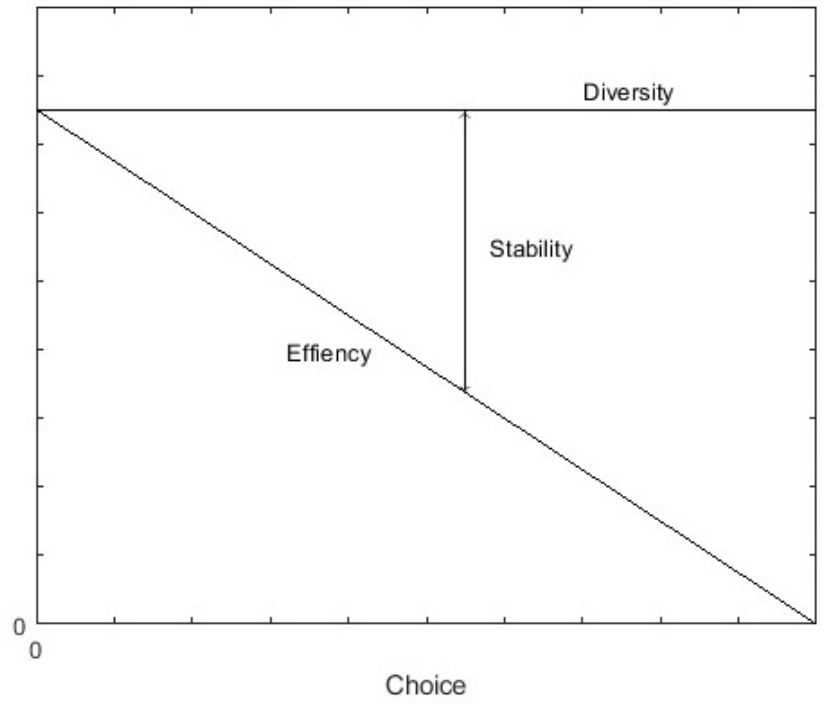
As can be seen, efficiency and stability are antagonistic: one comes at the cost of the other. If the freedom of choice of the species becomes larger, meaning they have a broader variety of interaction partners, the stability of the overall network grows *but* the efficiency of their interactions goes down. Visualisations in mathematical and ecological terms are presented in Figure 1.2a and 1.2b, respectively. There it is assumed that the deviation of the uniform distribution, and hence the diversity of the species in the network, remains constant [51].

1.3 Specialists versus generalists

The terms 'specialist' and 'generalist' were already mentioned in Section 1.2, but will be further developed here. Globally they can be distinguished by saying that the former only interacts with one (or a very select group of) species, while the latter has a lot of possible interaction partners, but there are more scientific approaches available. Here the Optimal Diet Model is elucidated (a kind of Optimal Foraging Theory), applied for a mobile predator feeding on stationary prey. Subsequently, this could easily be transformed to the setting of pollination, as pollinators are mobile and flowers are stationary. In the predation context, the following metrics are to be taken into account: E, the amount of energy that the prey provides to the predator; S, the search time (i.e. the time necessary to find the prey, which is dependent on the abundance of the prey and the ease of locating it); and h, the handling time (i.e. the time necessary to catch and consume the prey, starting from the point where the prey is found). The ratio E/h is called the profitability of a prey. We now assume $prey_1$ with energy E_1 and handling time h_1 and $prey_2$ with energy E_2 and handling time h_2, and assume that the profitability of the first one is the highest: $\frac{E_1}{h_1} > \frac{E_2}{h_2}$. When the predator encounters $prey_1$, it should always choose to eat it without considering the one with the lower profitability. However, if the predator encounters $prey_2$, it should reject it and search for $prey_1$, *except* when this is no cost-effective option. The latter means



(a) Trade-off in mathematical terms.



(b) Trade-off in ecological terms.

Figure 1.2: Trade-off in ecological networks
When the freedom of choice increases (i.e. every species has more possible interaction partners), the uncertainty over the established interactions increases. (a) The conditional entropy increases, with a decrease in mutual information as a result. Knowledge of X now reveals less information on Y (the interaction partner). (b) A greater conditional entropy can be translated to a greater stability, and a smaller mutual information to a smaller efficiency of the interactions [51].

that it would take too much time to find the more profitable one that it is actually not worth it. Hence we say that the predator should eat $prey_2$ if $\frac{E_2}{h_2} > \frac{E_1}{h_1+S_1}$ with S_1 the search time of $prey_1$. It is this equation that we need to define the two concepts of this section.
The equation can be rearranged to get: $S_1 > \left(\frac{E_1 h_2}{E_2}\right) - h_1$. S_1 can be seen as a threshold in the choice of a prey. Animals that have values of S_1 reaching the threshold are defined as generalists. They include a wide variety of preys in their diet. Animals that have values of S_1 below the threshold are defined as specialists. They are better off by exclusively eating one prey. A switch between these two feeding strategies can be made depending on the abundance of preys. Since it is always favourable to eat $prey_1$, the choice of eating this one is not dependent on the abundance of $prey_2$. However, since the choice of eating $prey_2$ is dependent on S_1, this choice is dependent on the abundance of $prey_1$. Hence, when the food of a specialist becomes scarce (i.e. the abundance of $prey_1$ is too low), a specialist could sometimes switch to become a generalist. The model can of course be extended for more than two preys, but this extension is not further treated here [48].
When the Optimal Fouraging Theory is applied to pollination, two things need to be taken into account. Firstly, nectar is assumed to be a non-depleting food resource, which is not valid for all prey types. Secondly, bees do not tend to maximize their net rate of energy gain (i.e. the previously defined profitability, the net energy gained per time unit), but rather their energy efficiency (i.e. energy gained per energy spent). The reason for this is that to maximize the former, tremendous loads of nectar should be carried in one flight. The weight of the nectar adds a significant cost to the bee's flight between flowers and can even shorten the lifespan of the bee. Therefore, the metabolic cost of the transport of nectar should be included in the model. Afterwards, all the same concepts based on thresholds can be applied. The maximization of energetic efficiency is just an adaptation to a limited flight-cost budget [54]. To continue, pollination and pollination networks are first to be more properly defined.

1.4 Pollination networks

Pollination is a specific type of biological interaction. Pollination literally means the transfer of pollen from a male anther to a female stigma, where pollen is a sex cell of plants containing its genes. These cells are essential for reproduction.

Darwin (1877) already described distylous flowering plants as species characterized by possessing two types of flowers, being the 'pin' flowers (with long styles and short stamens) and the 'thrum' flowers (with short styles and long stamens). Each individual of such plant species only bears one of these two types. This was probably

to prevent fertilization after accidental self-pollination so that in this way, a pollen transfer resulting in successful fertilization only occurs between different individuals. Later distyly evolved to dioecy, where male and female flowers no longer contain parts of the opposite sex. Presumably an evolutional change in pollinator types (e.g. shortening of the mouth parts) induced the inutility of the short styles and short stamens [8].

Just like other ecological networks, pollination networks can be represented as a graph or as an interaction matrix. Here, arrows in the graph are not necessary as a direction of the interactions is not really applicable. Matrices can be binary or quantitative, with the same advantages and disadvantages as mentioned above ((+) more information in the quantitative case but (-) uncertainty about abundances). Moreover, other sources about pollination mention a second problem with taking the number of visits as an interaction value. Identifying the main pollinator for a floral species is based on two components of animal activity: frequency of visits and effectiveness of pollen transfer to appropriate stigmas in each flower visit. Most studies only consider the first aspect because presence of visitors is more easily observed than the transfer of pollen. Pollinator effectiveness can be qualified by a variety of metrics: the number of pollen grains deposited per visit, the amount of both pollen deposited on stigmas and pollen removed from anthers, the frequency with which each visitor species contacts anthers and stigmas, fruit set per visit or seed set per visit [22]. Although newer methods have been developed to define the effectiveness, like DNA meta-barcoding, the component is often left behind [19]. This incurs the risk of misidentifying the main pollinator and mistaking the specialized system as a generalized system [22]. For previously and here mentioned reasons, taking binary values instead of visitation frequencies could be seen as a safe decision.
If the risk of misidentifying the main pollinator would really induce a problem is left aside, as this is quite contradictory in literature. Only a few studies compared the visitation frequency with the effectiveness of pollen transfer. Armbruster (1985) found that the most common visitor is mostly a poor pollinator. These pollinators could actually be seen as parasites from the plant's point of view, as they remove pollen that otherwise would be transferred more effectively to stigmas of other individuals. Here, the importance of the common pollinators is greatly overestimated, and the importance of less common but highly effective ones underestimated by only taking into account visitation frequencies [3]. Olsen (1997) on the other hand states that the most common visitors are also the most important pollinator [43].

Pollination is mostly of the mutualism type (see Section 1.1). However, the relation is rather asymmetrical as pollinators are typically more specialized than plants, so being more dependent on this plant's abundance [40]. The benefit for the plants is

the ability to mate with other individuals; the benefit for the pollinator is the collection of nectar and pollen, by which they transfer this pollen from the anthers to the next flower [2].

Pollination is quickly linked to bees, but in fact various types of pollinators are known. To start with the insects, there are of course bees (the domesticated honey bees and wild bees such as bumblebees and solitary bees), but also beetles (e.g. oil beetles, long horn beetles and swollen thigh beetles), wasps (e.g. fig wasps and orchid wasps), butterflies, moths (e.g. hawk moths), thirps and flies (e.g. hoverflies) [2]. Fig wasps are famous because of their co-evolution/co-speciation and intimate relation with fig plants. In total, more than 750 fig species are known, all owning their own pollinating wasp species of the Agaonidae family. In theory, the 'one-to-one rule' would apply for this pollinating interaction, but in practice this rule is sometimes violated. Still an intense mutualism exists between the species, as the plant depends on its wasp for pollination and the wasp on the fig for reproduction (their larvae feed themselves by galling fig flowers) [12].

Next to these invertebrate pollinators, three other vertebrate classes need to be considered: Mammalia, Reptilia and Aves. Mammals are mainly important in forests, where they help pollinating big trees. Some examples are rodents, flying squirrels and lemurs. However, the most famous mammalian pollinator is the fruit bat. This mostly nocturnal species has a diet containing fruit, nectar, seeds and leaves [28] and covers large distances, making him perfect for transferring pollen. Unlike many other bats, fruit bats locate their food not by echolocation but by sight and smell. Their eyes are therefore remarkably large, as can be seen in Figure 1.3a.
Examples of reptile pollinators are lizards and geckos. In desert regions where water is sparse, lizards (e.g. *Podarcis lilfordi*) can visit some cacti species to drink the nectar and are hence able to pollinate them. Lizards normally feed on insects, but feeding on nectar is an element of a variable set of feeding strategies found by *P. lilfordi*. (Other strategies include the consumption of seeds, fruits, and other parts of several plant species, as well as small crustaceans [44]).
Birds present in the pollination network are the sunbirds (family of the Nectariniidae) living in Africa, hummingbirds living in America and honey eaters living in Australia. These three groups are distantly related to each other. In some regions these birds co-evolved with plant species, making them more successful in nectar foraging than the insects competing with them. The co-evolution results in a morphological compatibility (see Figure 1.3b) as well as in a more reddish colour of the flowers (as birds and insects are more sensitive to different reflected wavelengths).

(a) Samoan flying fox (*Pteropus samoensis*).

(b) Purple-throated Carib hummingbird (*Eulampis jugularis*).

Figure 1.3: Pollinators of the Mammalia and Aves class.

1.5 Phylogeny and traits

To model a pollination network, information about the participating species can be highly useful. This information will then be stored in the feature vector of each node, as mentioned in Section 1.1. One can focus on species phylogeny, traits or a combination of both. The phylogeny and traits of the species comprised in the used pollination network are more thoroughly discussed in Chapter 3, but some examples of their relevance are shown here.

1.5.1 Phylogeny of species

Phylogeny focuses on the evolution of species and on how closely they are related to each other. It shows the relationships between groups of organisms and tries to recreate their evolution by means of common ancestors. Classification of species (e.g for angiosperms) is now strongly based on phylogenetic insights. This change was made when phylogenetic trees became more developed. Major clades were identified but those relationships were in conflict with the then prevailing classification (which was mostly based on visual similarities) [1].

Rafferty and Ives (2008) used Phylogenetic Linear Mixed Models (PLMM) for the statistical assessment of two statements: (1) whether closely related pollinators are more likely to visit plants with similar relative frequencies, and (2) whether closely related pollinators tend to visit closely related plants. As can be assumed, the models treated the quantitative strengths of pairwise interactions as the dependent variable, and incorporated phylogenies as anticipated covariances among these interactions, as independent variables.
The conclusion of the researchers was that pollinator phylogeny did not explain the

community composition. Closely related pollinators were not more likely to visit the same plant species, and the same pollinator is not more likely to be attracted to closely related plants (research question 2). Nonetheless, pollinators were affected by plant phylogeny, namely that closely related plants were likely to have similar visitation frequencies, regardless of the species ($p < 0.001$) (research question 1) [49].

This last statement is confirmed by the study of Vazquez et al. (2009). There they showed that the number of pollinators visiting different plants depended strongly on the phylogeny of the plants but only weakly on the phylogeny of the pollinators [63].

1.5.2 Traits of species

Traits denote are all characteristics of a certain species. They can be morphological, geographical, behavioural, etc. Sometimes phylogeny is also seen as a trait. Afterwards, one can discover which plant traits are likely to be responsible for attracting different pollinators.

Examples of plant traits

Rafferty and Ives (2008) investigated eight plant traits to determine which ones could explain the visitation behaviour of pollinators on different plant species. Of the eight traits, two traits involved phenology[3]: (1) phenological shift (i.e. whether plants are flowering significantly earlier) and (2) date of first bloom (i.e. the mean week of flowering onset) and six were morphological traits: (3) plant height, (4) flower color, (5) floral symmetry, (6) floral display size (i.e. mean number of flowers or iflorescences per plant), (7) nectar volume, and (8) nectar concealment (i.e. whether flowers have concealed nectar or not). To facilitate comparisons among the effects of plant traits, they standardized values for each trait to have a mean of 0 and a variance of 1. After the validation of their linear mixed model, four of the eight traits seemed to be significant, being date of first bloom ($p=0.047$), plant height ($p=0.002$), flower color ($p=0.009$) and floral symmetry ($p=0.046$) [49].

Simpson and Neff (1983) investigated floral morphology, floral colour, scent and reward chemistry as traits. They found (in contrast to the ones above) that blossom colour scored poorly in predicting pollinator visits while reward chemistry seemed more important. The amount and availability of reward may strongly limit the functional groups of attracted pollinators. The most common rewards are pollen and nectar. Pollen is a reward offered by e.g. plants with poricidally dehiscent[4] anthers, but is not available for all pollen-feeding insects. Exclusively bees that can vibrate their

[3] Phenology is the study of periodic plant and animal life cycle events and how these are influenced by seasonal and interannual variations in climate.

[4] Dehiscence is the splitting along a built-in line of weakness in a plant structure in order to release its contents at maturity.

flight muscles to buzz the flowers can collect this pollen, while other insects lack this access. Other less common types of rewards are oil, fragrance (that male euglossine bees collect to attract females), floral resin (that can be used for nest building), and a site for breeding, as mentioned in the very specialized fig/figwasp example of Section 1.4.

The apparent contradiction about blossom colour can easily be explained, as different floral traits could be associated with specialization at other taxonomic scales. Traits like floral colour are more important at a higher level (e.g. differentiation between bee and bird pollinators), while reward is more important at a lower taxonomic level (e.g. differentiation between different bee families/genera) [22].

Pollinator traits

Leigle et al. (2017) used a long list of pollinator traits to develop a recommendation model. Examples are long or short legs, body mass, feeding habits, living above or below ground, etc. Afterwards they could conclude that only three traits were necessary to create a good recommendation: the body mass of the species and two other traits based on phylogeny [16].

The importance of traits is illustrated well in Santamaria and Rodríguez-Gironés (2007). These people did research about whether plant-pollination networks can be described using trait complementarity rules and barrier rules. They used different models and different data to check if topological properties (like nestedness, degree distribution, etc.) of mutualistic networks can be predicted.
One of the good scoring models was the lognormal neutral model. In this one the probability that a plant-pollinator pair interacts is proportional to their relative abundances, and relative abundances were drawn from a lognormal probability distribution. However, they listed several reasons why this cannot be considered as a valid prediction model, despite the good fit. Firstly, assuming random interactions is not sufficient to reproduce network topology, as neutral models based on other propability distributions (e.g. uniform ones) give poor fits to the data. Secondly, the neutral model assumes that species abundance determines the frequency of interactions, but there is no proof that these two are significantly correlated to each other. Also an interpretation error occurs because it is not clear whether generalistic species are generalists because they are more abundant, or if they are more abundant because they interact generalistic and hence have more access to resources. Thirdly, considering random interactions implies that most (all) phenotypic characteristics of interacting species are irrelevant to determine patterns in the data. Other studies showed that phenotypic traits often prevent the happening of a certain interaction. Due to this so-called forbidden interactions, and due to previously mentioned lack of causal interpretations, random behaviour can be rejected.

What they did discover is that the combination of trait complementarity rules and barrier rules provided very good fits and good predictions for the network metrics. They included traits like phenology (flowering period of plants, flying period of insects), the nectar sugar concentration, the flower color and scent. They discovered that simple linkage rules lead to mutualistic networks with the same topological properties as observed in actual datasets. Beside the challenge of understanding the ecological processes that underlie these regularities, they showed that two to four linked traits were enough to predict topologic properties [53]. These are important conclusions and can be used as basis for later discussed prediction models (Chapter 4).

1.6 Climate change for plants/pollinators

Climate change is a hot topic in media and literature, and deserves some special attention in the context of plant-pollinator interactions. Many ecosystems in nature are affected by regional and global climate changes, and pollination networks are one of them. All kinds of effects can disturb ecological interactions (e.g. rainfall) but the main factor seems to be the temperature increase. Both plants and pollinators are affected by global warming, although the generally shorter life span of insects makes them more sensitive to this variability [30].

Mutualistic partners have synchronized their timing over the evolution. This development of increasingly narrower phenological matches is one of the aspects of co-specialization [39]. Now however, more mismatches are observed or predicted. Data has shown that many plants have reacted to increasing temperatures by flowering earlier. The flowering period appears to start earlier in the season, linearly correlated with the mean (increasing) temperature of the month of/months before flowering, and species flowering early in the season appear to be most sensitive. However, this is not true for all plant species as other data collections show that e.g. 20% of the species was not affected. Besides the (whether or not) earlier flowering, the length of the flowering season seemed less affected.

Pollinators show similar behaviour. In studies where butterflies were observed, a close relationship between first appearance dates and temperature was detected. The peak appearance came earlier *and* flight duration was prolonged during a warming period. The same goes for honeybees (e.g. *Apis mellifera*) where there is also a quite linear relationship between the date of first appearance and the temperature in the previous months. Honey bees can be considered as good indicators of climate change as they overwinter, and appear to react quickly to increases in spring temperatures.

Although the responses show the same trends, it cannot be avoided that sometimes temporal mismatches appear among the mutualistic partners: not all linear relationships are evenly sensitive; other environmental cues for plant flower inducing can be altered, leading to unexpected cue combinations and bizarre flowering times; early mismatches in the season can result in restricted nest development of bee species and can limit pollination services later in the season; etc.
Modelling outcomes infer that between 17 and 50% of all pollinator species suffer from disturbance in food supply due to temporal mismatches, depending on the phenological shifts applied. However, these values have to be treated with care as direct temperature responses and the occurrence of mismatches in pollination interactions may vary among regions.

Here a second mismatch can be introduced, namely a special mismatch. Not only temporal behaviour changes can be induced by global warming, but also changes in abundance and distribution. Both increased reproductive effort of plants as decreased flower abundance have been observed for increasing spring temperatures. Pollinators are then again influenced by both this global warming, and the changes in food availability by changes in plant population density. In history it has been observed that species distributions move towards equator regions and descend from mountains during cold glaciations, while the opposite is observed during warmer inter-glaciations. A lot of trees and pollinator species (like butterflies and bees) now show the dynamics of moving towards higher latitudes and altitudes, similar to what is expected in warming scenarios. Flies on the other hand show opposite patterns, possibly due to new dominant and competing species.

Again due to this sometimes similar but perhaps not equal effects, or sometimes opposite effects, both phenological (temporal) and spatial mismatches may occur. This has several consequences. Specialized pollinators are most likely to be left with no food due to new competing insects and are most sensitive to extension, but also generalist species could be pushed to diet shifts or diet reductions. Also, if some pollination interactions are uncoupled, the network has to establish new interactions. Considering long co-evolution of lots of species, this is not evident. The last frequently studies effect is anew quite contradictive. Some plants suffer from reduced pollen deposition through quantitatively less or qualitatively less efficient visits. Plants may suffer from limited reproduction due to insufficient pollination. On the other hand also opposite effects are observed, where supplemental pollination leads to a positive influence on the survival and growth rate of flowering species. An increased food availability per flower could maybe partly compensate the diet reductions due to mismatches in time and space.

The latter can be seen as a buffer mechanism, and some similar statements can be made. Naturally, ecological networks possess an inherent robustness in their structure. Mutualistic networks consist of highly asymmetric relationships, where a core of generalist species interact with each other, while most specialists interact only with these generalists. This was the nested structure of a plant-pollination network. The stability induced by nestedness makes networks more robust against perturbations caused by climate change. Although the buffering capacity, the loss of generalistic plant species in particular, may put other plants and pollinators at higher risk for extinction. And even when a dynamic structure of links tries to compensate for these occurrences, each mutualistic network can reach a tipping point and collapse under a disturbance [30].

CHAPTER 2

MODELLING TECHNIQUES

2.1 Machine learning and pairwise learning

Literally describing and implementing all (possible) driving forces for interactions in a network would be hard and time consuming. These interactions are dependent on lots of biotic and abiotic factors, on each other and probably on a lot of phenomena that are not completely understood yet. Here machine learning could be of use. Also secondly, models providing *those* interactions with a great chance of occurring in reality could shorten and optimize field searches. In other words, predictive models enable researchers to prioritize interactions for experimental validation [62].

Machine learning can be described as the science of finding stable patterns in data [57]. In fact, it is the field of computer science that gives computers the ability to learn without being explicitly programmed [52]. It provides algorithms based on training examples out of which knowledge can be discovered [47]. A small example (of the supervised learning type) to clarify the definition could be: assume there is a dataset of face images available and one wants to make a classifier that provides for every picture a label $y \in \{male, female\}$. This application would be hard to program in a traditional way since formally specifying a rule that differentiates male from female is not evident. An alternative is to give example pictures labeled with their gender, and let a machine automatically learn a rule. When this is done, the developed model can be used to provide labels for other pictures that were not included in the training dataset [6]. For biological networks, supervised learning approaches perform typically much better than unsupervised ones, as they take advantage of known interactions of the network and create a model based on their specific properties [56].

A specific direction in machine learning is the field of pairwise learning. In pairwise learning, the goal is to predict the label of a pair of objects (instead of a label for just one object, like the picture example above) where this pair of objects is called a dyad. A differentiation can be made between monadic and dyadic data, depending on the two objects comprised in the dyad. When for example an interaction metric between two pollinators should be predicted, the pair of objects can be denoted as

(u, u') with $u \in \mathcal{U}$, $u' \in \mathcal{U}$ ($\mathcal{U}$ is the set of all pollinators) and a label y. This is to be linked with the unipartite network of Section 1.1. When the interaction strength between a pollinator and a plant is of interest, the corresponding dyads are of the form (u, v) with $u \in \mathcal{U}$, $v \in \mathcal{V}$ ($\mathcal{U}$ is the set of all pollinators and $\mathcal{V}$ is the set of all plants), and again label y. This corresponds to the bipartite network of Section 1.1. y referred to as an interaction strength two times, but can in fact stand for any (binary, quantitative or textual) label [58].

The goal of pairwise learning is to learn a function $f(u, v)$ to make predictions of this label for new dyads. The prediction can be an estimate of the binary, quantitative or textual label, but can in some cases also be interpreted as a score indicating the confidence of the interaction occurring [59]. However, the term *new dyad* is not always evident. Therefore, four different settings are distinguished.

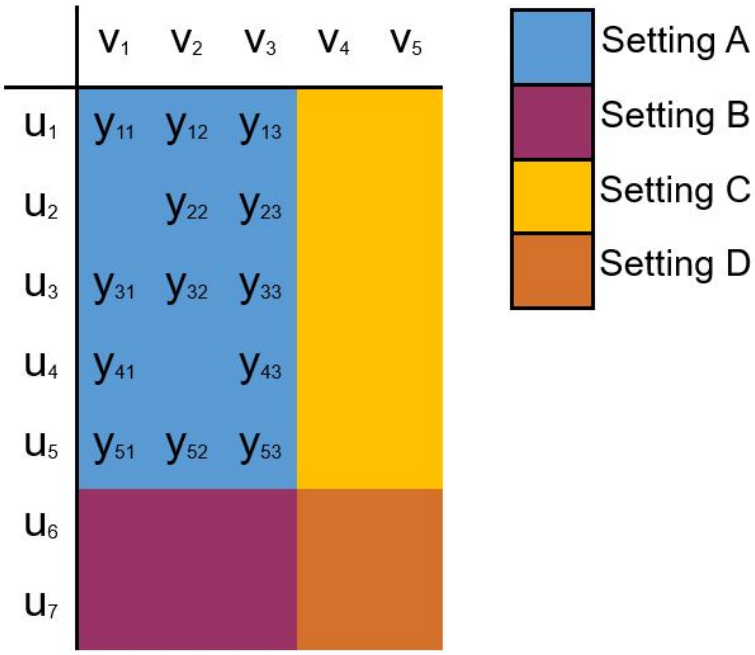


Figure 2.1: Four settings in new dyad prediction.

- Setting A (blue): both u and v were observed during training, but the label of this combination was missing.
- Setting B (purple): only v was observed during training; u is a new object that did not occur in any dyad of the training dataset.
- Setting C (yellow): only u was observed during training; v is a new object that did not occur in any dyad of the training dataset.
- Setting D (orange): both u and v are new objects that did not occur in any dyad of the training dataset.

Intuitively, it is easy to understand that the prediction of a label in setting A is much easier than a label in setting D. Therefore, it is advised to also separate validation tests or performance metrics over four settings, because otherwise over- or under-estimation of the model performance can occur. Frequently used validation methods are leave-out cross-validation schemes, but these are further discussed in Section 2.3 [58].

2.2 Techniques to predict plant pollinator interactions

As mentioned in Section 1.5.2, neutral trophic models take the probability of two species interacting is proportional to the product of their relative abundances [11]. Obviously this is a too simple approach to estimate the interaction network. Two co-occurring species do not necessary have to interact with each other and several other reasons rejected the use of these probability distributions [53]. Other manners can be used to obtain predictions for the interactions of species. In this section, two approaches will follow. The two are distinguished form each other by the use of information. The first one only uses the provided interaction matrix (see Chapter 3) while the second one includes much more data. It would be expected that the performance rises as more information is taken into account, but this is to be tested.

2.2.1 Collaborative filtering

When no additional information is known about the species of the training data (like the earlier mentioned phylogeny and traits), only the structure of the given dataset itself can be used to predict missing and new values, or to re-evaluate the given values.

Introduction to collaborative filtering (CF)

CF is a technique complementary to content-based filtering. Content-based filtering uses features of plants and pollinators (or e.g. features of items and customers in a recommendation system for online sites) to make predictions. Collaborative filtering, on the other hand, only uses the known preferences (so the binary interaction matrix or a qualitative item scoring matrix) to predict other interactions.

CF has several challenges to deal with. First of all, the data sparsity. Just as in commercial recommender systems, the interaction data of biological networks is very sparse, which complicates the making of predictions. Two problems are related to this sparsity of data. The most challenging one is the cold start problem. This occurs when

a new pollinator or plant enters the system, but as the features of this species are not taken into account, it is difficult to find similar species and predict interactions. The other problem with sparse databases is the so-called neighbour transitivity. In e-commerce recommendation, this refers to the problem in which users with similar tastes may not be identified as such if they have not rated the same items. All users only rate a small subset of the items and if this subsets do not overlap, possible similar users cannot be used for each others recommendations. In pollination datasets mostly, all possible plant-pollinator combinations are evaluated with an interaction value so this problem does not really apply. A possible solution for these sparse data are dimensionality reduction techniques such as Singular Value Decomposition [65] or Latent Semantic Indexing [32] in which unrepresentative or insignificant items are deleted and further predictions are done with the reduced dataset. Another solution would be a hybrid model in which content-based filtering and collaborative filtering are combined. As such, for example the bulk taxonomic information can be incorporated in the model and the cold start problem gives less issues. A second challenge for CF is the scalability. Both user-item databases as plant-pollinator databases can be very big which requires good computational resources. Also here, dimensionally reduction techniques can provide good solutions. The third challenge mentioned in Su and Khoshgoftaar (2009) is synonymy, referring to the fact that a number of the same (or very similar) items can be stored in the database with similar names but not as the same item, meaning they are considered as two completely different things. Here Semantic Indexing can help to reduce these items, by creating semantic spaces in which all the items would be closely related, e.g. the words 'film' and 'movie'. In biological contexts, one species can also possess different (official) names (as *Apis mefifera* and *Apis mellifica* (the European honeybee) or *Epilobium angustifolium* and *Chamerion angustifolium* (the great willowherb)), but still this is more limited than in sales applications. Finally, there is the grey sheep phenomenon. Here one has a specific user or pollinator whose preferences do not consistently agree or disagree with the rest of the group, so this one cannot benefit from the collaborative filtering. Next to these problems, a few other challenges are mentioned in Su and Khoshgoftaar (2009) (such as privacy limitations or shilling attacks from businesses rating their own products on e-sale sites), but these last ones do not apply in a biological context [60].

A specific CF-technique: a linear filter

This filter technique re-evaluates the binary interaction matrix, meaning that all (known) interactions are given a quantitative score of how likely they are to be positive. This [0,1]-score range then replaces the original zeros (negative) and ones (positive). The score is generated by a linear filter, which will be theoretically explained below, but first a small example is shown:

	Plant1	Plant2	Plant3	Plant4	Plant5	Plant6
Pollinator1	1	0	1	1	0	0
Pollinator2	0	0	0	1	0	0
Pollinator3	1	1	1	0	1	1
Pollinator4	1	0	0	1	0	0

Pollinator 2 is a very specific pollinator as it only interacts with plant 4 (specialist). When then e.g. the interaction [pollinator 2, plant 2] is evaluated, this zero is probably correct. Hence, the new score given by the filter (instead of 0) is 0.34, meaning the interaction is more likely to be negative. Pollinator 3 on the other hand is highly non-specific, interacting with almost every plant (generalist). Also, plant 4 is highly non-specific, interacting with almost all pollinators. Now when re-estimating the negative interaction value of [pollinator 3, plant 4], the generated score is 0.81 instead of the original 0. Being close to 1, this interaction is indeed very likely to be positive and can (depending on the threshold) be defined as a false negative of the dataset. The values should however not be confused with a probabilistic interpretation; there is no chance of 34 or 81% that an interaction happens. In addition, the distribution of the filtered values can be very tilted towards zero when the interaction matrix is sparse. Only the relative comparison between the generated values has a valid interpretation.

The few non-specific species could already be distinguished in the interaction graphs. As biological networks are typically non-random, they show a heavy-tailed distribution of node degrees. Several nodes, called hubs, have degrees greatly higher than the average. In such networks, a new node (without consideration of its features) is more likely to interact with a hub than with a less connected node [56]. This is similar to what was described in the theory of nestedness. In summary, negative interactions with a hub are more likely to be wrong. The reason why one would focus on false negatives is the way a biological network is constructed. This is mostly done by aggregating positive (i.e. observed) interactions. However, there is not always evidence that a negative interaction does really not occur in reality. There is a chance that the interaction is in fact positive, but was not observed during the build-up of the network [62].

Now the construction of the linear filter is explained. The algorithm comes from the work of Stock (2017) [58]. The original binary $(n \times m)$-interaction matrix is referred to as $Y = [Y_{ij}]$. The filtered interaction matrix will be called $F = [F_{ij}]$ and the cross-validating values will be called β. First, the values of F are constructed as a weighted average of the interaction value itself, the average of the interactions in its row, the average of the interactions in its column and the average of all the interactions in the

matrix:

$$F_{ij} = \alpha_1 Y_{ij} + \alpha_2 \frac{1}{n} \sum_{k=1}^{n} Y_{kj} + \alpha_3 \frac{1}{m} \sum_{l=1}^{m} Y_{il} + \alpha_4 \frac{1}{nm} \sum_{k=1}^{n} \sum_{l=1}^{m} Y_{kl}. \tag{2.1}$$

The use of this row and column information is intuitively logical considering the previously given example. The four α's are weights between zero and one, obviously summing up to one. The method works well for all α's being 0.25, but further tuning can be done.

If one now wants to validate every interaction Y_{ij} in the original matrix (i.e. generate the interaction value β), it would be ideal *not* to make use of the specific interaction of pollinator i and plant j and only use the other information comprised in the network. Only this way a proper estimation of the model performance can be achieved. β is therefore constructed based on the condition that when this value passes through the linear filter F, it remains unchanged. Then the previous expression is changed to:

$$F_{ij} = \alpha_1 Y_{ij} + \left[\alpha_2 \frac{1}{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} Y_{kj} + \alpha_2 \frac{1}{n} Y_{ij} \right] + \left[\alpha_3 \frac{1}{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} Y_{il} + \alpha_3 \frac{1}{m} Y_{ij} \right] + \left[\alpha_4 \frac{1}{nm} \sum_{\substack{k=1 \\ k \neq i}}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{m} Y_{kl} + \alpha_4 \frac{1}{nm} Y_{ij} \right]. \tag{2.2}$$

When both Y_{ij} and F_{ij} are first changed to β, one gets

$$\beta = \frac{\alpha_2}{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} Y_{kj} + \frac{\alpha_3}{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} Y_{il} + \frac{\alpha_4}{nm} \sum_{\substack{k=1 \\ k \neq i}}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{m} Y_{kl} + \left(\alpha_1 + \frac{\alpha_2}{n} + \frac{\alpha_3}{m} + \frac{\alpha_4}{nm} \right) \beta, \tag{2.3}$$

in which this structure can be recognised

$$\beta = \left[F_{ij} - \left(\alpha_1 + \frac{\alpha_2}{n} + \frac{\alpha_3}{m} + \frac{\alpha_4}{nm} \right) Y_{ij} \right] + \left(\alpha_1 + \frac{\alpha_2}{n} + \frac{\alpha_3}{m} + \frac{\alpha_4}{nm} \right) \beta. \tag{2.4}$$

Solving this equation to β leaves us with

$$\beta = \frac{F_{ij} - \left(\alpha_1 + \frac{\alpha_2}{n} + \frac{\alpha_3}{m} + \frac{\alpha_4}{nm} \right) Y_{ij}}{1 - \left(\alpha_1 + \frac{\alpha_2}{n} + \frac{\alpha_3}{m} + \frac{\alpha_4}{nm} \right)}. \tag{2.5}$$

Although Y_{ij} appears in Equation (2.5), β does *not* depend on this value. The setup for this formula makes sure that all dependencies clear each other out. Hence, this is an application of the well-known leave-one-out cross-validation (LOOCV). CV will follow in Section 2.3 as well.

The predicted interaction value β is replacing the original zeros and ones. In order to not have to define a threshold for β from which we define an $Y_{ij} = 0$ as being a false negative, ROC curves (Receiver Operating Characteristics) and corresponding AUC's (Area Under the Curve) are used. The meaning of this ROC curve and its AUC, as also their application on the dataset are presented in Sections 2.3 and 4.1, respectively. Validation experiments in Stock (2017) with 94 datasets with the LOO-computation gave an average AUC of about 80% for binary matrices, meaning that on average there is a chance of 80% that a missing positive interaction gets a higher score than a missing negative interaction [58].

In summary, negative interactions with high scores are natural targets for increased sampling effort, as they are most likely to occur in reality.

2.2.2 Kernel methods

Secondly, two-step kernel ridge regression (TSKRR) is discussed. This method will use the interaction matrix, as well as the traits/phylogeny of both species sets to make predictions. Predictions are again numerical values that (try to) approach the binary interaction matrix. Firstly, a short introduction to the different parts of TSKRR is provided.

Kernels

Kernel functions are mathematical tools to represent and manipulate objects in artificial high-dimensional feature spaces. By using kernels, linear models can be used for non-linear problems. It is assumed that there is a feature map $\phi : X \to H$, where H (the Hilbert space) is a suitable space to represent these objects. This Hilbert space H extends the methods from the two-dimensional Euclidean plane (or three-dimensional space) to spaces with any (in)finite number of dimensions. Then, the general idea of kernels is that in this high-dimensional space H, a simple linear model might suffice to describe the patterns in the data, instead of having to compute a very complex model in space X.

The feature mapping itself is in practice never really done. Computing the map is computationally expensive as H is an infinite-dimensional space, making the calculations hard or nearly impossible. A Hilbert space is only an abstract vector space, and it has the structure of an inner product. The inner product (or dot product) is used to apply the *kernel trick*. This means that the value of the kernel function of a pair of objects in the original object space is the same as the inner product of the items of the pair presented in the Hilbert space: $k(x, x') = \langle \phi(x), \phi(x') \rangle_H$, where $\langle a, b \rangle$ is used to indicate a dot product. By using this kernel trick, algebraic operations can be performed in space H without performing the feature mapping [55].

A common use of a given kernel function is in the form of a matrix, containing the kernel values of all possible pairs of objects. This is called a Gram matrix: $K = [K_{ij}] = [k(x_i, x_j)]$. Such kernel or Gram matrices are always symmetric (i.e. $k(x, x') = k(x', x)$) and positive semi-definite. Kernels quantify the similarity between objects.

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \vdots & \vdots & & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix} \tag{2.6}$$

Kernels can easily be used in a pairwise learning setting by defining so-called pairwise kernels $\Gamma((u, v), (\bar{u}, \bar{v}))$, where e.g. $\Gamma((u, v), (\bar{u}, \bar{v})) = k(u, \bar{u})\, g(v, \bar{v})$. These measure the similarity between two dyads (u, v) and $(\bar{u}, \bar{v})$ instead of between two separate objects [58].

Ridge regression

Ridge regression is a modification of the well-known linear regression, where not only the least-square error is used to estimate the coefficients of the regression, but a second penalty term is included. This penalty term consists of the L2-norm with a tuning parameter λ. The tuning parameter λ controls the relative impact of the penalty term on the estimates of the regression coefficients. When $\lambda = 0$, the penalty term has no effect and ridge regression will produce the least-square estimates. The profit of this extra term is that it shrinks the weights of less contributing variables (feature selection). As λ increases, the flexibility of the ridge regression fit decreases (and hence prevents unnecessary complexity/overfitting), but increases the bias [34]. Ridge regression can be formulated as:

$$f(x_i) \;=\; w_0 + w_1 x_{i1} + \ldots + w_p x_{ip} + \epsilon_i \;=\; \sum_{j=0}^{p} w_j x_{ij} + \epsilon_i$$

$$\hat{w} \;=\; \min_{w} \sum_{i=1}^{n} \Big(y_i - \sum_{j=0}^{p} w_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} w_j^2 \quad \Big(\text{i.e. } RSS + \lambda \sum_{j=1}^{p} w_j^2 \Big) \tag{2.7}$$

Kernel ridge regression

When combining the 2 concepts above, one can replace all feature vectors in the ridge regression expression with their mapping in H: $x_i \rightarrow \phi_i = \phi(x_i)$. In this case the number of dimensions can be much higher, or even infinitely higher than the number of data-cases. A long expression for the weights w can again be obtained, but we will focus on these expressions in the following subsection [67]. Actually, it boils down to finding that prediction function f (to fit a model in an 'imaginary' high-dimensional

feature space H) that minimises a similar twofold loss function.

$$\min_f \sum_{i=1}^{n} \Big(f(x_i) - y_i \Big)^2 + \lambda \, ||f||_H^2 \tag{2.8}$$

The second part is again the L2-norm, here in the Hilbert space, with weight λ [58].

Two-step kernel ridge regression

Two-step kernel ridge regression is conceptually quite straight-forward. Two ordinary kernel ridge regressions are combined: one for generalizing to new pollinators and one for generalizing to new plants (see Figure 2.2). In this way, a prediction for *new dyads* can be made. TSKRR can be used for any of the four discussed settings.

In a first step, a prediction is made for known plants and new pollinators. Next, a second KRR is used to make predictions for new plants, using the predicted labels from the first model. The order of the two regressions is purely arbitrary; it does not matter if the first step uses a model for new plants and the second step one for new pollinators, or the other way around [58].

The fact that two successive kernel ridge regressions are performed, implies that two separate kernel matrices can be used instead of having to work with similarities between dyads. We hence need a kernel matrix $k(u, \bar{u}) : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ containing the similarities between the pollinators and a kernel matrix $g(v, \bar{v}) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ containing the similarities between the plants, without having to include pairwise kernels like $\Gamma((u, v), (\bar{u}, \bar{v}))$.

A short mathematical outline of TSKRR is provided. In this context, n pollinators u, m plants v and the $(n \times m)$ interaction matrix Y are available. The model for any dyad (u, v) that has to be learned is of the form

$$f(u, v) = \sum_{i=1}^{n} \sum_{j=1}^{m} W_{ij} \, k(u, u_i) \, g(v, v_j) \,, \tag{2.9}$$

with k the pollinator kernel, g the plant kernel and W the $(n \times m)$ matrix of weights. These are the model parameters (weights) that have to be estimated.
When shifting to the notation of Gram matrices ($K = [k(u_i, u_j)]$ and $G = [g(u_i, u_j)]$), the parameters of the TSKRR can be obtained by

$$W = (K + \lambda_u 1_n)^{-1} \, Y \, (G + \lambda_v 1_m)^{-1} \,, \tag{2.10}$$

where 1_n and 1_m respectively are the $(n \times n)$ and $(m \times m)$ identity matrices and where λ_u and λ_v are two regularization parameters.

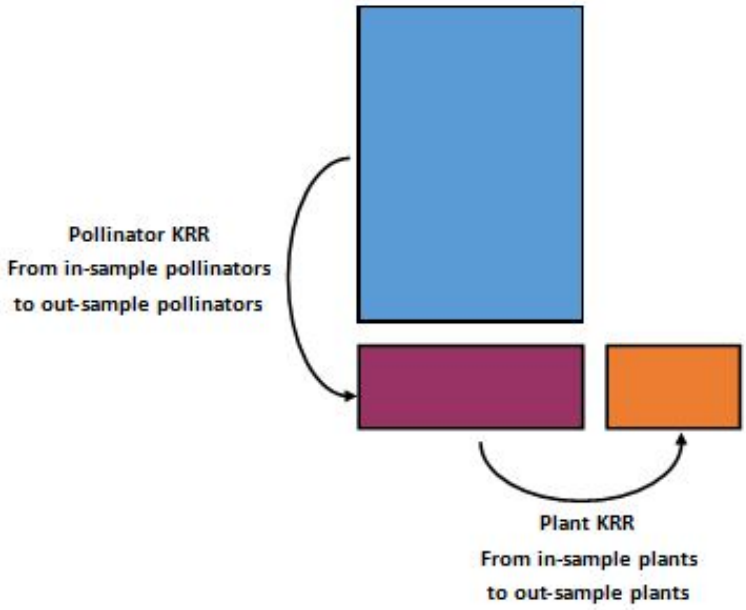


Figure 2.2: Illustration of the TSKRR principle.

The $(n \times m)$ matrix F with the model's predictions can be obtained as

$$F = K\,(K + \lambda_u 1_n)^{-1}\; Y\; (G + \lambda_v 1_m)^{-1}\, G \tag{2.11}$$

or as $F = H^k\, Y\, H^g$, when the form of hat matrices is introduced:

$$H^k = K\,(K + \lambda_u 1_n)^{-1} \qquad \text{and} \qquad H^g = (G + \lambda_v 1_m)^{-1}\, G\,. \tag{2.12}$$

The TSKRR will be used in Chapter 4, where hence the interaction matrix Y and kernel functions will be used as input.

2.3 Performance evaluation

Models are only effective when their functioning is adequate. Several options to estimate a model's performance are available and some of them will be highlighted here.

Performance metrics

Common criteria to evaluate binary predictions are the accuracy (i.e. the number of correctly predicted pairs divided by the total number of pairs) or, equivalently, the error rate (i.e. one minus the accuracy). However, ecological networks typically deal with highly imbalanced data as non-interacting pairs often far outnumber interacting ones. Accuracy is not appropriate in such situations because it greatly favors the majority class (a simple model just predicting all pairs as non-interacting would receive a high accuracy, although this obviously is not a good model). Alternative measures

are based on a confusion matrix. In the case of binary classification, this matrix is a (2×2) matrix where the columns and rows represent respectively the actual and the predicted classes (positive and negative), with each cell containing the number of pairs corresponding to these classes [56]. Note again that in a classification context, an interaction (1) is defined as the positive class and a non-interaction (0) as the negative. This is not to be confused with the biological meaning of positive and negative effects on the interacting species [58]. We assume that all pollination networks are of the mutualism type.

	Actual positive (P)	Actual negative (N)
Predicted positive (predP)	True positive (TP)	False positive (FP)
Predicted negative (predN)	False negative (FN)	True negative (TN)

Table 2.1: Confusion matrix for binary predictions.

Several metrics can be derived from this matrix to evaluate the performance of a model, among which:

- The true positive rate (TPR), also called the sensitivity or the recall, is equal to the number of true positives divided by the number of actual positives: $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$.
- The true negative rate (TNR), also called the specificity, is equal to the number of true negatives divided by the number of actual negatives: $\frac{\mathrm{TN}}{\mathrm{FP+TN}}$.
- The false positive rate (FPR), corresponding to 1-specificity, is equal to the number of false positives divided by the number of actual negatives: $\frac{\mathrm{FP}}{\mathrm{FP+TN}}$. In many biological networks however, the number of interactions is much lower than the number of non-interactions. It is therefore important to achieve a low FPR because even moderate FPR can easily lead to much more false positive predictions than true positive predictions, and hence a very low precision.
- The false negative rate (FNR), also called the miss, is equal to the number of false negative divided by the number of actual negatives: $\frac{\mathrm{FN}}{\mathrm{TP+FN}}$.
- The precision is equal to the number of true positives divided by the number of predicted positives: $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$.
- The rate of positive predictions (RPP) is equal to the number of predicted positive divided by the total number of examples: $\frac{\mathrm{TP+FP}}{\mathrm{P}+N}$ or $\frac{\mathrm{predP}}{\mathrm{P+N}}$.

These measures should be combined to give a global picture of the performance of a model, e.g. sensitivity and specificity or precision and recall. Other, less used, metrics are the correlation coefficient Q^2, the F-score and the average normalized rank. Their definitions can be found in literature [56].

Graphical assessment

Beside single values, visualisations with curves are also possible. The three most known curves are the ROC curve, the PR curve and the lift chart. When the number of positive examples is much smaller than the number of negative ones, as it often happens in biological networks, there is not much difference between the ROC curve and the lift chart, so this last one is omitted.

- ROC stands for Receiver Operating Characteristics. ROC curves plot the True Positive Rate as a function of the False Positive Rate, when varying the confidence threshold. More specifically, the predictions are sorted from the most confident to the least confident and the threshold is varied from the maximum to the minimum confidence score. Each value of the threshold corresponds to a different confusion matrix and thus a different pair of values of the TPR and FPR. All these different pairs of values together construct the ROC curve. Every ROC curve goes through the points (0,0) and (1,1). The curve of a 100% perfect classifier would make a right-angle trough the point (0,1), while a completely random classifier would make a ROC curve coinciding with the diagonal (the line directly connecting the points (0,0) and (1,1)). Both can be seen in Figure 2.3. Logically, one wants a classifier with a ROC curve as close as possible to the one through (0,1). A way to quantify this goal is with the area under the ROC curve (AUROC), which is equal to 1 for a perfect classifier and 0.5 for a random one. ROC curves allow to compare classification methods and work for each possible ratio of expected positive and negative predictions. However, if one knows that the ratio between positives and negatives will be very low when applying the classification model, then one is typically only interested in the bottom-left part of the ROC curve. In fact in such cases, PR curves are more suitable to give an overview of the whole computed model.

- PR stands for Precision vs. Recall. As the name reveals, this curve plots the precision as a function of the recall (equal to the TPR), when varying the confidence threshold. A perfect classifier would give a PR curve passing through the point (1,1), while a random classifier would have an average precision equal to $\frac{P}{P+N}$, see also Figure 2.3. All PR curves end at the point $(1, \frac{P}{P+N})$ corresponding to predicting all pairs as positive. Similarly as the AUROC, also the PR curve can be summarized in one value, namely the area under the RR curve (AUPR). A drawback of PR curves is that they are much more sensitive to false negatives in the true dataset. On the other hand, they can be used this way to experiment with the fraction of false negatives [56].

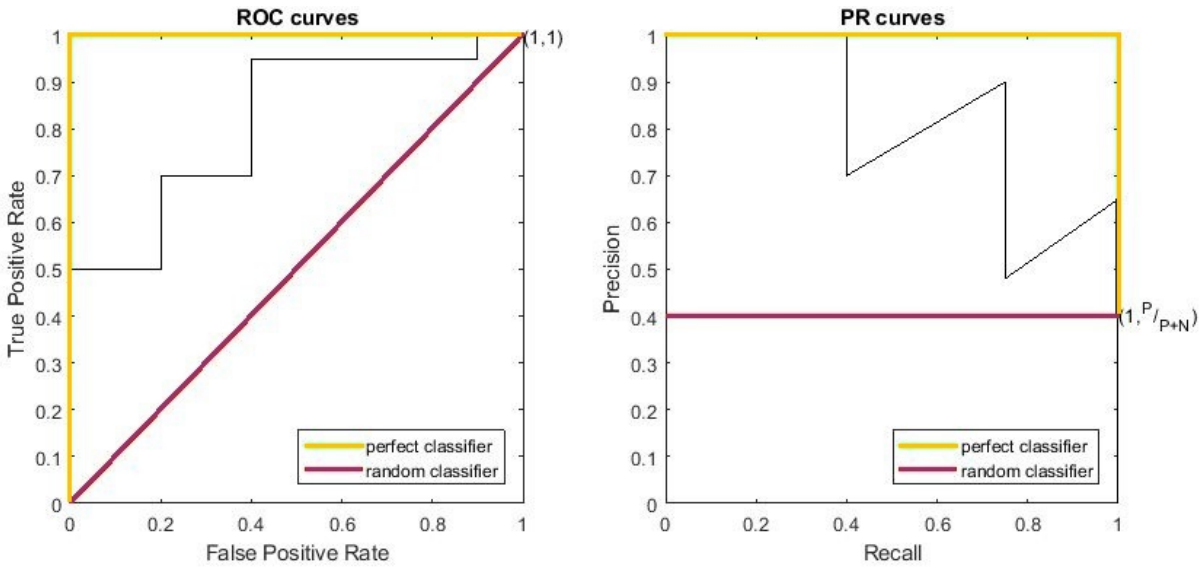


Figure 2.3: ROC and PR curves of a 100% perfect classifier, a completely random classifier and an example curve.

Cross-Validation

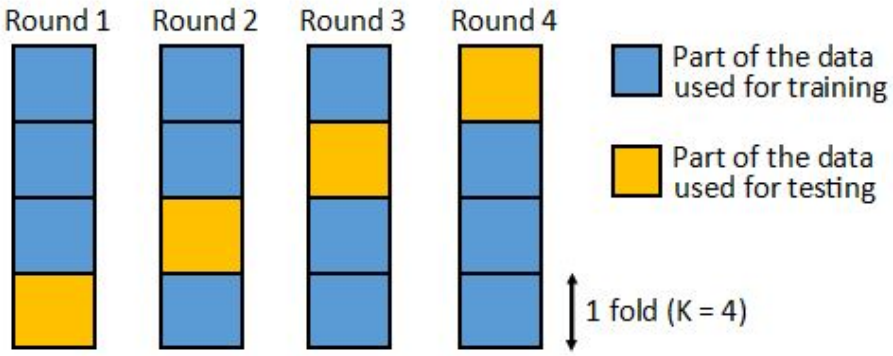
Classification methods are typically validated using cross-validation (CV). CV is a statistical evaluation method that divides the data in two segments, being a training set and a test set. The model is trained on the examples of the training set and subsequently validated with the examples of the independent test set. In typical CV, the training and validation sets must switch in successive rounds such that each data point has a chance of being validated against. This is the basic form of K-fold cross-validation [50].

Regular K-fold CV divides the items in K folds, trains the model on $K-1$ folds and evaluates the model with the data points of the remaining fold. This principle is illustrated in Figure 2.4a for four folds. However, in the setting of pairwise learning, an important difference is to be noticed. Pairwise learning models take dyads as an input instead of single items. Since the objects comprised in the dyad come from two different sets, there are two dimensions in which items can be left out of the training set. In other words, the model may have learned from either both objects, only one object or none of the objects of the test instance during training, depending on how training and test instances were selected. This results in four different K-fold CV schemes instead of the classical one-dimensional one. To start, the four *prediction* settings of Figure 2.1 in Section 2.1 are repeated, applied to the pollination network:

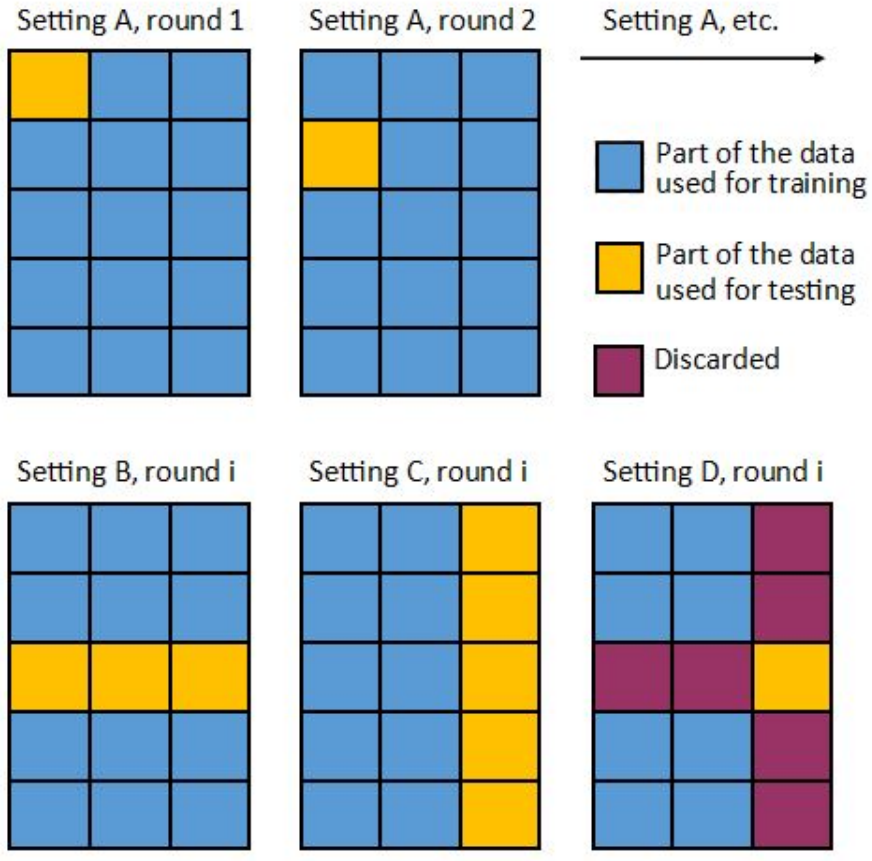
(A) The plant and the pollinator species were both part of the training data, but the value of this specific interaction was missing and hence needed to be predicted.

(B) The pollinator was part of the training dataset, but the plant is a new species.

(C) The plant was part of the training dataset, but the pollinator is a new species.



(a) Illustration of the regualar K-fold cross-validation scheme on a one-dimensional dataset, with K=4.



(b) Illustration of the four K-fold cross-validation schemes on a pairwise dataset, with K=15.

Figure 2.4: Cross-validation schemes, focusing on the difference between a one- and a two-dimensional dataset.

(D) Both plant and pollinator species were not included in the data on which the model was built.

Similarly, four CV settings can be determined. All the settings can be visualized in Figure 2.4b.

(A) CV for setting A is the most straight-forward. The dyads of the original set are divided in K folds; $K-1$ folds are used to train the model while the last one is used for validation. This means that when predicting a test dyad, this exact plant-pollinator combination is new to the model, but interaction information of both species is included in model training. Performance estimated with this cross-validation scheme is therefore the most optimistic.

(B) In this validation method, all dyads involving a particular pollinator are omitted for model training and included in the test set for performance estimation.

(C) Similarly as in CV-setting B, all dyads involving a particular plant species are omitted for model training and included in the test set for performance estimation.

(D) Here one has the intention to evaluate predictions on completely new dyads, so none of both objects may already be included in the training set. This means that not only the particular fold used for validation (containing the new dyads) but also every fold containing one of this dyad's objects needs to be discarded during training. When predicting this fold, both the plant and pollinator species are unseen by the model, making performance estimation in this CV scheme the most stringent [62]. In this way, for each of the K iterations, part of the data will never be used (i.e. not in the validation set and not in the training set). Quantified for K folds, this boils down to a fraction of $(K-1)^2/K^2$ of the original dataset that is used for training, $1/K^2$ for validation and $2K(K-1)/K^2$ remaining unused [58].

Afterwards, all K validation outcomes per setting can be summarized/combined into one performance metric. When designing a new supervised network inference method, it is recommended to communicate performances for all four possibilities separately, as a method can work well for one case but less good for another.

2.4 Optimal transport

The previous parts of Chapter 2 all focused on the prediction of interactions, and the evaluation of the models doing so. Here a second topic is approached, based on the interaction behaviour of species in a network. The practical use of the optimal transport theory is demonstrated in Chapter 5, but first a more theoretical overview is provided.

Optimal transport is a mathematical theory that makes it possible to measure distances between functions, or more general, objects, by enforcing several (e.g. mass) conservation laws. The first published optimal transport problem is referred to as *Monge's problem* (later called the *Monge-Kantorovich mass transport problem* [15]) and is about the transformation of a landscape. An original landscape a with defined relief characteristics (little hills and valleys) needed to be converted to a desired landscape b with another relief. Transporting soil from one place in the landscape to another is associated with a cost value and of course the total cost of transporting needed to be minimized, but with the condition that the total mass of all soil was conserved. Mathematically, we define X as a subset of $\mathbb{R}^2$, a and b as two functions of X, and $c(\cdot,\cdot)$ as a convex distance. Now the problem consists of finding a function T (called the transport map) from X to X that transports landscape a into landscape b, while minimising the product of the amount of transported earth $a(x)$ with the transported distance $c(x, T(x))$, or

$$\min_{T:X\to X} \int_x c(x, X(T))\, a(x)\, dx$$

$$\text{subject to the conservation} \quad \forall B \subset X : \int_{T^{-1}(B)} a(x)\, dx = \int_B b(x)\, dx\,. \tag{2.13}$$

When the transport occurs between two completely different sets (so e.g. not from landscape to landscape) the transport map T maps a set X to a set Y. Algorithms based on this idea can be used in various settings for numerous applications. One of the settings is the semi-discrete setting, where a continuous resource or substance is transported to a discrete number of items, places, etc. [37]. Applied to a pollination network, this boils down to the problem of having a distribution of plants containing nectar, and defining the optimal transportation to a finite set of insects, such that each insect receives its desired amount of nectar (visualized in Figure 2.5). When the assumption is made that each plant produces the same amount of nectar and each insect consumes the same amount of nectar, then e.g. the amount of nectar needed for one insect species is proportional to the relative abundance of this species in the total population of pollinators.

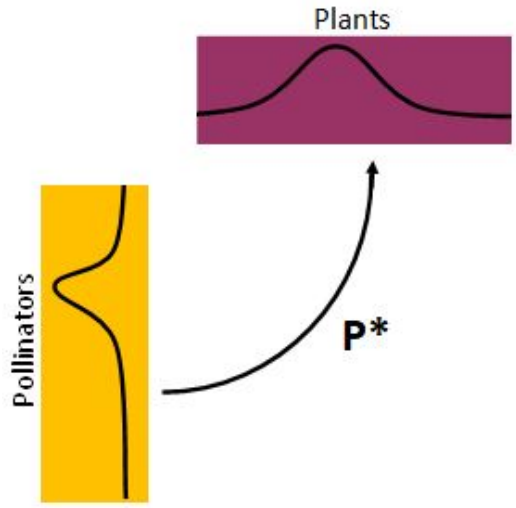


Figure 2.5: Find the optimal distribution/map to match the (yellow) pollinator distribution to the (purple) plant distribution.

Again switching to mathematical notation, this problem can be translated to:
r = vector containing the distribution of insects (n-dimensional)
c = vector containing the distribution of plants (m-dimensional)
M = a cost matrix, i.e. defining how 'unlikely' it is that a specific insect and plant interact with each other. This is in fact the negative of the pollinators' preferences.
Out of this we get a polyhedral set containing all valid partition matrices P of plants over pollinators:

$$U(r,c) = \left\{P \in \mathbb{R}^{n\times m} \mid P\,1_m = r,\ P^T\,1_n = c\right\} \tag{2.14}$$

where 1_m and 1_n are respectively an m- and n-dimensional vector of ones.

Now, the following optimization problem needs to be solved

$$d_M(r,c) = \min_{P\in U(r,c)} \sum_{i,j} P_{ij}\,M_{ij} \tag{2.15}$$

to minimise the distance (cost) and hence find the most optimal distribution P*.
A shorter notation is possible using the Frobenius dot-product, where $\sum_{i,j} P_{ij}\,M_{ij}$ is written as $\langle P, M\rangle_F$, leaving the formula as

$$d_M(r,c) = \min_{P\in U(r,c)} \langle P, M\rangle_F\,. \tag{2.16}$$

A modification can be made based on the information theory of Section 1.2. There it was stated that an interaction network can be more stable if the (conditional) entropy becomes higher (i.e. there is more uncertainty) and species are to interact with more partners. This can here be achieved by inducing (obligating) a minimal evenness in

the solution. The optimal partition matrix P^* then forces the pollinators to visit a bit of all plants instead of just focusing on one or a few favourite plant species. The solution is smoothened out following the maximum-entropy principle [14]. The regularization using the maximum-entropy principle is quite intuitive and has in fact been favoured over the one without entropy in, for example, the use of transport theory to predict traffic patterns [69]. In this paper entropy maximizing is approached as probability maximizing, which is equivalent considering the structure of P^* and the formula of entropy.

The evenness in the transport matrix P^* is quantified by a tuning parameter λ. The smaller this parameter, the more even the partition of pollinators over plants becomes. The formula of the distance is then adjusted to

$$d_M^\lambda(r,c) = \langle P^\lambda, M\rangle_F = \min_{P\in U(r,c)} \langle P, M\rangle_F - \frac{1}{\lambda}h(P) = \min_{P\in U(r,c)} \sum_{i,j} P_{ij}M_{ij} - \frac{1}{\lambda}h(P), \qquad (2.17)$$

with $h(P) = -\sum_{i,j} P_{ij}\log P_{ij}$, being the entropic regularization term. Note that this is the exact same equation as Equation (1.3) defining entropy (except that the one there is defined in bits by using $\log_2$).
Indeed, the smaller the parameter λ, the higher the weight of the entropy becomes, the more the distance is punished for an uneven distribution. d_M^λ is called the dual-Sinkhorn divergence. The solution is derived by taking the Lagrangian of Equation (2.17). Setting its partial derivative to zero gives the solution P^* of the minimization problem.

Experiments require two sets of species distributions and a cost matrix defining the cost of the interaction between every two species of the separate sets. The algorithm gives two outputs, namely the optimal transport matrix between the two distributions P^* and the corresponding Sinkhorn distance d_M^λ. The mathematical meaning of the two is already clear: the matrix is the most ideal mapping of the first distribution to the second, and the distance is the overall cost that the just defined mapping pattern would cause, but is now focused on their biological meaning.
The optimal transport matrix defines the part of its interactions every pollinator should establish with each plant (i.e. the 'portion of its visits'). This is dependent on the abundance of the plants, the abundance of the insect species itself, the abundance of the other insect species capable of pollinating the same plants, and on the cost matrix per possible interaction. P^* defines the optimal interaction behaviour of the pollinators. The distance then is the overall cost of this defined interaction behaviour. As it is more convenient to speak of a maximization of preference instead of minimization of cost, the sign of the distance can be switched and this value can hence be interpreted as a 'satisfaction index'. The interaction matrix with the lowest cost is now

the interaction matrix that matches best with the initial species' preferences. It tries to fit as good as possible the pollinators' behaviour if they were free to choose their interaction partners, but still meets the constraints set by the available distributions and the entropic restriction term.

CHAPTER 3

EXAMINING THE DATA

3.1 Quick overview

This work is entirely based a single dataset, provided by *FlorAbeilles*, a project of *Laboratoire Pollinisation et Ecologie des Abeilles de l'Unité Abeilles et Environnement de l'INRA d'Avignon* [26]. The original dataset contains 305 pollinators and 452 plant species. It is a binary interaction matrix, meaning there is only a distinction between interactions and non-interactions without providing qualitative information (a recommended choice considering Section 1.4). The matrix is very sparse, having a positive value density of only 1.10%. This implies that the network predominantly consists of specific interactions, combined with a few generalists. Multiple visualisation techniques can be performed to introduce the FlorAbeilles dataset. One of them is with the software package *Bipartite* in R, of which an obtained image is shown in Figure 3.1. To visualise the data sparsity, heatmaps can be used. The first heatmap of Figure 3.2 represents the original interaction matrix, with all species in alphabetical order. The most striking row is on about 1/4th of the figure. This pollinator is *Apis mellifera*, the famous Western or European honey bee, see Figure 3.3. *Mellifera* literally means honey-bearing, referring to the fact that this bee produces a large volume of honey as stockpile over the winter. This species can interact with a large variety of plants and is able to adapt itself to the local environments as they spread geographically [68].

The second heatmap of Figure 3.2 is obtained by swapping rows and columns respectively, in such way to indicate the nestedness of the network. Nestedness and modularity were defined and visualized in Section 1.2. Referring back to these plots, the nested structure of the network is quite clear. By using Equation (1.1), we obtain an $\eta^{(R)}$ of 0.0718, $\eta^{(C)}$ of 0.1312 and hence an overall η of 0.1015. Next to this value, other calculations based on Section 1.2 *(Information theory)* can be made:

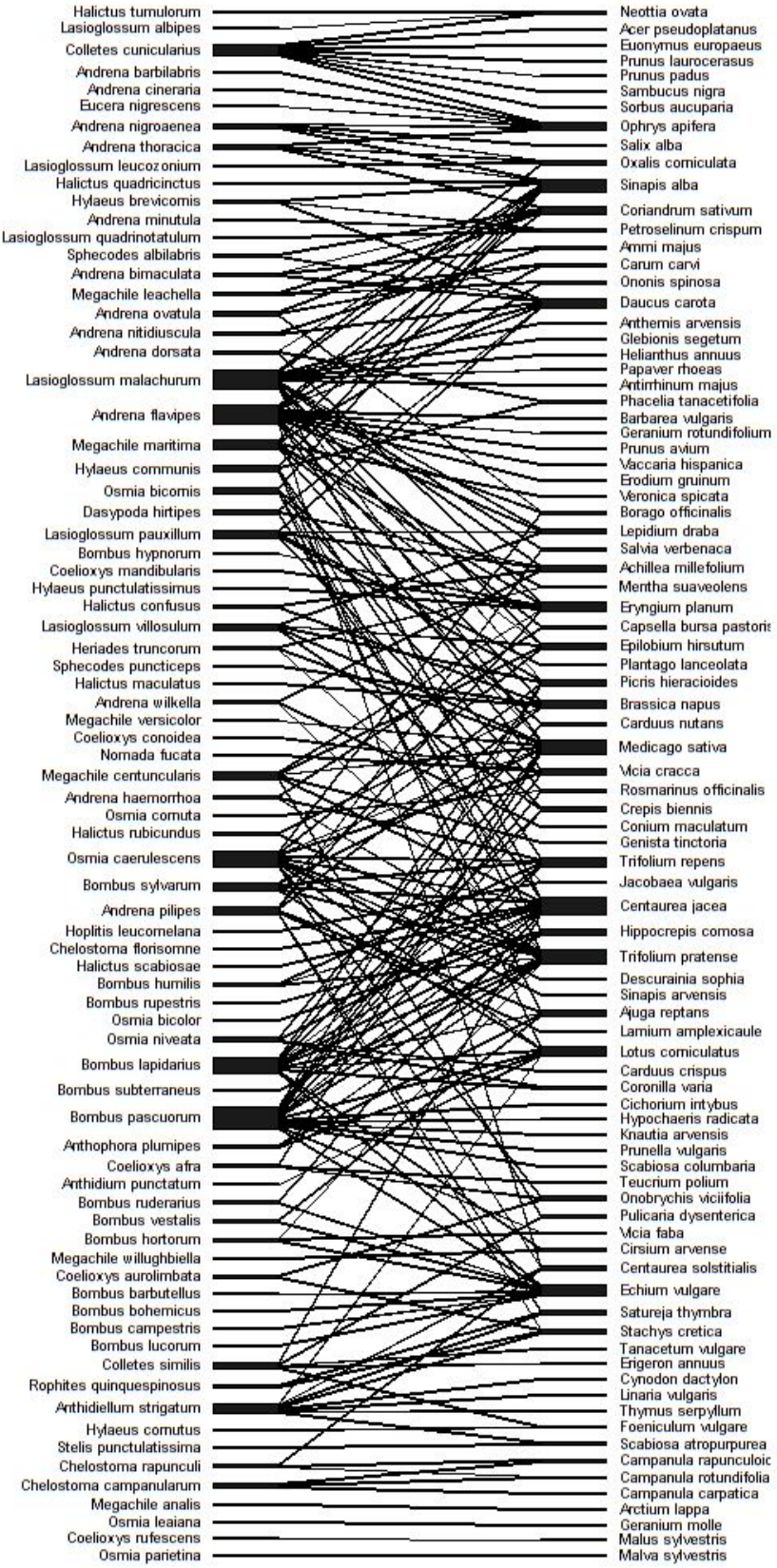
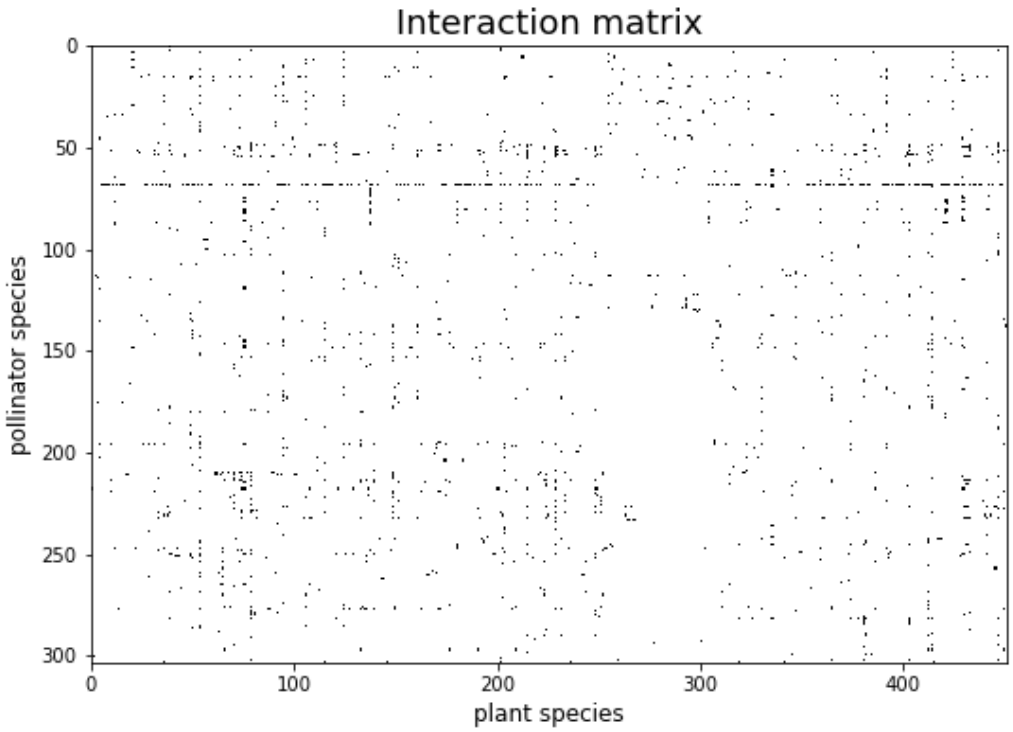
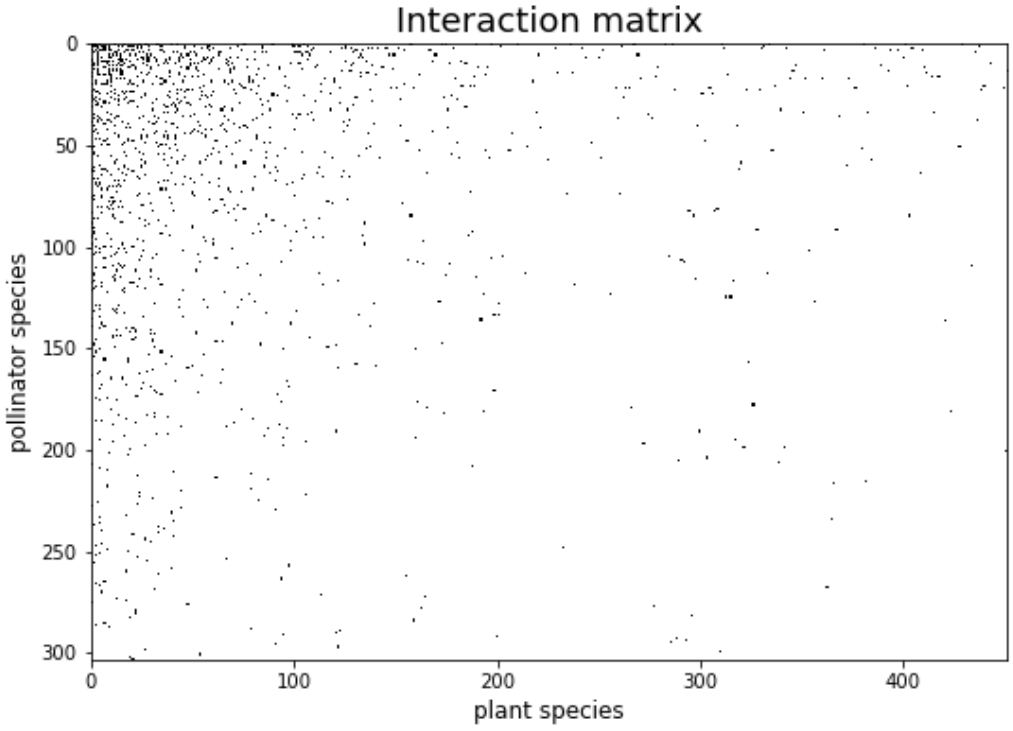


Figure 3.1: A first visualisation of the dataset. The left names represent pollinators, the right ones represent plant species. For clarity not all species are shown, but part of the hubs and part of the specialized species are included in the image.



(a) Interaction matrix of the species in alphabetical order.



(b) Interaction matrix of the species indicating nestedness.

Figure 3.2: Interaction matrix visualisation using heatmaps. Every black cell denotes an interaction, every white cell a non-interaction.

Figure 3.3: European honey bee *(Apis mellifera).*

H_B	7.2139
H_P	7.9230
$H_{B,P}$	10.5651
$H_{B\|P}$	2.6421
$H_{P\|B}$	3.3512
$VI(B;P)$	5.9933
$MI(B;P)$	4.5718
ΔU	1.9359

Table 3.1: Calculated values of the Information theoretic metrics (in bits), on the original dataset.

3.2 Information theory

In Table 3.1, the determined values of the Information theory can be found. The row species (i.e. the pollinators (bees)) denote variable B, the column species (i.e. the plants) denote variable P. Two main facts can be extracted from this table. Firstly, we can see that the conditional entropy of B given P is smaller than that of P given B. This means that when the plant is known, less uncertainty about the pollinator species remains than the other way around. As stated in Chapter 1, a conditional entropy can be interpreted as the expected number of binary questions that have to be asked to determine the particular species of an interaction, when the interaction partner is known. When a plant, for example, only interacts with one pollinator, no questions have to be asked and the conditional entropy given this plant is zero. When a plant can interact with two different pollinators, the conditional entropy given this plant is one, as one question suffices to determine the interaction partner of a specific interaction. With the defined metrics above, we could hence say that when the plant is known, on average 2.64 questions are necessary to determine the interacting pollinator. When on the other hand the pollinator is known, on average 3.35 questions are necessary to determine the particular plant species the pollinator has visited. There is more uncertainty left when the pollinator species is known, from of which we can conclude

that the pollinators in this network interact in a more generalistic way than the plants. When compared to the earlier mentioned statement of Morales-Castilla et al. (2015) (that a plant-pollinator relation is rather asymmetrical as pollinators are typically more specialized than plants), these are not in line. However when checked, the average number of interaction partners a plant has in this dataset is 3.35 (the median is 1), while the pollinators in the database on average interact with 4.97 plants (and the median is 2). Here, it is true that identifying the plant species when the pollinator species is known is harder than conversely.

Secondly, conditional entropy can be seen as a measure of network stability (recall Equation (1.12)). The two discussed conditional entropies can be added up to the variance of information, still a measure for the stability. As the contribution of $H_{P|B}$ is bigger than $H_{B|P}$, mainly the pollinators effectuate this stability.
Also, the variance of information and the mutual information can be compared, still with the same trade-off in mind. The larger the mutual information, the better the information transfer from B to P and vice versa. A high mutual information shows a rather limited number of interaction partners and hence a high efficiency of the interactions. When on the other hand the conditional entropies are larger, the number of potential interaction partners increases (there is more uncertainty), but the overall network becomes more stable. Here the variance of information dominates the mutual information so stability dominates efficiency.
The other metrics are hard to discuss, as they cannot be compared to other values. The joint entropy $H(B, P)$, for example, can be seen as a measure for diversity, but as there is no other network to compare with, no conclusions can be made.

As *Apis mellifera* (the Western honeybee) is the most striking species of the network, the properties are calculated a second time without this row. No huge differences can be seen but still there are some changes. ΔU drops a little, as the species distribution moved somewhat in the direction of a uniform distribution. Though, the biggest change is $H_{P|B}$ dropping to 2.8649, hence lowering the variance of information. This would imply a loss of network stability and could be considered as a negative impact. However the story for *Apis mellifera* is quite different than for other bees. The amount of human managed honeybee stocks has almost tripled during the past decades. The main reason for this is that they are highly efficient pollen and nectar foragers, capable of interacting with many agricultural crops. Despite this benefit for food cultivation, the drawbacks of this enormous increase in honeybees start to be revealed. As most productive agricultural crops are mass-flowering intensively, but only during a short period, the managed bees have to exploit other resources too temporarily. They form a direct competition for wild bee species and force them to shift their diets towards less profitable or scarce resources. Outcompeting local wild bees is not a de-

Domain:	Eukaryota
Kingdom:	Animalia
Phylum:	Arthropoda
Class:	Insecta
Order:	Hymenoptera
No taxon:	Aculeata *(Ants, bees and stinging wasps)*
No taxon:	Anthophila *(Bees)*
Superfamily:	Apoidea
Families:	Seven families are covered by bees:
	Andrenidae
	Apidae
	Colletidae
	Halictidae
	Megachilidae
	Melittidae
	Stenotritidae

Table 3.2: Taxonomy of the pollinators.

sired effect, especially since the best pollination service is provided by a combination of *Apis mellifera* and wild bees. Next to that, also plant species can be affected. A too high visitation frequency can lead to less success in reproduction, e.g. by preventing pollen tube development. So although *Apis mellifera* can provide great pollinating services to crops, still the trade-off has to be made between this and the effects of honeybee spillovers on wild plants and pollinators [39].

3.3 Species comprised in the network

3.3.1 Species phylogeny

The pollination network obviously consists of pollinators and plants. As stated above, many types of pollinators exist, but the used dataset only contains bees. Bees are considered to be a clade, called Anthophila. A clade is a group of organisms that consists of a common ancestor and all its lineal descendants. It represents a single branch of the tree of life [22]. The taxonomy of bees and hence of all pollinators in the FlorAbeilles dataset is presented in Table 3.2. The taxonomic data originated from the site bugguide.net [10].

For the plants present a similar overview is made (Table 3.3), but not classified till the family level as done above. This is because the plants in the dataset are not as closely related to each other as the pollinators, already giving a large set of different orders. Taxonomic data of plants came from the USDA Plants dataportal [61].

This way of using taxonomy gives a broad overview of which species are closer related to one another than others, but a better way of presenting this is with a phylogenetic

Domain:	Eukaryota	
Kingdom:	Plantae	
Subkingdom:	Tracheobionta	
Superdevision:	Spermatophyta	
Devision:	Magnoliophyta	
Class:	Liliopsida	
	Orders:	Commelinales
		Cyperales
		Liliales
		Orchidales
		Typhales
Class:	Magnoliopsida	
	Orders:	Sapindales
		Apiales
		Asterales
		Campanulales
		Capparales
		Caryophyllales
		Celastrales
		Cornales
		Dipsacales
		Ericales
		Euphorbiales
		Fabales
		Gentianales
		Lamiales
		Linales
		Malvales
		Myrtales
		Papaverales
		Plantaginales
		Plumbaginales
		Polygonales
		Ranunculales
		Rhamnales
		Rosales
		Rubiales
		Salicales
		Sapindales
		Scrophulariales
		Solanales
		Theales
		Violales

Table 3.3: Taxonomy of the plants.

tree. A phylogenetic tree is an estimate of the relationships among taxa and their hypothetical common ancestors. Today most phylogenetic trees are built from molecular data, meaning DNA or protein sequences [29]. Here, DNA sequences were used. The sequences came from the public data portal of the BoldSystems database [9], were downloaded in the FASTA format and were further processed in MEGA to generate a phylogenetic tree with all respectively phylogenetic branch distances. MEGA bases this algorithm on the maximum likelihood estimator. The chosen statistical method for the tree is the neighbour-joining tree.

For insects, the DNA sequence of the *COI* (Cytochrome-c-oxidase) gene is used to build a tree. This is a mitochondrial gene that takes part in the aerobic respiration. It generates an oxidase protein that catalyses the four-electron reduction of molecular oxygen to two molecules of water, and then utilizes the obtained energy to pump protons across the inner mitochondrial membrane [24].

For plants (or more specifically angiosperms), both the *matK* and the *rcbL* gene are commonly used. The *matK* (Megakaryocyte-Associated Tyrosine Kinase) gene is coding for a protein with corresponding name. This protein is an intron maturase, a protein that splices introns. The protein is thought to play a significant role in the signal transduction of hematopoietic cells[1] [25]. It is also able to inactivate Src family kinases, and may play an inhibitory role in the control of T-cell proliferation [4].
The *rcbL* gene then is the Large subunit of the RuBisCO gene (opponent of the Small subunit (rbcS)). RuBisCO (Ribulose-1,5-bisphosphate carboxylase/oxygenase) is a famous enzyme involved in the first major step of carbon fixation in plants. This is a process by which atmospheric carbon dioxide is assimilated in the Calvin cycle; i.e. converted to and stored as energy-rich molecules such as glucose [17]. It is thought to be the single most abundant protein on earth [13].

A small piece of the pollinator tree is shown in Figure 3.4. For the full trees of pollinators and plants is referred to a Google Drive folder
`https://drive.google.com/open?id=1o5D-xidcLJZRZkSv_45C5dhLLeKw-q_Y`.

The generated phylogenetic trees consist of external nodes (the tips) that represent the actual sequences that are available today, internal nodes that represent hypothetical ancestors, and branches that connect nodes to each other. The lengths of the branches represent the amount of change that is estimated to have occurred between a pair of nodes [23]. Note that the tree is shown in the 'Topology only' - mode, meaning that the length of the branch lines on the figure are unrelated to branch lengths (i.e. the difference in DNA). Only the value above the branches is representative for the actual branch length. The reason for this is that some nodes are separated by

[1] Haematopoiesis is the formation of blood cellular components.

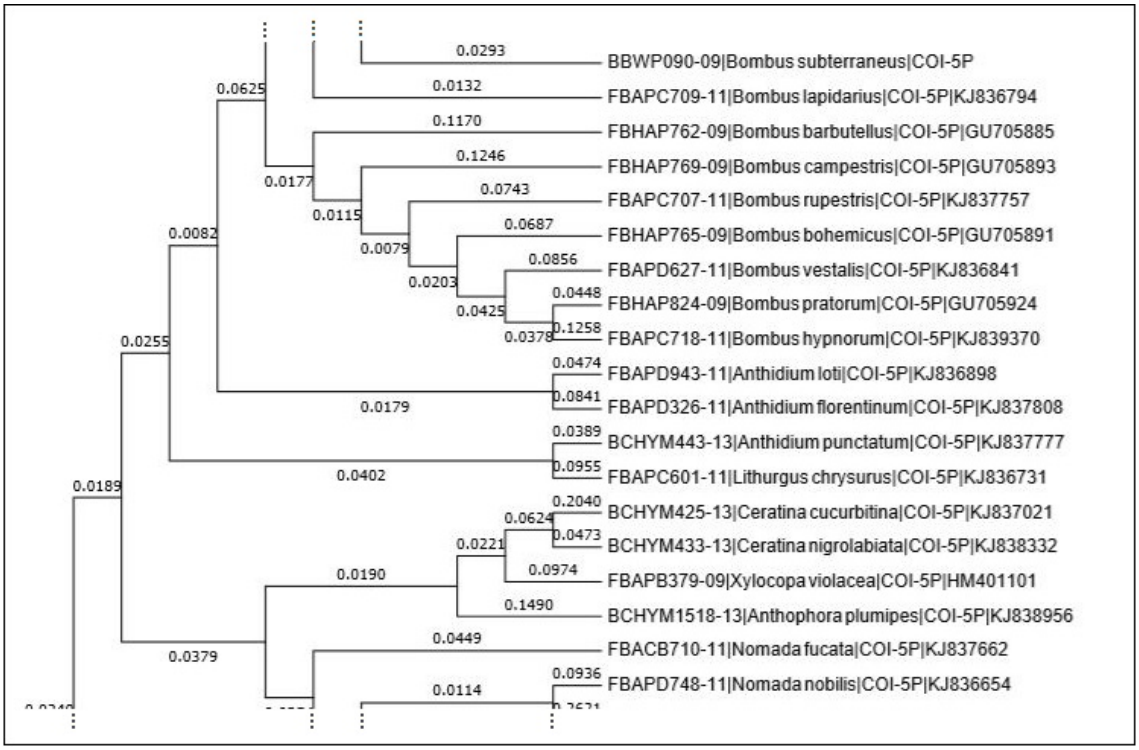


Figure 3.4: Part of the pollinator phylogenetic tree (made in MEGA with the neighbour joining tree method).

very short branches, whereas others are separated by very long ones. Sometimes it becomes impossible to visualise these short lines and, moreover, it gives the tree a messy look. When the actual branch lengths are replaced by their value, the tree only contains the topological information and the lines can be arbitrarily chosen in such way to give the tree a more elegant look.

3.3.2 Species traits

Characteristics of plants and pollinators can be denoted as traits, as introduced in Section 1.5. Not of all species traits were found, but the others are presented here. Data of the plant species came from the books *Veldgids Nederlandse Flora* of H. Eggelte [18] and *Geïllustreerde Flora van Nederland* of E. Heimans, H. Heinsius and J. Thijsse [31]; data of the pollinator species from *Veldgids Bijen voor Nederland en Vlaanderen* of S. Falk [20].

Plant traits

The first two gathered plant traits are the *Growth habit* of the species, ranging from herb, graminoid, subshrub, shrub to tree, followed by their (proportional) *Minimum, Maximum* and *Mean height*. These two traits can be seen in Figures 3.5 and 3.6, respectively. Because of some large trees as e.g. *Sorbus aucuparia* (the European mountain-ash), *Sambucus nigra* (the European elderberry) or *Euonymus europaeus*

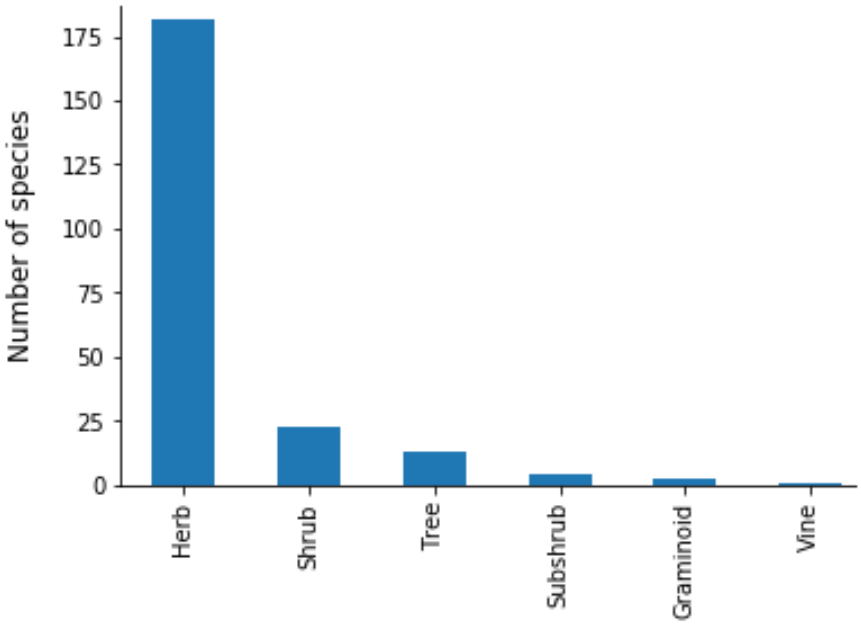
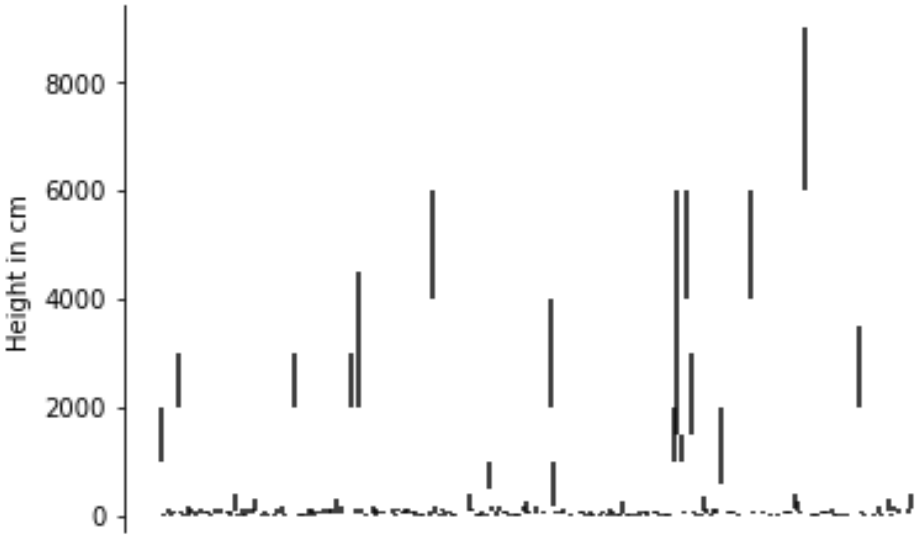


Figure 3.5: Distributin of *growth habits*.

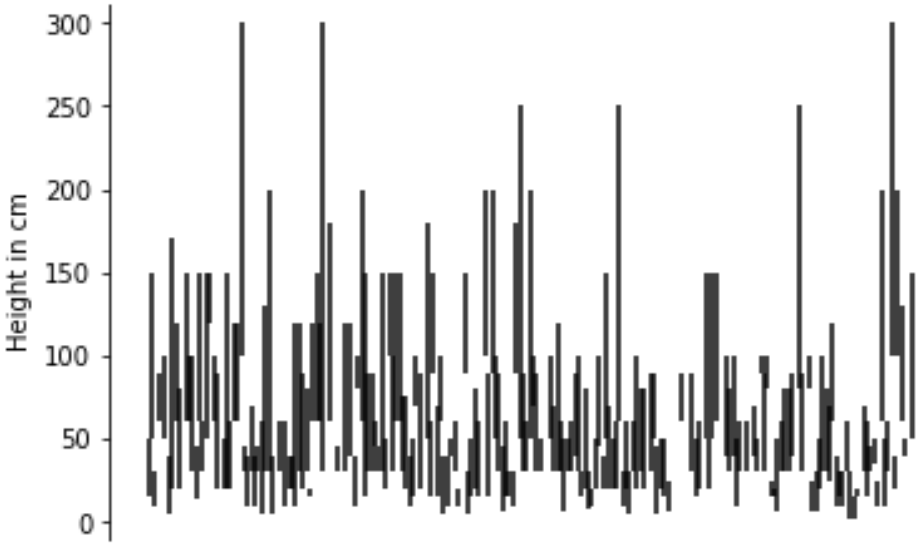
(the European spindle), the plot becomes quite unclear for the herb heights. Therefore, an adapted plot is made in Figure 3.6b where all plants with heights above 3 m are left out.

Next an overview of the *Blooming period* is given in Figure 3.7. This is done by binary labels for each month and each plant species, denoting whether the plant is flowering in this month or not. As can be seen (and expected), the summer months (June, July and August) are most common, but some species remarkably deviate from this. Examples are *Erodium gruinum* (long beaked stork's bill) blooming very early (from February to May), or *Hedera helix* (the common ivy) blooming very late (from September to November). Blooming all year long is very rare but also possible, like for *Senecio vernalis* (Eastern groundsel).

Two other characteristics of plants are their *Duration* and their *Category*. The duration of the plant can be annual, biennial or perennial. Annual plants are those whose entire life cycle occurs within one growth season, like many common garden plants. During this time, which can last from a few weeks to a few months, the plant will develop roots, stems and leaves and will die afterwards. In order not to go extinct, the plant only has this one growth season to also produce seeds. Seeds are the only things that allow these species to grow new annual plants the next season. The seeds are dormant (meaning they are not active) until the correct time of year, during which they will develop and go through their entire life cycle.
Biennials are plants that take two years to complete their entire life cycle. In the first year these plants are only vegetative, meaning that they do not produce reproductive



(a) Height of all plants.



(b) Height of plants below 3m.

Figure 3.6: Distribution of *plant heights*.

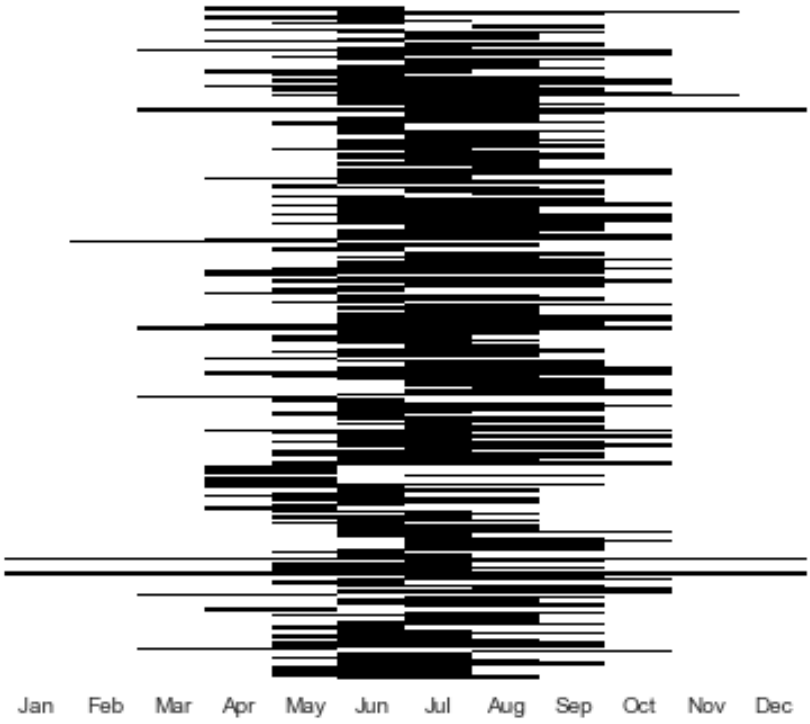


Figure 3.7: Distribution of *blooming periods*.

structures. In this period they grow roots below ground and a small rosette of leaves near the surface. At the end of this vegetative stage, the above-ground part of the plant may die or not but the roots remain. In the second year of growth, the stem elongates and flowers and seeds are produced. These seeds produce new biennials that will start their first growth season the year after.

At last, there are perennial species, mostly shrubs and trees, which persist for many growth seasons. Their vegetative (or juvenile) phase can be short (like biennials) but can also last for a few years. The species can be evergreen or deciduous, depending on the fact whether their foliage stays throughout the year or is dropped after every growing and blooming period.

The regulation mechanism of flower inducing is significantly different in perennial species compared to annual/biennial species. The floral promoter in annual/biennial plants induces all the above-ground meristems to flower in the same season, whereas in perennial plants a sophisticated regulation system (consisting of many different factors) finetunes flower inducing so that only a proportion of the meristems will be transformed into flowers at a certain time. This actually means that perennial plants are able to regulate their flower inducing in a quantitative way and hence partition their resources between reproductive and vegetative sinks, according to prevailing conditions. The partitioning of resources and the vegetative growth is required because of their long life span and need for competition with other species.

Consequences in e.g. orchards: trying to manipulate the flower inducing and shorten

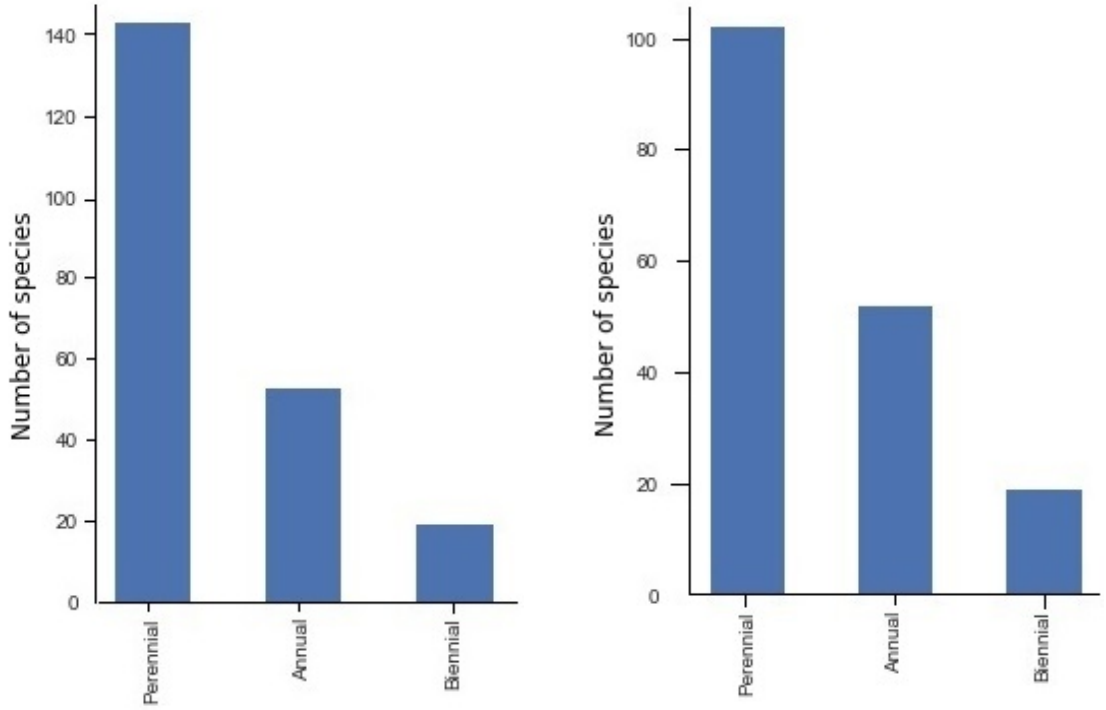
the juvenile period for perennial fruit trees will require more sophisticated techniques and knowledge compared to annual/biennial plants [5].
A bar plot of the duration of the dataset species is made in Figure 3.8a. Since all subshrubs, shrubs and trees are consistently perennials, a second plot of the duration of the herbs only is made. The ratio between both plots of course changes when leaving out part of the growth habits but the overall trends stays. Perennial is the most common duration type, biennial the least common one. Also, in both Figures 3.8a and 3.8b, seven species[2] are left out as for these, multiple growth categories were possible.

For *Category* the possibilities are monocot or dicot. Monocots denote flowering plants (angiosperms) whose seeds typically contain only one embryonic leaf or cotyledon. The largest family in this group is the family of the Orchidaceae. Dicots on the other hand have two cotyledons in their embryonic stage. A linked characteristic is that flower parts of monocots come in multiples of three, while flower parts of dicots come in multiples of four or five. Molecular phylogenetic research has shown that monocots form a monophyletic group, while dicots do not share a common ancestor. Therefore, the term 'dicot' is more used in the manner of 'not being monocot' [42].
Important to mention is that not all plant species can be labelled with one of these two categories. Conifers for example are none of both, as they are no flowering plants.
In the FlorAbeilles dataset, all plants are flowering. Of them, only thirteen are monocots, while all others are dicots.

The next trait is the *Flower colour* of the plant species. Similarly to the blooming period, the colour variable is changed into eight dummy variables, being white, pink, red, orange, yellow, green, blue and purple. Most plants only have 1 blossom colour, but sometimes there are multiple possibilities for one species. A clear example is *Antirrhinum majus* (the common snapdragon), being able to flower in white, pink, red and yellow (see Figure 3.9). Here, the values of the dummy variables are [1, 1, 1, 0, 1, 0, 0, 0]. A barplot of all flower colours is to be found in Figure 3.10. As can be seen there, white and yellow are the dominant flower colours in this dataset.

Further used traits were *Phyllotaxis* and *Flower symmetry*. Phyllotaxis is a categorical variable, telling how the leaves are placed on the plant stem. The basic phyllotactic patterns in nature are either opposite, whorled, alternate or basal.
In opposite phyllotaxis, leaf primordia grow one by one, but two successive leaves always grow on the opposite side of each other. This opposite pattern is further divided into the opposite distichous and the opposite decussated phyllotaxis. The difference is clarified in Figures 3.11a and 3.11b. In the first one, every leaf always grows 180°

[2] *Berteroa incana* (annual or biennial), *Centaurea diffusa* (annual or biennial), *Digitalis purpurea* (biennial or perennial), *Foeniculum vulgare* (biennial or perennial), *Geranium molle* (annual or biennial), *Jacobaea vulgaris* (biennial or perennial) and *Picris hieracioides* (biennial or perennial).



(a) Duration of all plants.

(b) Duration of herb species.

Figure 3.8: Distribution of *durations*.

(a) White (b) Pink (c) Red (d) Yellow

Figure 3.9: Flower colours of *Antirrhinum majus*.

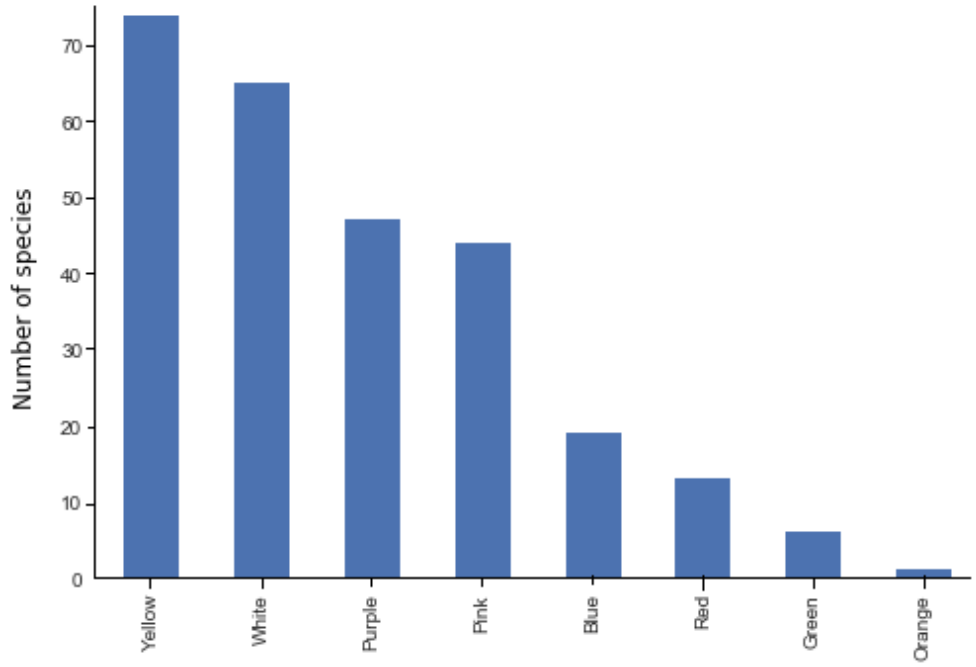
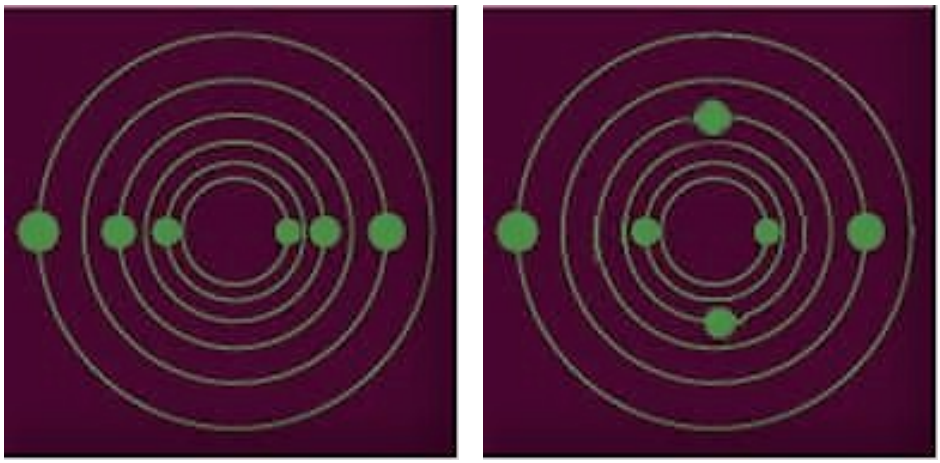


Figure 3.10: Distribution of *flower colours*.

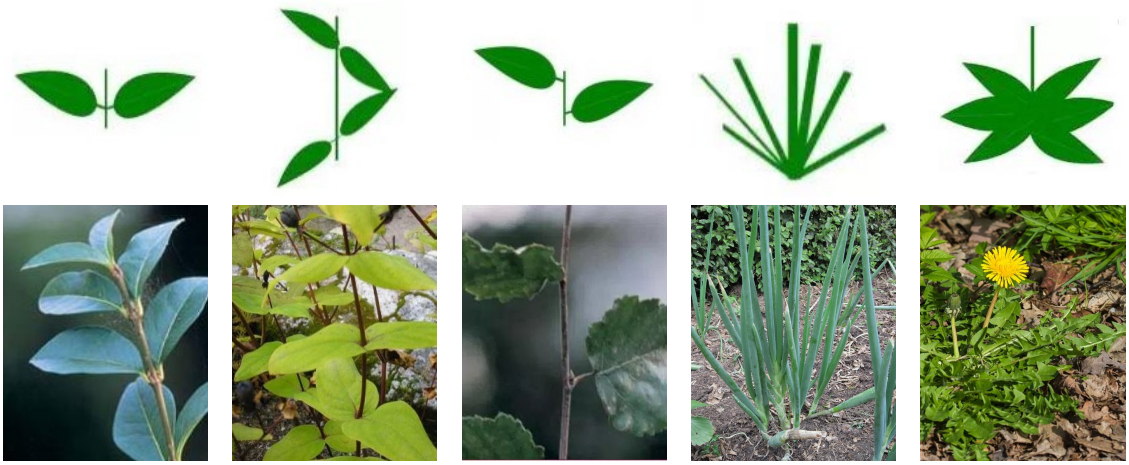
past the previous one, resulting in two clear rows of leaves. The second one can make angles of 90°, 180° and 270°, resulting in perpendicular successive leaf pairs. In whorled phyllotaxis, at least three leaf primordia grow at the same node. This results in rosettes of leaves along the stem. This did not occur in the investigated dataset. Alternate phyllotaxis is the most common type in nature (approximately 80% of the plant species). There leaf primordia can grow one per node, or two or more leaf primordia can grow at the same node, but each leaf is always at a constant divergence angle of 137.5° from the previous one. This gives the impression that all leaves are spread quite randomly over the stem. The last type considered is the basal one. Here leaves are not placed over the stem, but all start near the surface. A special type of this basal phyllotaxis is a ground rosette.
How all these types look, is made clear in Figure 3.12. In the dataset 64% of the species had an alternate phyllotaxis, 6% an opposite structure being distichous, 25% an opposite structure being decussated, six species having a basal phyllotaxis and three species having a rosette one.

For the *Flower symmetry* three types are considered, being asymmetrical, bilaterally symmetrical and versatile symmetrical. The difference is the number of symmetry axes. An asymmetrical flower has no symmetry axis; a bilaterally symmetrical flower has one, and a versatile symmetrical flower has multiple. In the found traits 20% of the flowers were asymmetrical, 38% were bilaterally symmetrical and the other 42% were formed versatile symmetrically.



(a) Opposite - distichous. (b) Opposite - decussated.

Figure 3.11: Two types of the opposite phyllotaxis.



(a) Distichous (b) Decussated (c) Alternate (d) Basal (e) Gr. rosette

Figure 3.12: *Pyllotaxis*.

The last three traits are *Position of the ovary, Number of styles* and *Number of stamens*. The ovary is the part of the pistil that contains the ovules. An ovule on its turn is that what becomes a seed after fertilization [21]. In general, ovary positions are classified as superior or inferior. When the ovary is superior, it lies above the attachment of the stamens, petals, and sepals. Such flowers are called hypogynous. When the ovary is inferior, it lies below the attachment of the outer floral whorls. These are epigynous flowers [41]. In the dataset 121 species had n superior ovary, 76 species an inferior and one species (namely *Euonymus europaeus* (the European spindle)) had a partially inferior ovary, which is an intermediate form.

The pistil is the female organ of a flower, consisting of an ovary, style, and stigma. The ovary is already discussed above and the stigma is the part of a pistil or style that receives the pollen. The style is then the elongated part of the pistil between the ovary and the stigma [21]. The number of styles ranged from 0 to 6, with a median of 1 and a mean of 1.35.
The stamen (plural stamina or stamens) is the pollen-producing reproductive organ of a flower. As being the pollen-bearing organ of the flower, it is the male organ in the angiosperms [21]. The number of stamens ranged from 1 to 40, with a median of 5 and a mean of 7.46.

Table 3.4 provides an overview of all used plant traits and a specific example. The example species chosen is *Ligustrum vulgare* (the wild privet or common privet).

Growth habit	Categorical	Shrub
Minimum height (cm)	Numerical	50
Maximum height (cm)	Numerical	200
Mean height (cm)	Numerical	125
Blooming period	Dummy variables	[0,0,0,0,0,1,1,0,0,0,0,0]
Duration	Categorical	Perennial
Category	Categorical	Dicot
Flower colour	Dummy variables	[1,0,0,0,0,0,0,0]
Phyllotaxis	Categorical	Opposite decussated
Flower symmetry	Categorical	Versatile symmetrical
Position ovary	Categorical	Superior
Number of styles	Numerical	1
Number of stamens	Numerical	2

Table 3.4: Overview of all used plant traits and the specific example of the wild privet.

Pollinator traits

The first trait is *Voltinism*. Univoltine insects produce one generation per year, while bivoltine insects produce two generations per year. The bivoltinism does not have to be symmetric. When the resource availability peak is not in the middle of the year but e.g. falls early in the season and then decreases, the first generation has a shorter larval feeding stage than the second generation. The bivoltine life cycle is

Figure 3.13: Distribution of *flying periods*.

more likely to be superior to the univoltine one if (1) growth is fast, (2) the suitable growing season is long, (3) the biomass loss during nonlarval stages is small, and (4) the egg size is small [33]. For most insects of the dataset this seems not to be the case, as 87 species have are univoltine and only 11 are bivoltine.

Just like plants are only flowering for a limited time period, each pollinator also has a limited *Flying period*. A similar plot as for the plants is made, based on binary variables for each month-insect combination. The result can be seen in Figure 3.13. Overall, it can be said that flying periods are longer than blooming periods, but also (and logically) the most popular months are May, June, July and August. Still, both March/April and September/October are very abundant in the graph.

The next two traits are *Nesting type* and *Status*. The first one is quite clear (denoting the place and/or category of the insect's nest), and the second one is about the abundance of the species. The possibilities range from disappeared, possibly disappeared, very rare, rare, quite rare, quite common, common, very common to increased. A plot of both variables can be found in Figures 3.14 and 3.15 respectively. For a few species two nesting types were possible, so they contributed to two bars of the plot. As can clearly be seen, nesting on the ground is the most common nesting type. In the Status-plot no specific trend is observed, as common and rare have the same frequencies and the variations on these alternate[3].

The last pollinator trait is the *Size* of the bees. A similar plot as the plant heights is made in Figure 3.16. The smallest pollinator was *Hylaeus brevicornis* with a size of 4 to 5 mm, the biggest one was *Bombus rupestris* ranging from 20 to 24 mm.

[3]Possibly *Status* is a less indicating variable, as the occurrence of the species comes from Belgian data, while the FlorAbeilles dataset comes from France. Still, the status can be considered to be quite proportional.

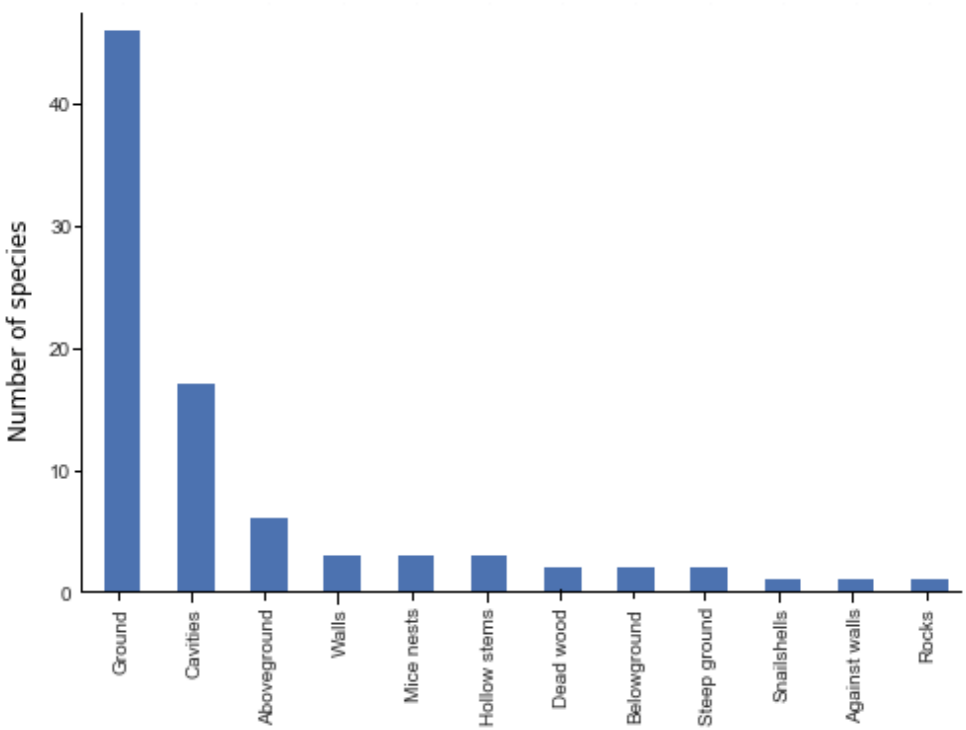


Figure 3.14: Distribution of the *nesting types*.

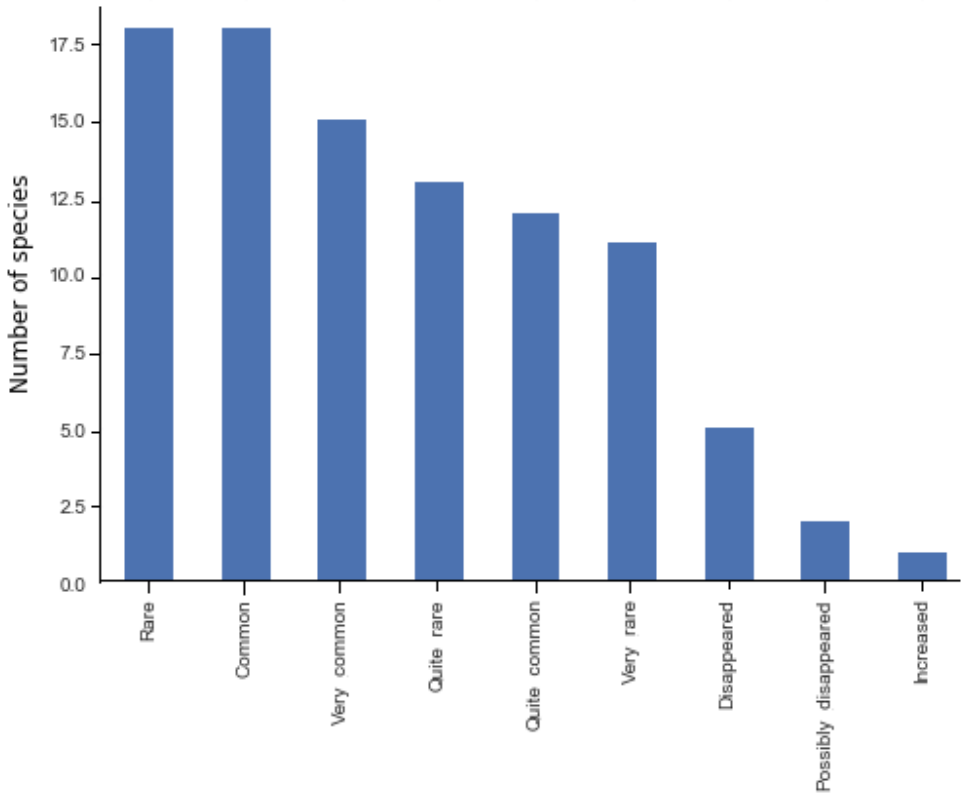


Figure 3.15: Distribution of the *statuses*.

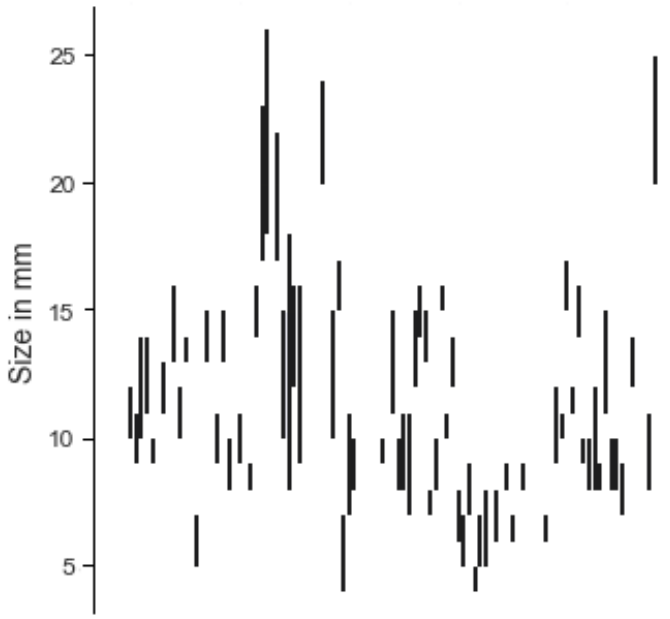


Figure 3.16: Distribution of *pollinator sizes*.

Table 3.5 provides an overview of all used pollinator traits and a specific example. The example species chosen is *Andrena flavipes* (the yellow legged mining bee).

Voltunism	Categorical	bivoltine
Blooming period	Dummy variables	[0,0,1,1,1,1,1,1,1,0,0,0]
Nesting type	Categorical	Ground
Status	Categorical	Very common
Minimum height (mm)	Numerical	11
Maximum height (mm)	Numerical	13

Table 3.5: Overview of all used pollinator traits and the specific example of the yellow legged mining bee.

Again, the traits of the species comprised in the network can be found in the same Google Drive folder as mentioned on page 48. There also, the original binary FlorAbeilles dataset is stored.

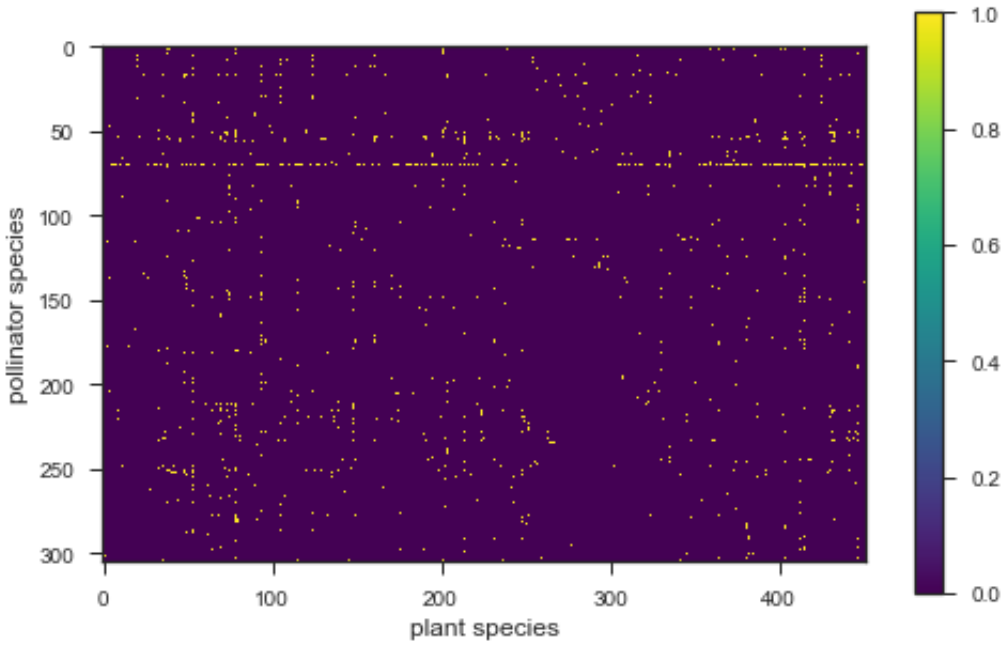
CHAPTER 4

PREDICTION OF PLANT - POLLINATOR INTERACTIONS

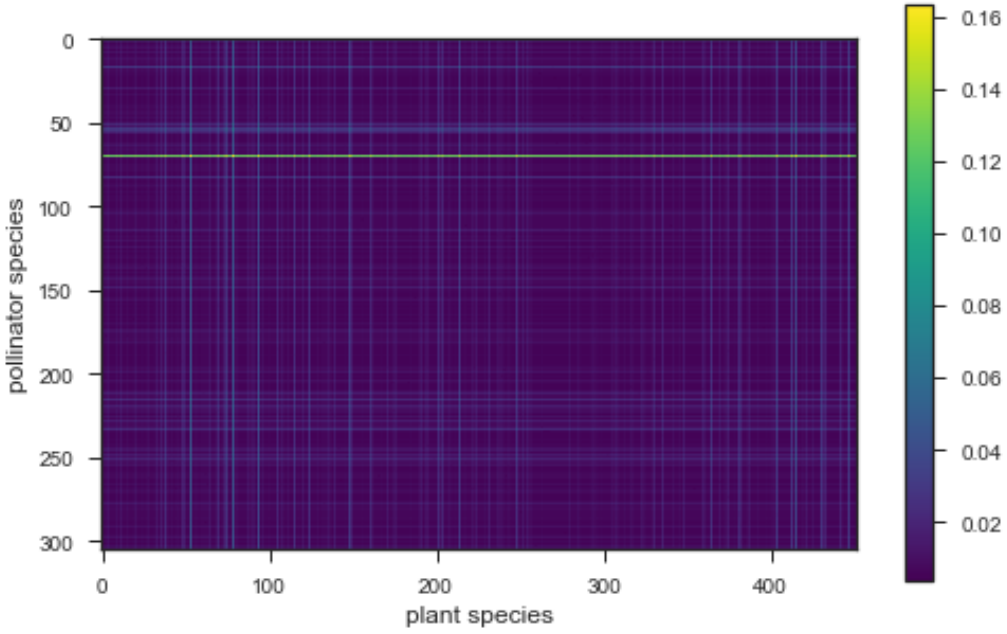
4.1 Linear filter

When not using any features of the plants or pollinators, predicting labels (interacting or non-interacting) can be done using a linear filter. This method generates a value between zero and one, related to how likely it is for the interaction to happen in reality. With Equation (2.5), a leave-one-out cross-validation can be performed, where the original value is always left out and the interaction estimation is purely based on the rest of the matrix (i.e. the info comprised in the network). Full explanation was given in Section 2.2.1.

Firstly, the filter was applied to the data with all α's being 0.25. The ROC curve based on the original values Y and the filtered values F gave an AUC of 0.9979, but as stated above, this formula still contains the respective original values. Therefore we will only focus on the LOO-values from now on, because only these can give a proper validation and proper estimation of the performance. The generated ROC curve (based on Y and β) gave an AUC of 0.8390, which is relatively high compared to the score of 0.5 of a random classifier. AUC (area under the ROC curve) is a measure for the ability of a model to rank true interactions higher than non-interactions, independent of prediction score threshold, but this concept was already introduced in Section 2.3. A new heatmap is plotted in Figure 4.1b, showing the distribution of these generated interaction values. In theory all values lie distributed between zero and one, but the large number of zeros in the original matrix (sparsity of 98.90%) pulls these values remarkably down. As can be seen in the figure, species already having more interactions in Y now obtained 'high' LOO-values for all of their possible interactions. This corresponds to what was stated earlier: it is less feasible for a generalistic pollinator to *not* interact with a certain plant than it is for a very specialized pollinator.



(a) Heatmap of the binary interaction matrix.



(b) Heatmap of the generated LOO-values (i.e. β).

Figure 4.1: Heatmaps before and after filtering (already leaving out the respective original value).

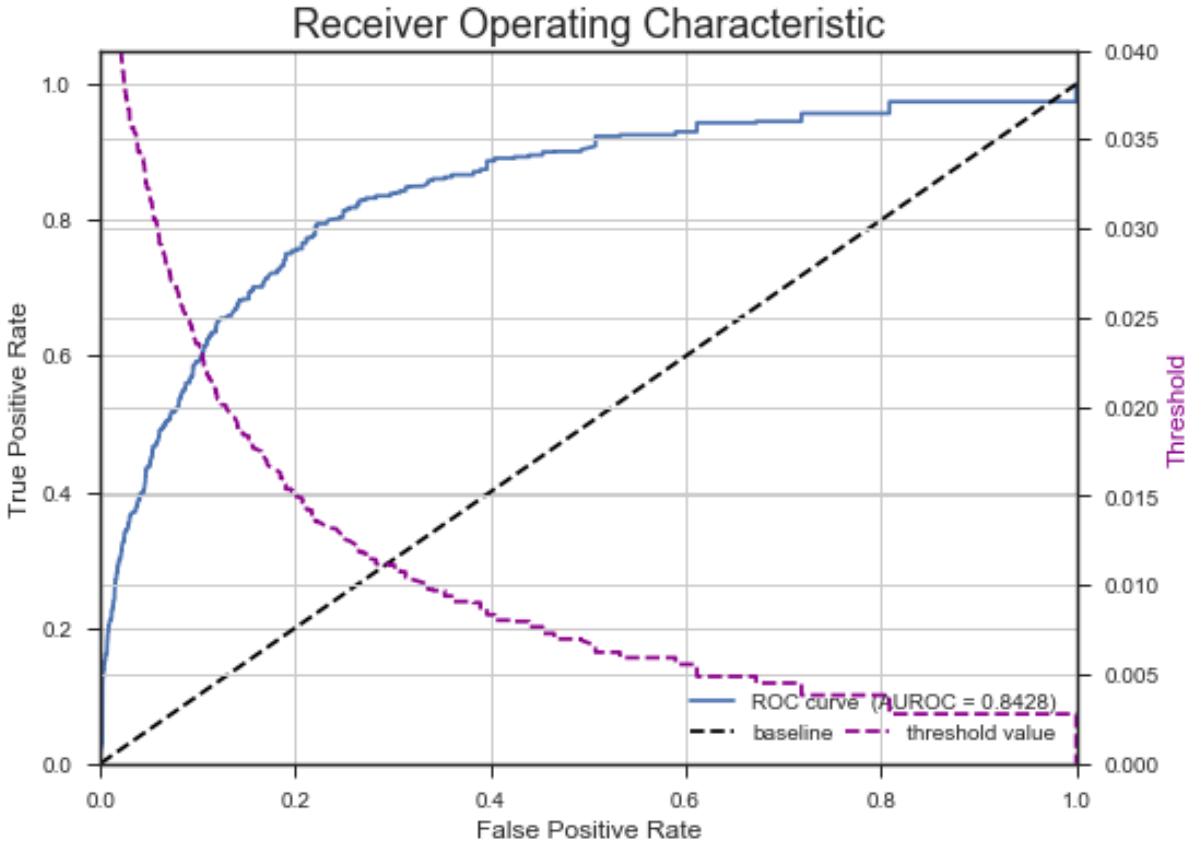


Figure 4.2: The ROC curve of the linear filter model, with $[\alpha_1, \alpha_2, \alpha_3, \alpha_4] = [0.05, 0.50, 0.45, 0.00]$ and the ROC curve of a random classifier on the left axis, together with the threshold values used to construct the blue ROC curve on the right axis.

Furthermore, different values for the parameters α were tested. Normally hyperparameters are optimized with a type of cross-validation, so since the β interaction value is computed as a leave-one-out cross-validation, different α's can be filled in and the AUC based on these scores can be optimized. All α's between zero and one with a step size of 0.05 were tried, always making sure the four alphas summed up to one. This resulted in 1761 possible parameter combinations, all generating their own AUC. The obtained values lie between 0.6570 and 0.8428, with this highest AUC corresponding to $[\alpha_1, \alpha_2, \alpha_3, \alpha_4] = [0.05, 0.50, 0.45, 0.00]$. The top-26 highest AUC's all left α_4 as zero. The following 23 highest AUC's all left α_1 as zero. This implies that the terms in Equation (2.1) with α_2 and α_3 are the most valuable for the model. A further evidence for this statement is that the lowest 424 AUC values corresponded with a parameter combination in which either α_2 was zero, α_3 was zero, or both (resulting in the worst models). However, still 1125 of the total 1761 combinations gave an AUC above 80%, so it can be concluded that the exact combination of α's is not of major importance, as long as enough weight is assigned to the second and third term.

The ROC curve of the model with this optimal parameter combination is plotted in Figure 4.2, together with the performance curve of a random classifier. Also a third line can be seen on the plot, denoting the thresholds taken to separate interacting and non-interacting pairs. With a ROC curve analysis, an optimal threshold can be determined, i.e. an optimal cut-off to classify β, based on the original interaction matrix. Many approaches are available to determine the optimal threshold for a classifier, e.g. taking the point with maximal accuracy, the average predicted probability/suitability approach, the sensitivity-specificity sum maximization approach, the sensitivity-specificity equality approach or the approach based on the shortest distance to the top-left corner (0,1) in ROC plot. Information on these methods and their differences can be found in Liu (2005) [38]. Firstly, one of these theoretical approaches is chosen, namely the last mentioned one of the closest point to (0,1) (which was the point that only a perfect classifier contains). When computed, this point corresponded with a threshold of 0.014, which can now be seen as an optimal threshold for the generated β's. Note that this value is quite low compared to all computed scores, meaning that a lot of interactions will be classified as positive by the filter. The distribution of β can be seen in Figure 4.3. Also, Figure 4.4 shows the distribution of β by the means of a bloxplot. The highest point marked on the ROC curve next to the boxplot is the point closest to (0,1). This corresponds to a threshold of 0.014.
A second possibility would be to intuitively choose the *best* point of the ROC curve, e.g. a point where not excessively many positive predictions are made but where all these predictions are correct. This is based on what was said in the section about performance: *In many biological networks, however, the number of interactions is much lower than the number of non-interactions. It is therefore important to achieve a low FPR because even moderate FPR can easily lead to much more false positive predictions than true positive predictions, and hence a very low precision*. The chosen point is also marked on the ROC in Figure 4.4,i.e. the lowest one. This corresponds with a threshold of 0.053.

In the same figure, the boxplot of the computed β's is shown to scale the thresholds. The first quartile (Q_1), median (M) and third quantile (Q_3) are respectively 0.00383, 0.00661 and 0.01292, meaning that 50% of the scores lies between these outer values. The distance between Q_1 and Q_3 is also called the interquartile range (IQR). Note that the plotted maximum in the boxplot is $\beta = 0.02652$ instead of the real maximum of 0.23291. This is because all values greater than $Q_3 + 1.5$ IQR are considered as outliers.

Although the AUC of 0.8428 seemed already promising, a ROC curve is only a theoretical performance estimation. Therefore, a second practical evaluation was conducted. Web of Life [66] is an online database portal where a lot of ecological networks can be

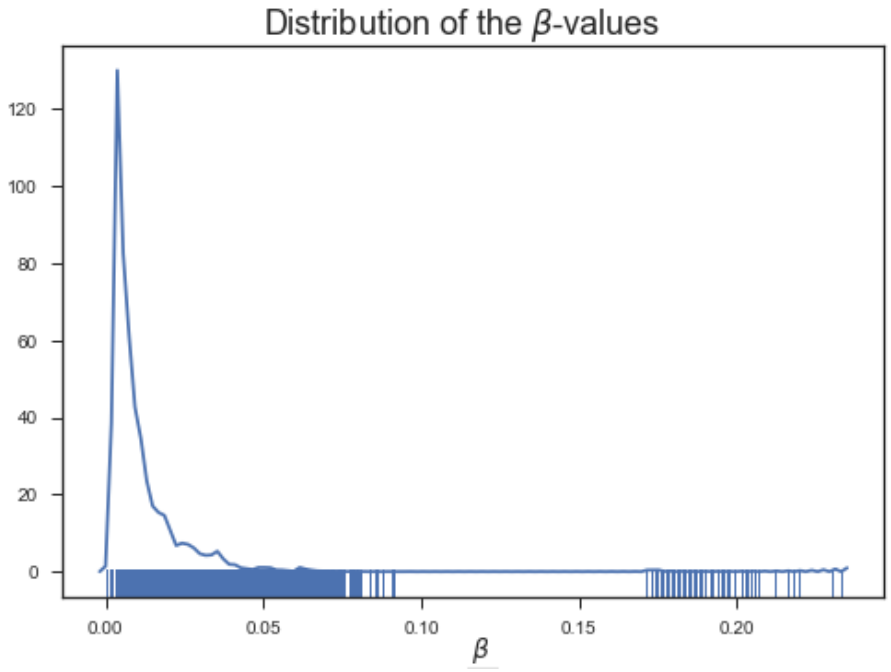


Figure 4.3: Distribution plot of all 137 860 computed β's.

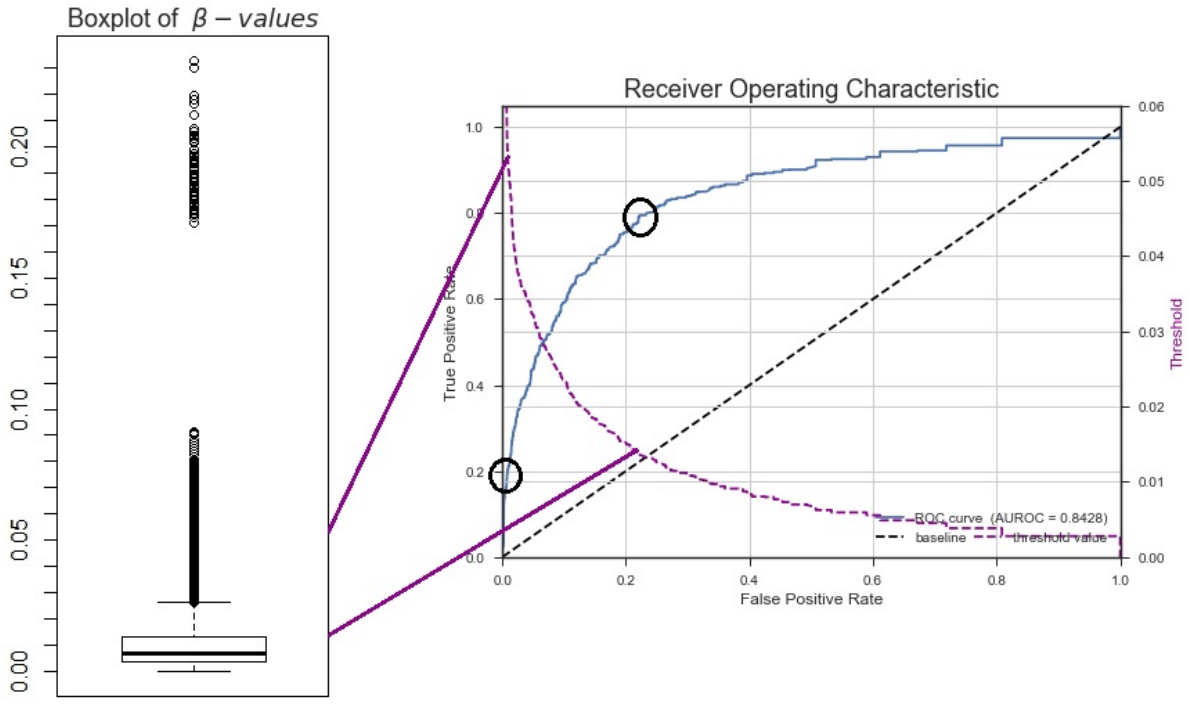


Figure 4.4: Overview of the computed β's (left plot), the ROC curve of the linear filter (with optimal parameters $[\alpha_1, \alpha_2, \alpha_3, \alpha_4] = [0.05, 0.50, 0.45, 0.00]$) plus the corresponding thresholds to construct the ROC curve (right plot) and the relative position of the selected optimal thresholds in the β-range (connecting the left and right plot). The chosen points on the ROC curve are highlighted. The respective thresholds are 0.014 and 0.053.

Plant	**Pollinator**	**Original Y-value**	**Computed β-value**	**Value in the Web of Life dataset**
Eryngium campestre	*Anthidiellum strigatum*	0	0.0625 → false negative?	1 → indeed false negative
Scolymus hispanicus	*Lagioglossum malachurum*	0	0.0384 → false negative?	1 → indeed false negative
Cirsium arvense	*Bombus pascuorum*	0	0.0339 → false negative?	1 → indeed false negative

Table 4.1: Evidence for detecting false negatives with the linear filter model (threshold 0.014), based on a real life dataset.

found. Among these numerous files, both binary and weighted pollination datasets from around the world are present. Firstly we focused on one binary pollination network, namely a dataset of Greece (from the Aristotelian University in Thessaloniki [45]) containing quite some similar species as the dataset used in this work. Having defined an optimal threshold for the model, one can now search for *false negatives* in the original matrix Y. Important to mention: this term is not completely similar to what is mentioned in a confusion matrix (cfr. Table 2.1)! Normally, a false negative (FN) is a value/category referring to the predicted matrix. A prediction is classified as FN if it is positive in reality (i.e. in Y), but is (wrongly) predicted as negative by a classification model. Then the predicted value is incorrectly assumed to be negative. Analogous to this, we now want to find false negatives of the *original* dataset. By this we mean interactions that could be incorrectly considered as negative during field research, but are possible to occur in reality. For this, we will use predictions of the filter model. Instead of typically assuming that Y is 100% correct and looking for mistakes in the predictions, we will use the predictions to possibly find missing interactions in Y. The considered values will be those plant-pollinator combinations for which Y denoted a zero, but where the linear filter would predict a positive interaction value (i.e. β was larger than 0.014 - if we focus on the first method). After searching these previously mentioned species combinations in the dataset of Greece, indeed some false negatives were detected. Examples are shown in Table 4.1.

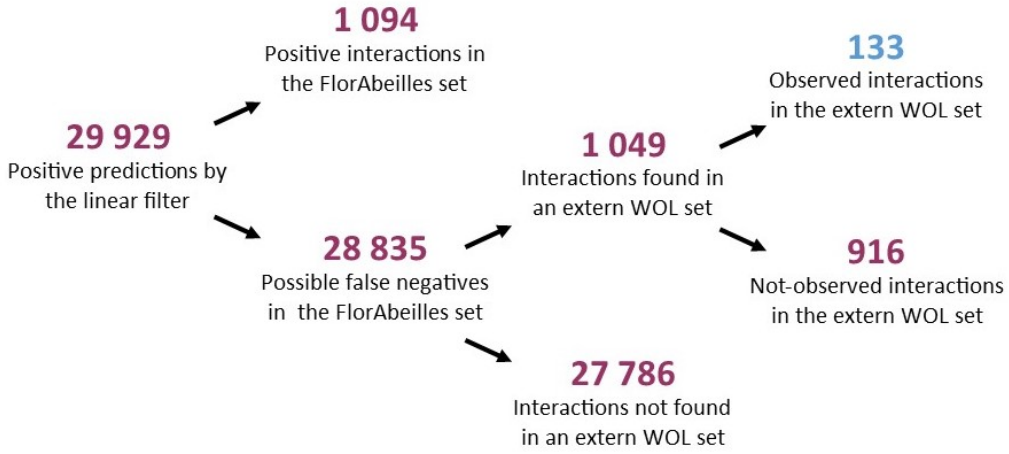
From this table, one can conclude that some predicted interactions were not observed during the timespan of composing the FlorAbeilles dataset, but actually can occur in nature as they are present in the dataset of Greece. Considering that Greece and France are not that far-removed and both plant and pollinator species are present in these countries, it is possible that these interactions are not non-happening in France, but were just missed during field research. This underlines the usefulness of machine learning models in ecology. The fact that most datasets are constructed based on field observations, it is hard to find evidence that a certain interaction is surely non-happening. Some negative interactions in a dataset can be false negatives and can

with the use of these models be prioritized for further research in the environment where the dataset was created.

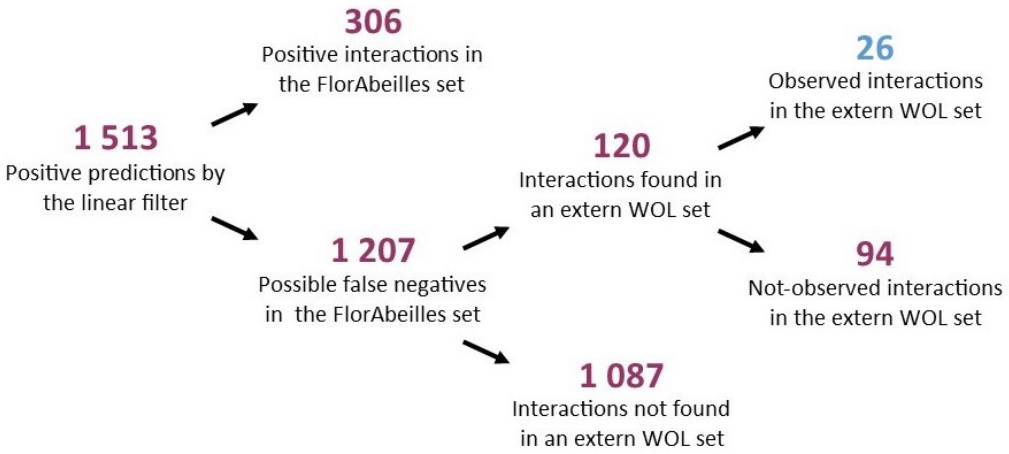
Secondly, all the pollination datasets of Web of Life were considered. In total there were 40 binary and 143 weighted ones, distributed over all continents. This of course already means that 'validation' of a missing interaction is less straightforward, as occurrences of species may differ. However still, if a predicted interaction is present in another dataset, this nonetheless means that this one is biologically possible.
Of all predicted interactions ($\beta > 0.014$) only the possible false negatives of Y are examined, as we assume that the positive values of Y are all correctly observed. This then boils down to 28 835 interactions. The reason for this sizable number is that with the construction of the filter, the more generalistic the species, the greater the chance of having positive predictions with all 452 plants or all 306 pollinators. The earlier mentioned statement can again be stressed: it is less feasible for a generalistic pollinator to *not* interact with a certain plant than it is for a very specialized pollinator. The example species of Table 4.1 are also all highly non-specific[1]. Of the 28 835 examined plant-pollinator combinations of Y, only 4.099% was also present in another file of Web of Life. (Not many species of the FlorAbeilles dataset corresponded to species comprised in these geographic widely distributed pollination files. Only 13 of the 183 files could be used.) Still, of this 4% (i.e. 1049 species pairs) that was present in another file, 133 interactions were negative in Y but positive in the other file. This means that at least 133 interactions were not detected when creating the original dataset, but are possible to occur in nature. These do not necessarily all have to happen in the neighbourhood of the original dataset, but can again be prioritized in field research in this area.

If this last practical validation is re-done with the second threshold of 0.053, the number of predicted interactions obviously goes down quickly. Now the filter only generates 1513 positive predictions (instead of 28 835) of which 1207 interactions are possible false negatives; the others are already positive in Y and are hence considered to be correct. Now 9.942% of these plant-pollinator combinations was present in another Web of Life-file so could be externally checked. 26 of the 120 interactions were found to be positive in the other files, so are proven to be biologically possible. Again, these pollinations do not automatically happen in France, but confirm the statement quoted in the theory section: negative interactions with high scores are natural targets for increased sampling effort, as they are most likely to occur in reality. The prioritized field research is a crucial concept in this chapter. It is also good to know that some generalistic species are able to interact with more partners

[1] Pollinators *A. strigatum, L. malachurum* and *B. pascuorum* resp. have 26, 25 and 24 interactions; plants *E. campestre, S. hispanicus* and *C. arvense* resp. have 20, 7 and 5 interactions. These are all high numbers for pollinators and plants respectively, considering the dataset.



(a) Scheme based on the first threshold of 0.014.



(b) Scheme based on the second threshold of 0.053.

Figure 4.5: Visualization of the external validation experiments.

than initially expected, as this can have a positive influence on the network stability (cfr. Section 1.2).

4.2 Two-step kernel ridge regression

The second model includes more information of thecomprised objects/dyads. To use the collected traits *and* phylogeny in models, the cross-section of the species with known traits and the species with available *COI*- or *matK*-sequence is going to be used. In this way, the same plants and pollinators are included in all data files and the predictive power of a model with traits can be compared to one using phylogeny. The result is a subset of 96 pollinators and 193 plants[2].

For two-step kernel ridge regression, the interaction matrix and two similarity matri-

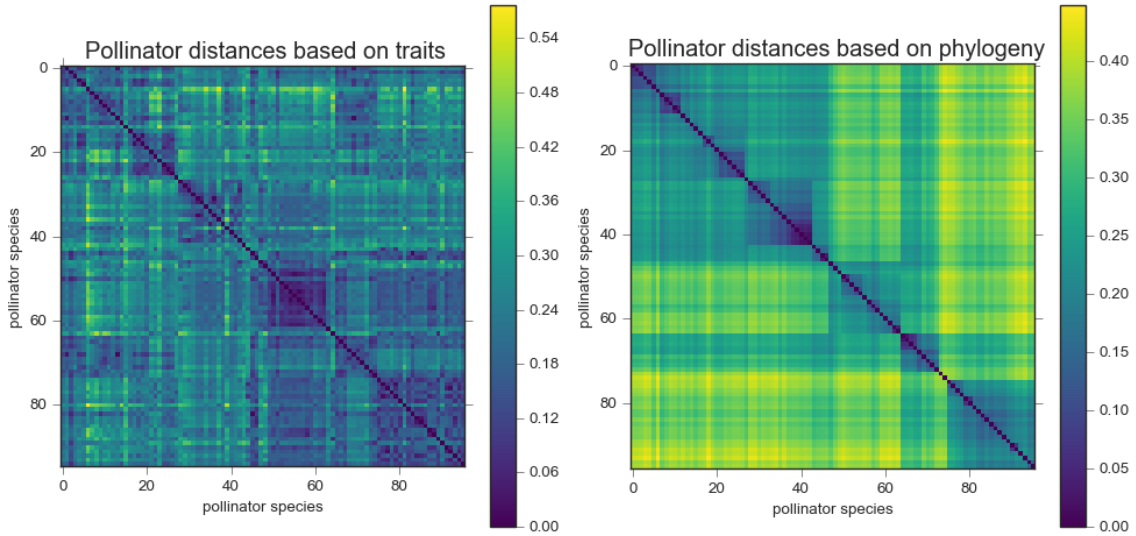
[2]It could be important to know that *Apis mellifera* is not included in the subset. This pollinator is manually introduced in a lot of places and is able to interact with nearly all plants. This could positively influence the performance metrics (as these predictions will be very accurate), but this could give a small overestimation of the model's real performance.

ces are necessary. The latter is realized by the use of kernels. Note that one kernel matrix contains similarity values of only *one* set of species, never mixing with the other set. We will hence end up with four Gram matrices:

- K^T: a (96×96) similarity matrix of pollinators, based on traits,
- K^P: a (96×96) similarity matrix of pollinators, based on phylogeny,
- G^T: a (193×193) similarity matrix of plants, based on traits, and
- G^P: a (193×193) similarity matrix of plants, based on phylogeny.

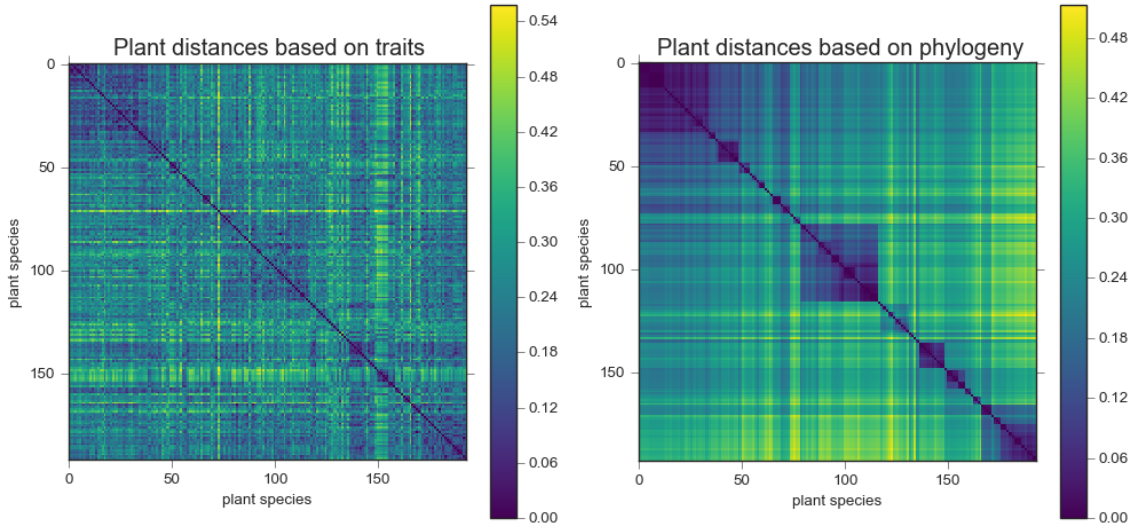
To create those kernels, we start with dissimilarity matrices which are later transformed to similarity matrices. The files containing the *traits* of plants and pollinators can be converted to two square dissimilarity (or distance) matrices by using the *Vegan* package in R. The distance metric specified is the 'Gower' distance, frequently used for ecological datasets. The reason for this is that many datasets of traits contain as well numerical values (e.g. height of the plants, size of the pollinators,...), binary values (e.g. monocot/dicot, univoltinism/bivoltinism,...) as categorical values (e.g. five options for growth habit, twelve options for nesting type,...). Most distance metrics (e.g. the Euclidean distance) have difficulties coping with this combination, but the Gower metric is designed to handle such data, without requiring any recoding for multistate or quantitative characters [27]. It first rescales every column separately by dividing each entry by the range of the corresponding variable, after subtracting the minimum value. Hence each variable is scaled to a range of $[0, 1]$. Afterwards the distance between two items is the average of all the variable-specific rescaled distances.
The *phylogeny* of plants and pollinators is currently stored in the produced phylogenetic trees. To export these, two Newick files are generated in MEGA. A Newick file can be processed in Python using the *Phylo* module. This module of BioPython reads the file with the branch lengths, determines the terminal nodes (i.e. the species) and calculates the total distance between them using all these branch lengths.

Now the four distance (or dissimilarity) matrices are constructed. Of course these matrices are symmetrical, as the distance between species i and j is the same as between species j and i. Heatmaps are plotted for visualization in Figure 4.6. For convenience, the same order of species is used for rows and columns, so a clear diagonal of distance zero can be noticed (the distance of a species to itself is zero). These cells are dark blue. When paying attention, one can also see that the darker blue squares of more similar species in the phylogeny matrices can be found back in the ones based on traits. This means that closer related organisms based on DNA show corresponding features.



(a) Heatmap of the pollinator distances, based on traits.

(b) Heatmap of the pollinator distances, based on phylogeny.



(c) Heatmap of the plant distances, based on traits.

(d) Heatmap of the plant distances, based on phylogeny.

Figure 4.6: Heatmaps of dissimilarity.
Note: The 96 pollinator and 193 plant species are plotted in the order produced by their respective phylogenetic trees. In this way it is clear to see that species close to each other in this matrix (i.e. close to each other in the tree) are more similar. These plots are more legible than those where the species were sorted alphabetically (as in the original interaction matrix). With regard to similarity/dissimilarity, alphabetical sorting is almost random.

As all distances lie distributed between zero and one, the similarity matrix can just be computed by taking '1-dissimilarity'. The other tried option to go from dissimilarity to similarity was Non-metric Multi Dimensional Scaling (NMDS). This also gave decent results but was in fact a superfluous step that can cause a possible loss of information. Still, this technique can provide great alternative visualisations of the distances, but this is left behind. The kernels themselves are solely based on a simple math operation on the computed dissimilarities.

One last thing will be done before staring with the actual TSKRR. As the name suggests and as explained earlier, TSKRR is based on regressions. As all the kernel evaluations are centered around zero, the regression can be hampered. The most simple solution is adding an intercept, by adding 1 to every element of the four kernels. This is not to be confused with adding the identity matrix to a kernel, which would make all the species relatively more distinct form each other and give biased results. For the combined model where all available data is used (so where the pollinator and plant kernel are the sum of their respective trait-based and phylogeny-based kernels), this intercept of 1 is not necessary.

Now we have an interaction matrix (which is a subset of the original matrix Y containing the pairwise binary interaction values for the 96 and 193 species) and the respective square kernels to conduct the regressions. Two-step kernel ridge regression can be performed. The package *xnet* in R contains all formulas described in Section 2.2.2. It is based on the paper of Stock et al. (2018) [59] and is developed by Joris Meys.
Three initial models are trained, i.e. one using the two trait-based kernels, one using the two phylogeny-based kernels and one using the two summed kernels. The regularization parameters λ_u and λ_v of the TSKRR predictions (introduced in Equation (2.11)) are firstly set to 0.1. By comparing the obtained predictions after training to the original interaction matrix, the fit of the training data can be determined. The trait-based model realizes an AUC of 1, the phylogeny-based model an AUC of 0.9998 and the combined model an AUC of 1. What is more important, is the model performance on a test set. As stated earlier, the training and test sets can be divided in four different ways. The performance estimations of these four prediction settings are mentioned in Table 4.2, using the four cross-validation schemes of Figure 2.4b in a leave-one-out framework. Shortcuts to calculate these metrics are available and are included in the *xnet* package. Considering that a test AUC of 0.5 resembles random guessing and 1 represents a perfect classifier, these values are not bad to start from.

One would expect that the performance decreases from setting A → (B,C) → D as prediction becomes harder when less information or examples of the dyad are included in the training set. (Setting A observes both plant and pollinator species in

Prediction setting	Trait-based model	Phylogeny-based model	Combined model
Setting A	0.8248	0.8304	0.8768
Setting B	0.7317	0.7491	0.8405
Setting C	0.7955	0.7111	0.8190
Setting D	0.6930	0.5664	0.7693

Table 4.2: Performance (AUC) when using leave-one-out cross-validation in four different schemes, on the initial trained models (i.e. with all λ's being 0.1).

the training set, setting B and C observe one of both species and setting D has to make predictions for completely new dyads). Also, as there are almost two times more plants than pollinators included, the model probably generalize better to new plants. The performance of setting C is hence expected to be slightly higher than that of setting B. When checked, the values of Table 4.2 tend to follow these trends, although some deviations are observed.

Depending on the setting it is sometimes more beneficial to make predictions based on the species' traits or on the species' phylogeny. For the prediction of interactions of a new pollinator species, it seems to be slightly better to focus on the phylogeny of all comprised species of the network. The interactions of a new plant species, on the other hand, seem to be more reproducible using a model based on the plants' and pollinators' traits. This nonetheless does *not* mean that the interaction behaviour of pollinators is more determined by this pollinator's phylogeny and a plant's interaction behaviour more by its traits, as e.g. the phylogeny-based model is based on the phylogeny of both sets (plants *and* pollinators). In addition, this observation only applies for the models trained with this arbitrarily chosen parameter set.

Next to the two separate models, the overall results imply that the combined model performs best for all settings, which is not illogical as this one includes all available information about the comprised species. The next paragraphs will sometimes only focus on this combined model.

To improve the model's performance, their regularization parameters can be optimized. For the final estimate of those parameters, *all* data points are used. A range of $[10^{-4}, 5.10^{-4}, 10^{-3}, 5.10^{-3}, 10^{-2}, 5.10^{-2}, 10^{-1}, 0.5, 1, 5, 10]$ is tried for both parameters, leading to $11^2 = 121$ possible λ-combinations. The best parameter-combination for the four different settings of the combined model is shown in Table 4.3, together with their highest reachable AUC. These AUC-values also fulfil the expected trends explained above.

The optimal parameter combinations for other models are shown in Table 4.4, without the highest reachable AUC's with leave-one-out tuning. The maximal AUC's of the trait-based model are very similar to those of the combined model, but those of the phylogeny-based model are somewhat lower.

Prediction setting	**Optimal (λ_u, λ_v)-combination**	**Corresponding AUC with LOO**
Setting A	$(1,\ 1.10^{-1})$	0.9041
Setting B	$(1,\ 5.10^{-2})$	0.8622
Setting C	$(5,\ 1.10^{-1})$	0.8714
Setting D	$(5,\ 5.10^{-2})$	0.8289

Table 4.3: Optimization of both regularisation parameters of the combined model, based on the highest reachable LOO-AUC and all species of the interaction matrix.

Prediction setting	**Trait-based model**	**Phylogeny-based model**	**Combined model**
Setting A	$(1,\ 1.10^{-1})$	$(2.10^{-2},\ 1.10^{-3})$	$(1,\ 1.10^{-1})$
Setting B	$(1,\ 1.10^{-1})$	$(1.10^{-1},\ 10)$	$(1,\ 5.10^{-2})$
Setting C	$(5,\ 1.10^{-1})$	$(1.10^{-1},\ 5.10^{-1})$	$(5,\ 1.10^{-1})$
Setting D	$(5,\ 1.10^{-1})$	$(5,\ 1.10^{-1})$	$(5,\ 5.10^{-2})$

Table 4.4: Optimization of both regularisation parameters of all models, based on the highest reachable LOO-AUC and all species of the interaction matrix.

However, rationally it would not be fair to say that our new models now perform with these certain (higher!) AUC's, as they were optimized for those specific data. Therefore, nested CV is the solution. Nested CV splits the data in training, tuning and test sets. The general idea is that the part used for the ultimate validation is never mixed with the optimization process. The principle is explained in Figure 4.7 for the example of setting B, with four outer folds and eight inner folds. In this case, we will use four outer folds for settings B (row wise) and C (column wise), and sixteen outer folds for settings A and D. Before allocating the dyads to different folds, the rows and columns of the interaction matrix and corresponding kernels are randomly shuffled. The tuning of the parameters in the inner folds will be done with the leave-one-out approach, meaning that the number of inner folds depends on the number of dyads left in the outer training data.

Nested CV is performed here to give an honest estimation of the model performance. Four or sixteen times the optimal (λ_u, λ_v)-combination is selected based on the highest LOO-AUC possible for this outer training fold. Again for both parameters the range $[10^{-4}, 5.10^{-4}, 10^{-3}, 5.10^{-3}, 10^{-2}, 5.10^{-2}, 10^{-1}, 0.5, 1, 5, 10]$ is tried, producing 121 λ-combinations. Then these regularisation parameters are used to train the specific outer training fold, and the performance of this model on the independent test set is computed. When repeated for all outer folds, this will leave us with four or sixteen external AUC's per setting, of which the average or median will give an honest estimation of the performance for this setting. It is intuitive to see that these values are more sincere than values where the whole dataset would be used to do both optimization and validation. The output of the nested cross-validation experiments for the combined

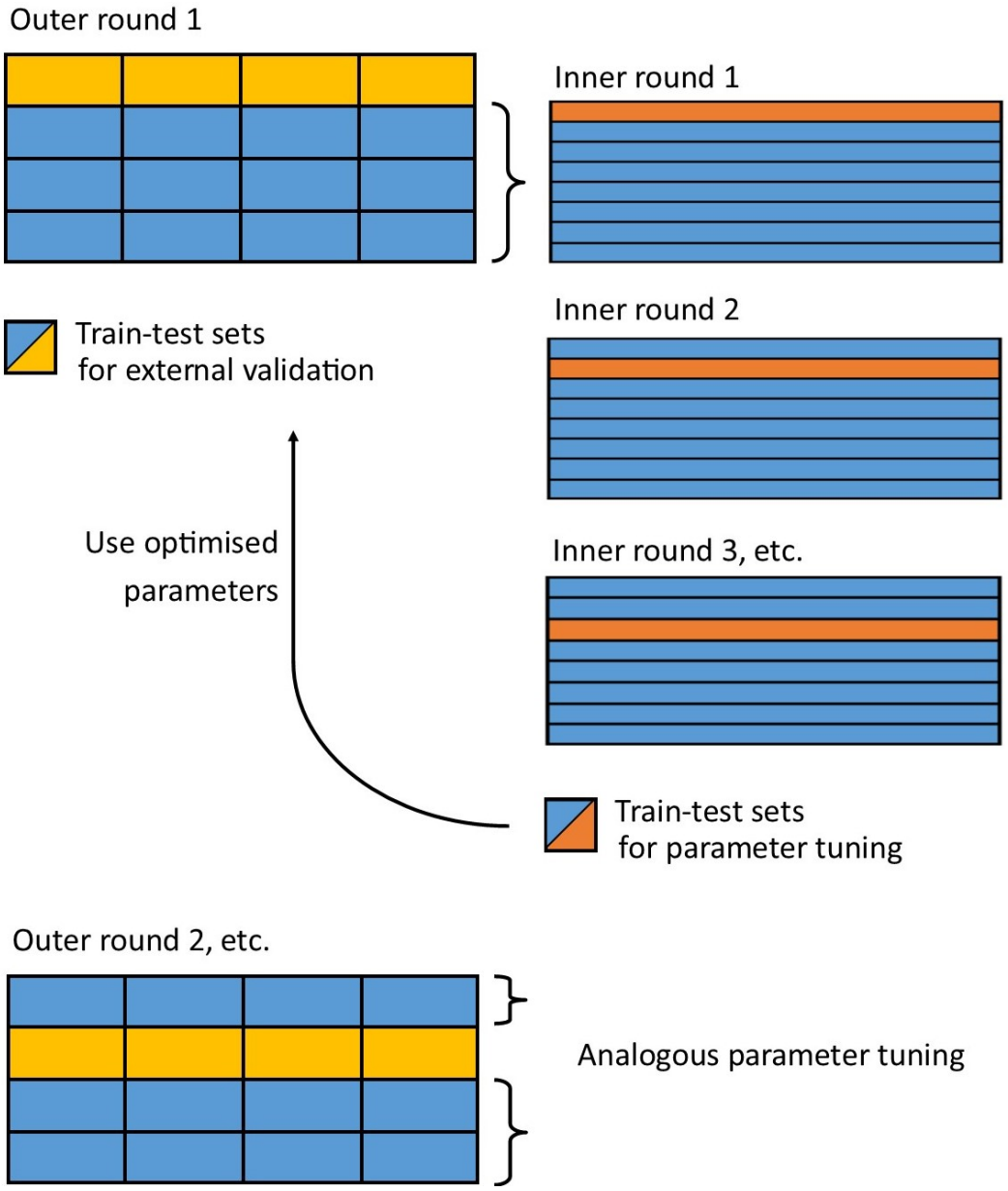


Figure 4.7: Demonstration of nested cross-validation on setting B, with four outer folds and eight inner folds. The yellow data is always completely new for the model, and is never combined with the optimization process. The other three settings show a similar pattern of exclusion, which can be composed by looking at Figure 2.4b.

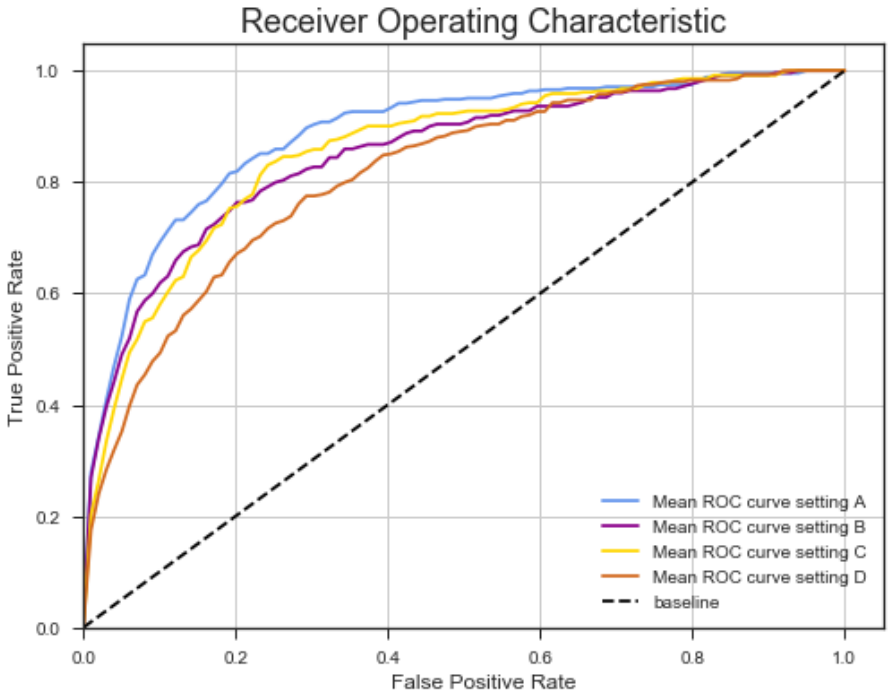


Figure 4.8: The four ROC curves generated by the nested-cross validation experiments for the combined model. Corresponding AUC's can be found in Table 4.5.

model can be found in Table 4.5. The representative AUC-values (i.e. the medians of all performances per setting, displayed in bold) did of course decrease in comparison with those of Table 4.3, but are still notably high. The values of Table 4.3 were anyway never meant to denote an actual performance. These were only used to determine the best overall regularisation parameters. The newly obtained performance estimates clearly follow the proposed patterns. The expected difference between setting B and C is hence confirmed. The same experiments are conducted for the two other TSKRR models. The outcomes are visualized in the next section of this chapter.

The ROC curves of the different prediction settings for the combined model are given in Figure 4.8. These curves are all a combination of four or sixteen different ROC curves, based on mean true positives and mean true negatives.

As a conclusion, we can say that with this model predictions for all types of *new* pollination interactions can be made. Performance estimations show that predictions are (way) better than random guessing, so the initial goal of building a model is already met and reasonable predictions can be made. Still one needs to realize that this model can impossibly describe all cause-effect relations related to pollination. A number of straightforward traits were used to characterize the incorporated species, but these can be further extended with knowledge from biology. By incorporating more and more information, other patterns in data can be discovered and more precise predictions can be made.

Prediction setting	**Fold**	**Optimal** ($lambda_u, \lambda_v$) **combination after training**	**Corresponding AUC on test set**
Setting A	1	(1, 0.1)	0.9046
	2	(1, 0.1)	0.9711
	3	(1, 0.1)	0.8921
	4	(1, 0.1)	0.9244
	5	(1, 0.1)	0.8420
	6	(1, 0.1)	0.9232
	7	(1, 0.1)	0.8489
	8	(1, 0.1)	0.7380
	9	(1, 0.1)	0.9345
	10	(1, 0.1)	0.8931
	11	(1, 0.1)	0.88841
	12	(1, 0.1)	0.8114
	13	(1, 0.1)	0.9312
	14	(1, 0.1)	0.8831
	15	1, 0.1)	0.8985
	16	(1, 0.1)	0.8921
			0.8926
Setting B	1	(1, 1.10^{-2})	0.8564
	2	(1, 0.1)	0.8264
	3	(5, 0.1)	0.8178
	4	(1, 0.1)	0.8806
			0.8487
Setting C	1	(5, 0.1)	0.687
	2	(5, 0.1)	0.8645
	3	(5, 0.1)	0.8076
	4	(5, 5.10^{-2})	0.8694
			0.8666
Setting D	1	(5, 1.10^{-2})	0.8221
	2	(10, 5.10^{-2})	0.8754
	3	(5, 5.10^{-2})	0.8208
	4	(5, 5.10^{-2})	0.8776
	5	(5, 5.10^{-2})	0.8425
	6	(5, 5.10^{-2})	0.9531
	7	(5, 0.1)	0.7404
	8	(5, 0.1)	0.7907
	9	(5, 5.10^{-2})	0.6813
	10	(10, 5.10^{-2})	0.7542
	11	(5, 0.1)	0.7592
	12	(5, 0.1)	0.6006
	13	(5, 5.10^{-2})	0.8302
	14	(5, 5.10^{-2})	0.8298
	15	(5, 5.10^{-2})	0.7895
	16	(5, 5.10^{-2})	0.8506
			0.8260

Table 4.5: Outcome of nested cross-validation experiments on the combined model. For each setting and each fold, the result of the optimization process in the training set is given (as parameter combination) together with the performance on the independent test set. The value in bold is always the median of all performances in this setting.

However, all articles described in Section 1.5.2 only needed a small number of traits to lead to good fits. Of course, e.g. Santamaria and Rodríguez-Gironés (2007) aimed at predicting topologic network properties, while here specific interactions are to be predicted. Hence, more data can be of use. The model uses data gathered from different sources (because ecological, morphological,... properties of species are not always neatly listed in tables or openly accessible) and can be used to prioritize field searches and detect false negatives. This detection is in fact a more useful goal to start with than being capable of perfectly predicting imaginary pollination networks. Models do not have to replace field searches, and new data from researchers (of e.g. changing abundances or interaction behaviour) will always be necessary, but in this way machine learning can help to detect gaps in the collected interaction information. Next to this finetuning of datasets (stated as the main purpose of this work), being able to predict which pollinators can interact with which plants - based on biological characteristics - may be helpful in other areas too. Studies can be done concerning the fate of several agricultural crops in times of climate change, the fate of pollinator abundances if invasive species are introduced, studies about co-extinction or rewilding ecological communities, and so on...
Biology and models can be mutually beneficial for each other.

4.3 Overview of all performances

In this summarising section, all performances are visualized. The linear filter model can only predict setting A as β is computed by a leave-one-out modus without withdrawing entire rows or columns. The three TSKRR models (the trait-based, phylogeny-based and combined model) on the other hand can make predictions for all four settings.

To make a comparison, the values should be as objective as possible. Therefore, the linear filter will be visualized without parameter optimization, i.e. with all α's being 0.25. Besides, the model is trained again for the same subset of 96 pollinators and 193 plants, to not be able to attribute eventual differences to the comprised species. For the kernel-based models the outcome of the nested cross-validation experiments will be visualized. As there are no 'standard' parameters for this type of ridge regression, these cannot be used. The tuned models where the maximal AUC was computed are strongly overfitted and were only used to determine the optimal overall parameter combinations. The most honest estimation of performance is hence with a tuning step, but with the performance prediction on an independent test set, never mixed with the optimization set. Everything can be seen in Figure 4.9. All values (except the first one) are the medians of four or sixteen AUC's.

Where the combined model scored best in Table 4.2 (where leave-one-out was performed on the whole dataset with arbitrary chosen parameters), the trait-based model seems to reach these performances after nested cross-validation. The difference can probably be attributed to the regularisation parameters. In Table 4.2 both λ_u and λ_v were chosen to be 0.1, while here an optimization process is included. If the trait-based model/combined model performs better with other parameters, this will be displayed here. Still, it was already stressed that the chosen optimal parameter combination is not connected to the set on which the performance is calculated, but other orders of magnitude may lead to better results in general for several models. Overall it can be said that, although the phylogeny-based model can be used to make rough predictions, it has limited added value to the combined model if both the trait-based and combined model are used in optimal mode to make predictions.

Possibly, all performances of the TSKRR-experiments can be increased by incorporating more of the networks' species and more information about these species. Still, both conclusions of Sections 4.1 and 4.2 show the ecological benefit of the models.

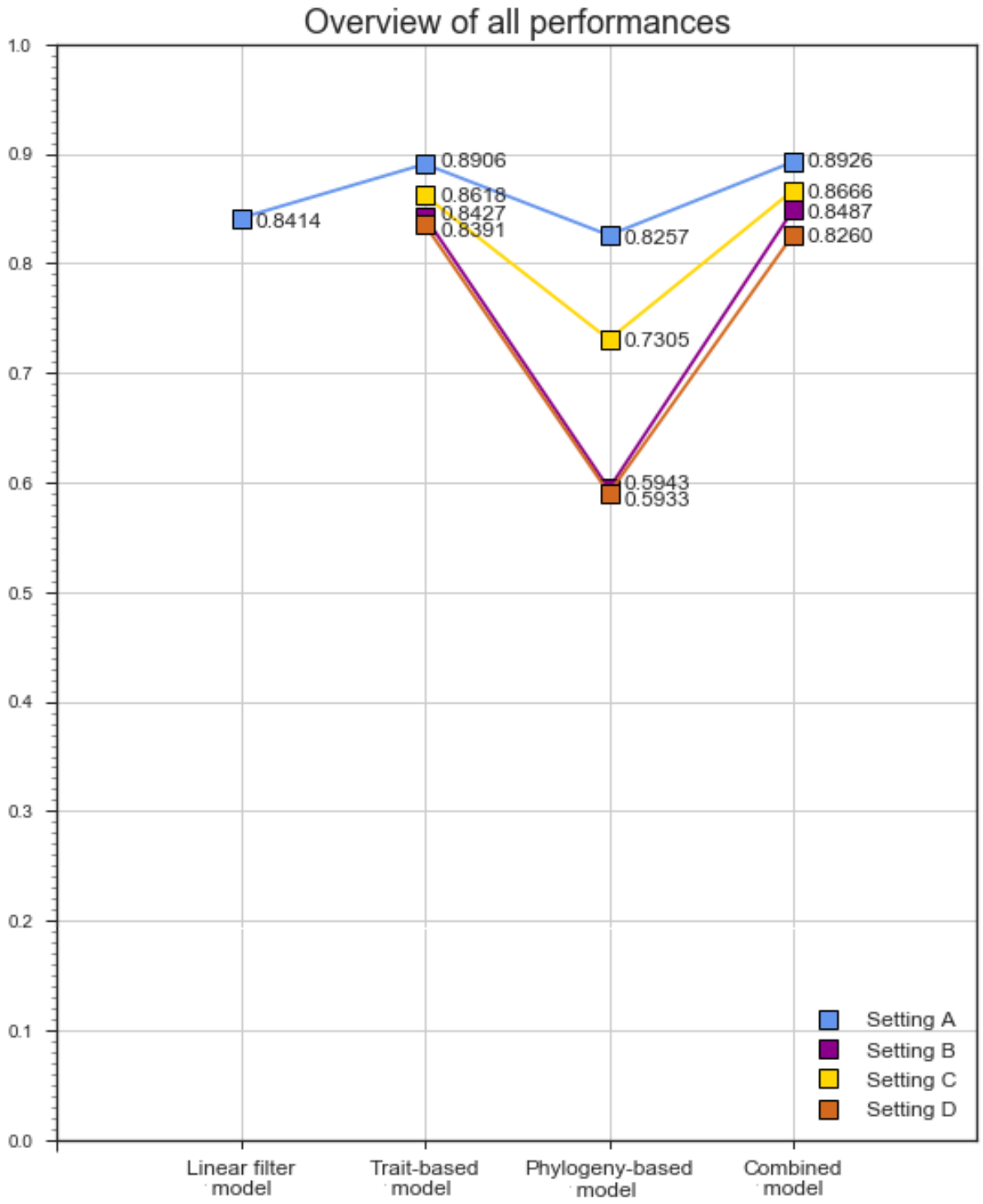


Figure 4.9: Overview of all computed AUC-values. For both the linear filter model as the three TSKRR-models, the most honest performance estimate is taken. All four models are based on the same species subset of 96 pollinators and 193 plants.

CHAPTER 5

OPTIMAL TRANSPORT

This last chapter is also centered around the main dataset of this work, but is a smaller one and does not focus on prediction specifically. It treats a more mathematical view on interaction networks and interaction behaviour. Up till now, we have a dataset of interactions occurring in nature and optimal transport theory providing us formulas to define the optimal partition of interactions. We could hence use the provided binary matrix as the optimal transport cost matrix and calculate the partitions for known plant/pollinator distributions. For these distributions external datasets will be necessary, but this will follow.

To begin with a short recall of the theory, the calculation of the most optimal division of plants over pollinators will always be done using the Sinkhorn algorithm (of which a full explanation was given in Section 2.4). Therefore a distribution of plants, a distribution of pollinators and a specific cost matrix are necessary. The cost matrix was the opposite of the species' preference matrix ($cost = -pref$), determining how 'unlikely' the interactions are. The cost matrix can be binary (containing just two numbers) or quantitative. In the binary case, the exact value of these numbers does not matter, as the algorithm scales the parameters and still generates the same optimal distribution. The values can e.g. be 0 and -1, or 1 and -1. The quantitative case, on the other hand, gives a weight to the cost of each possible interaction. The obtained P^* defines the optimal interaction behaviour of a set of pollinators.

5.1 Toy experiments

First, some simple experiments with binary cost matrices are performed as an introduction. When pollinators like *Amegilla garrula* or *Megachile lefebvrei* and plants like *Acanthus spinosus* or *Ophrys aranifera* are considered, the preference matrix will be very sparse as these species all only have one interaction. When on the contrary optimal transport is done with pollinators like *Apis mellifera* or *Anthidium florentinum*, and plants like *Thymbra capitata* or *Calendula arvensis*, the preference matrix will approach a homogeneous matrix with ones, as their interactions are numerous. Next to

$$\begin{bmatrix} 1&0&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&1 \end{bmatrix}$$

Specialized preferences

$$\begin{bmatrix} 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&1&1 \end{bmatrix}$$

Homogeneous preferences

$$\begin{bmatrix} 1&1&0&0&0&0&0&0&0&0\\ 1&1&0&0&0&0&0&0&0&0\\ 0&0&1&1&1&1&0&0&0&0\\ 0&0&1&1&1&1&0&0&0&0\\ 0&0&1&1&1&1&0&0&0&0\\ 0&0&1&1&1&1&0&0&0&0\\ 0&0&0&0&0&0&1&1&1&0\\ 0&0&0&0&0&0&1&1&1&0\\ 0&0&0&0&0&0&1&1&1&0\\ 0&0&0&0&0&0&0&0&0&1 \end{bmatrix}$$

Blocked preferences

$$\begin{bmatrix} 1&0&0&0&0&0&0&0&0&0\\ 1&1&0&0&0&0&0&0&0&0\\ 1&1&1&0&0&0&0&0&0&0\\ 1&1&1&1&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0\\ 1&1&1&1&1&1&0&0&0&0\\ 1&1&1&1&1&1&1&0&0&0\\ 1&1&1&1&1&1&1&1&0&0\\ 1&1&1&1&1&1&1&1&1&0\\ 1&1&1&1&1&1&1&1&1&1 \end{bmatrix}$$

Triangle-formed preferences

Table 5.1: Different preference matrices.

these two extremes, intermediate preference matrices are possible. Two common intermediate interaction patterns in nature are a block-type and a triangle-type. These are clearly based on the nestedness and modularity properties of a mutualistic network (see Section 1.2). The four types are illustrated in Table 5.1 for a set of ten plants and ten pollinators.

To start, it is assumed that all species have the same abundance, meaning for ten species that each species holds 10% of the total population. Later some experiments are done when a new invasive species would e.g. make up 30% of the population. As stated earlier, it is more convenient to speak of a maximization of preference instead of minimization of distance, hence the sign of the distance is switched so it can now interpreted as a 'satisfaction index'. When λ is high, there is a low contribution of the entropy term to the Sinkhorn distance, or the overall satisfaction. The generated value is not sensitive to the fact if the distribution $P*$ is even or if each pollinator just visits its favourite plant(s). When on the contrary λ is low, entropy becomes more important. The amount of evenness in the outcome will greatly influence the generated value. Of course, a high overall satisfaction is desired. It defines how 'happy' the species are with this partition. The first eight experiments were done. A plot of the outcomes and the corresponding dimensionless satisfaction is shown in Table 5.2. Always (except for the homogeneous one because the partition is already even), the overall satisfaction is higher (better) for a high λ. This is logical as when λ is low, the

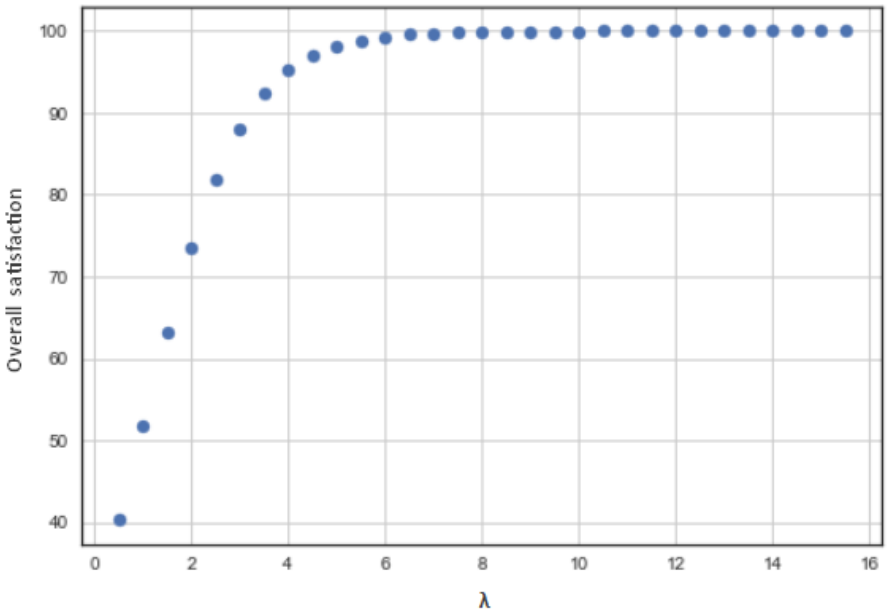


Figure 5.1: Graph of the overall satisfaction in function of λ, based on the block preference matrix and two even species distributions.

optimal distribution will mandatory shift to a more even distribution. This matches the species' preferences less than when they would be free to choose their interaction partners, decreasing their overall satisfaction. However, a maximum (or higher) entropy can still be seen as a beneficial property of the ecological network, as stated in Section 2.4.

An extra graph is plotted in Figure 5.1 to show the influence of λ. The graph is made based on the block preference matrix. Here again, the maximization of preference is used so the y-axis should be interpreted as a satisfaction index. Higher numbers of overall satisfaction are preferred, but also distributions with lower values but high entropy could be desired. The x-axis contains increasing λ-values. As expected the curve goes up as λ increases, meaning an always lower contribution of the entropy term in the distance calculation.

When an invasive species is added to these example experiments (e.g. $plant_6$ now has an abundance of 38% in the network instead of its original 10%), all optimal distributions change accordingly. Not all eight barplots are shown, but the one of the triangle preference matrix, with a λ of 10 is given in Figure 5.2. The overall satisfaction of this distribution is 83.940, which is approximately 10 units lower than the corresponding overall satisfaction in the case of a homogeneous plant distribution. This means that with one highly abundant species, the cost of the generated pollination network is higher, hence decreasing the overall satisfaction of the species.
The graph can easily be explained by looking at the triangle matrix. Only pollinators

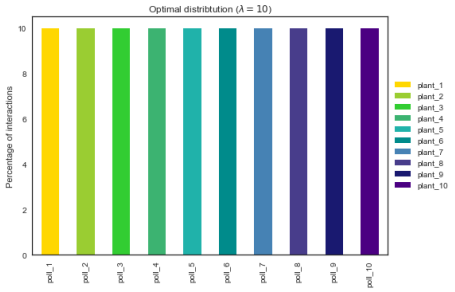
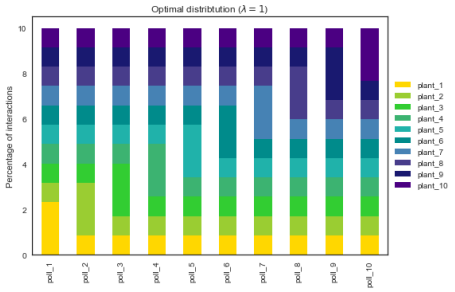
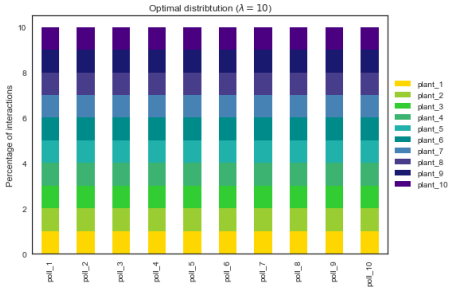
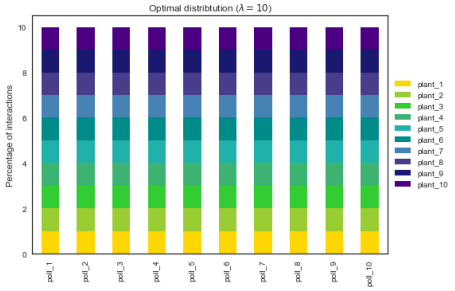
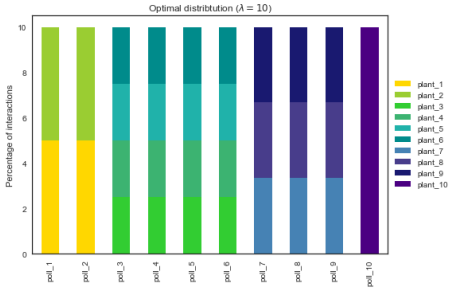
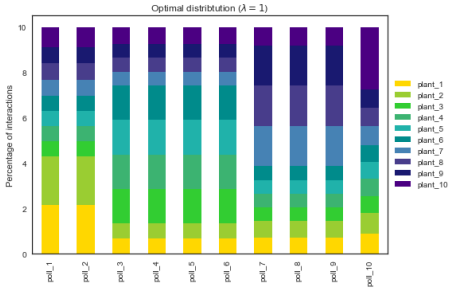
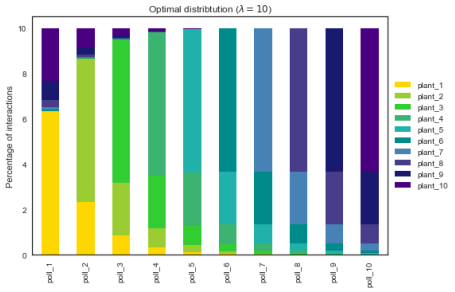
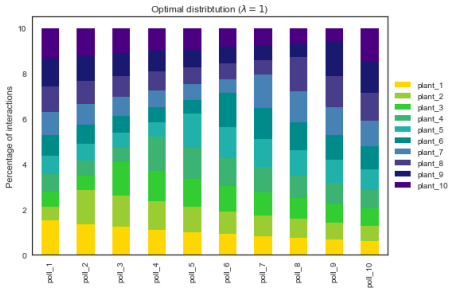
	High lambda, i.e. small entropy contribution	Low lambda, i.e. large entropy contribution
Specialized	 Overall satisfaction = 99.959	 Overall satisfaction = 23.197
Homogeneous	 Overall satisfaction = 100.000	 Overall satisfaction = 100.000
Blocks	 Overall satisfaction = 99.987	 Overall satisfaction = 51.717
Triangle	 Overall satisfaction = 94.185	 Overall satisfaction = 63.114

Table 5.2: Optimal transport partitions for the four possible preference matrices, for a uniform distribution of both plants and pollinators and for $\lambda = 10$ and $\lambda = 1$.

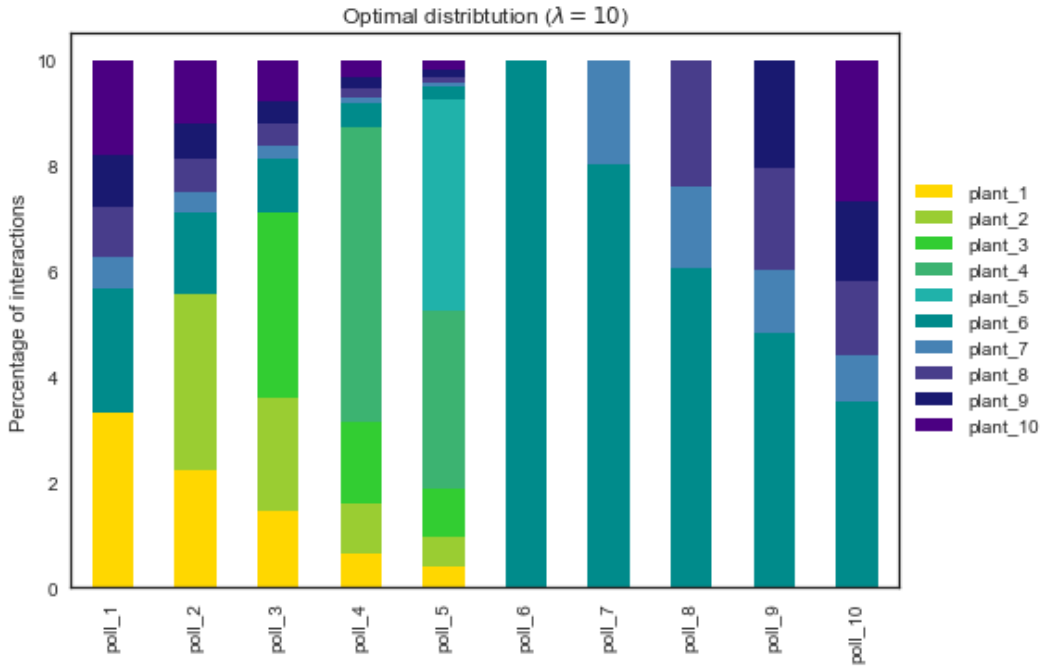


Figure 5.2: Optimal transport partition with one invasive plant species, a triangle-type preference matrix and $\lambda = 10$. The overall satisfaction is 83.940.

6 to 10 had a positive interaction with $plant_6$, meaning that they have the largest portion of visits to this plant. As λ is quite high, not much evenness is induced in the optimal division, meaning that not all pollinators *have* to visit $plant_6$ with the highest fraction of their plant visits. When λ is lowered to one, this would be the case. Then for all pollinators the highest fraction of visit is to $plant_6$, even when the first 5 pollinators did not have a positive interaction with this plant in the preference matrix. In this case, the overall satisfaction goes down remarkably (to 61.269).

Information theory

Also the metrics of Chapter 1 (Section 1.2) can help to interpret the formed partition matrices. There is focused on the scenarios where the preference matrix is perfectly specialized and perfectly homogeneous, because these are best to illustrate the behaviour of the metrics. As in Chapter 3: the row species (i.e. the pollinators (bees)) denote variable B, the column species (i.e. the plants) denote variable P. Always, a heatmap of the optimal partition matrix P^* is added as visualisation of the calculated values. In Table 5.3 the four combinations of a uniform species distribution and one dominant species are shown. The introduction of one/more dominant species can immediately be seen at the rising ΔU, denoting the deviation of two uniform marginal distributions.

With the specialized preference matrix, all conditional entropies stay quite low; the choice of an interaction partner is rather limited. Therefore the occurring interactions are more efficient (which can be seen from the high mutual information), but

the network is not very stable (there is a low variance of information). When a more abundant species is present, the conditional entropies increase. So does the variance of information (which is the sum of both) and hence the overall network stability. The reason is clear to see in the heatmaps: there are now more possible interactions which can effectuate this stability. Yet, as the trade-off described, the efficiency per interaction goes down.
For the homogeneous preference matrix, P^* behaves in a different way. The conditional entropies do not change as much as in the previous case when adding an abundant species. The uncertainty over the interaction partner when a species is known always stays high. This can also be seen in the plots: almost all possible species pairs can interact with each other. The conditional entropy values are not only quite stable but also high (meaning the network is very stable), but the interactions are not efficient. When every species reacts completely generalistic and no specific interactions occur, it is intuitively logical that those interactions are less efficient than in the first case.

When λ is changed to 1, only the first four matrices change (as was already known from Table 5.2), so the homogeneous preference case behaves as expressed above. In the specialized preference case, the conditional entropies increase when lowering λ, the variance of information of course increases too, and the mutual information shrinks. Also this is logical, as more evenness induces more possible interactions. The uncertainty over the interaction partners rises and the network becomes more stable. This was in fact the initial aim of introducing evenness, as stated in the introduction of the entropy term with weight $\frac{1}{\lambda}$.

5.2 Optimal transport on the pollination dataset

After these demonstrations of the optimal transport principle and the analysis of the behaviour of P^*, the toy matrices can be replaced by real life data. To define optimal partitions, we will use the original matrix Y as a preference matrix but we still need abundances of both plant and pollinator species. For this, the weighted files of Web of Life can again be used. However, the same issue as in Section 4.1 occurs: there are very few similar species in the other pollination files. Only four weighted files contained corresponding interaction pairs. Of them one file is chosen and the marginal distributions are taken as species abundances. A subset of Y is produced, based on these species. The used abundances are shown in Figure 5.3. The optimal transport visualization in Figure 5.4.

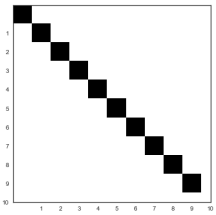
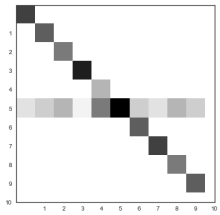
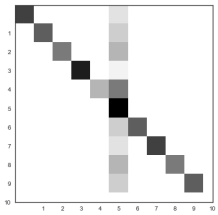
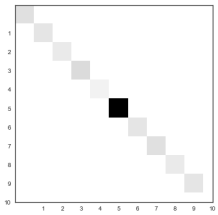
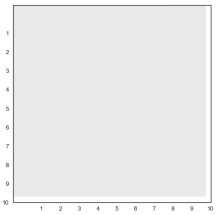
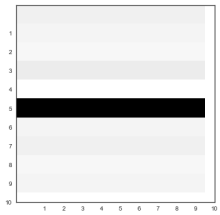
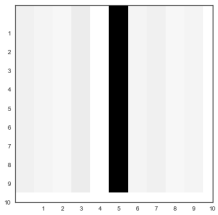
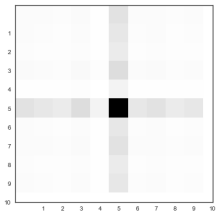
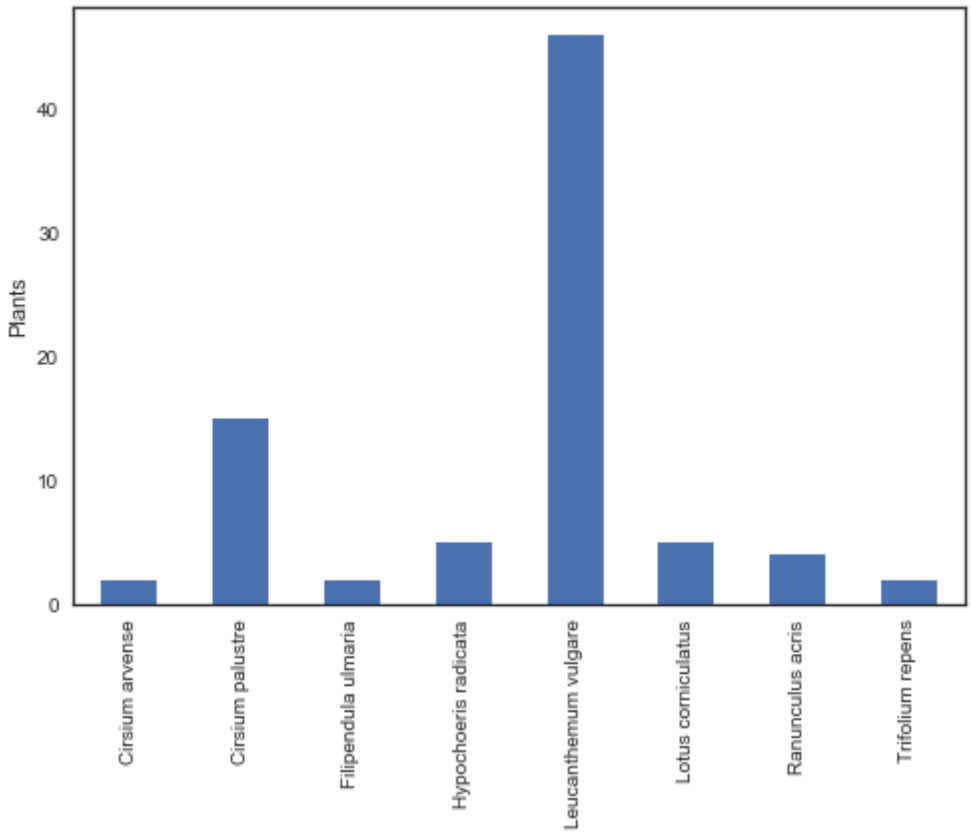
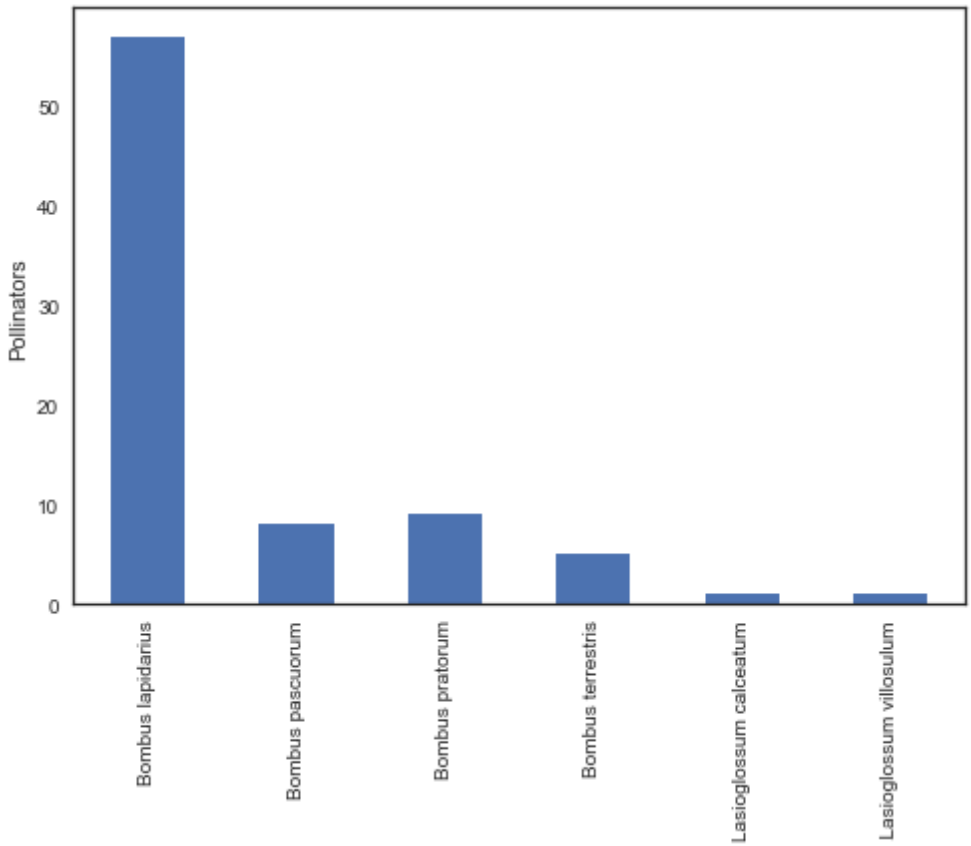
	Uniform poll. distr. Uniform plant distr.	Uniform poll. distr. Dominant plant	Dominant poll. Uniform plant distr.	Dominant poll. Dominant plant
Specialized				
	$H(B) = 3.32$ $H(P) = 3.32$ $H(B,P) = 3.33$ $H(B\|P) = 0.01$ $H(P\|B) = 0.01$ $VI(B;P) = 0.01$ $MI(B;P) = 3.32$ $\Delta U = 0.00$	$H(B) = 3.32$ $H(P) = 2.90$ $H(B,P) = 4.07$ $H(B\|P) = 1.17$ $H(P\|B) = 0.75$ $VI(B;P) = 1.92$ $MI(B;P) = 2.15$ $\Delta U = 0.42$	$H(B) = 2.90$ $H(P) = 3.32$ $H(B,P) = 4.07$ $H(B\|P) = 0.75$ $H(P\|B) = 1.17$ $VI(B;\|P) = 1.92$ $MI(B;P) = 2.15$ $\Delta U = 0.42$	$H(B) = 2.90$ $H(P) = 2.90$ $H(B,P) = 2.91$ $H(B\|P) = 0.01$ $H(P\|B) = 0.01$ $VI(B;P) = 0.01$ $MI(B;P) = 2.90$ $\Delta U = 0.84$
Homogeneous				
	$H(B) = 3.32$ $H(P) = 3.32$ $H(B,P) = 6.64$ $H(B\|P) = 3.32$ $H(P\|B) = 3.32$ $VI(B;P) = 6.64$ $MI(B;P) = 0.00$ $\Delta U = 0.00$	$H(B) = 3.32$ $H(P) = 2.90$ $H(B,P) = 6.23$ $H(B\|P) = 3.32$ $H(P\|B) = 2.90$ $VI(B;P) = 6.23$ $MI(B;P) = 0.00$ $\Delta U = 0.42$	$H(B) = 2.90$ $H(P) = 3.32$ $H(B,P) = 6.23$ $H(B\|P) = 2.90$ $H(P\|B) = 3.32$ $VI(B;\|P) = 6.23$ $MI(B;P) = 0.00$ $\Delta U = 0.42$	$H(B) = 2.90$ $H(P) = 2.90$ $H(B,P) = 5.81$ $H(B\|P) = 2.90$ $H(P\|B) = 2.90$ $VI(B;P) = 5.81$ $MI(B;P) = 0.00$ $\Delta U = 0.84$

Table 5.3: Calculated values of the Information theoretic metrics (in bits) on the optimal partition matrix P^*, for $\lambda = 10$. The cost matrices are of the specialized or homogeneous type, the abundances vary from a homogeous distribution to one dominant species and all combinations of them. Note that the color bar for each P^*-heatmap has a different range per matrix, but black is always the highest frequency of visits while white is the lowest.



(a) Plant abundance distribution.



(b) Pollinator abundance distribution.

Figure 5.3: Species distributions taken for optimal transport.

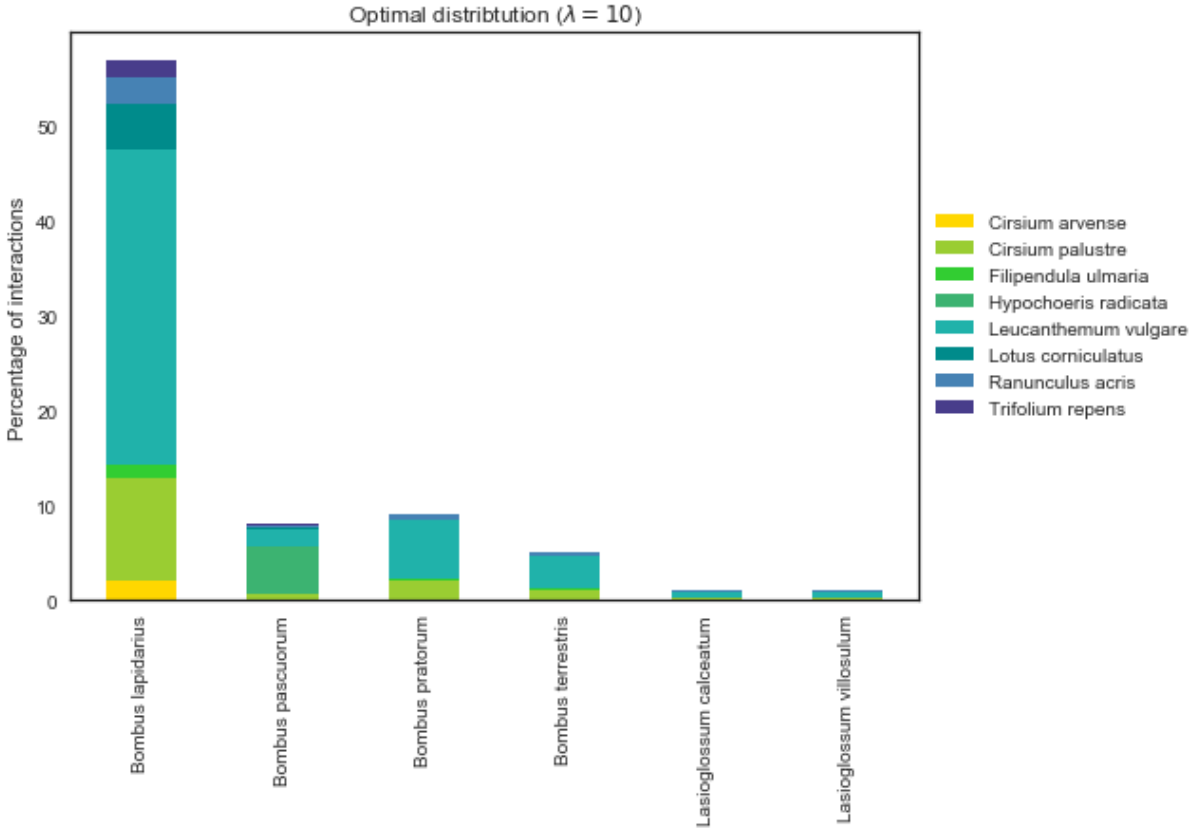


Figure 5.4: Optimal transport partition for a Web of Life dataset (with a $\lambda = 10$).

H_B	2.0200
H_P	1.4434
H_{BP}	3.1820
$H_{B\|P}$	1.7386
$H_{P\|B}$	1.1619
$VI(B;P)$	2.9005
$MI(B;P)$	0.2814
ΔU	2.1216

Table 5.4: Calculated values of the Information theoretic metrics (in bits) on the optimal partition matrix of the real life example.

When the information theoretical metrics are applied to this optimal P*-matrix, we get the values of Table 5.4. As can be seen, ΔU is very big. This is logical as both the pollinator and the plant distribution greatly deviate from a uniform distribution. Twice a dominant species overrules the abundances of the other species. Secondly, the variance of information clearly exceeds the mutual information. This means that the generated partition matrix is stable of nature. $H_{B|P}$ has a bigger contribution to this variance of information than $H_{P|B}$, so mostly the plants effectuate the stability. There is more uncertainty left when the plant species is known, so a plant species has a wider variety of interaction partners to chose from. Another way of comparing these two metrics is as an expected number of binary questions that have to be asked to determine the species, when the interaction partner is known. When the

plant is known, on average 1.73 questions are needed to determine the interacting pollinator. When on the other hand the pollinator is known, on average 1.16 questions are needed to determine the particular plant species the pollinator has visited. This leaves us with the same conclusion. Also, when compared to the statement that a plant-pollinator relation is rather asymmetrical as pollinators are typically more specialized than plants of Morales-Castilla et al. (2015), these are in line. To end with the familiar trade-off, when the stability of the network is high, this comes at the expense of the interactions' efficiency. The mutual information of this network (expressing the information transfer of B to P and vice versa, or the reduction of uncertainty) is low.

This was just a small application of optimal transport, but this theory can offer bigger insights than this. The theory for example can help to discover an underlying binary interaction preference, based on an observed weighted network. When quantitative matrices are assembled by counting visits or by more advanced techniques, the structure of the preference matrix can be determined for different tuning parameters. In this way, by repeating this for a lot of weighted sets with the same species, a consensus can be made for the underlying preferences. By tuning the regularisation parameter λ and changing the plants' and insects' preferences, the best matching interaction matrix can be found. This is a mathematical way of determining the species' most preferred interactions, and will of course be different than just making the weighted matrix binary. Undoubtedly, optimal transport offers many other potential applications.

CONCLUSION

Mathematics, modelling and ecology, three disciplines that can be linked to each other in several ways. Metrics of the Information theory as entropy and its derivatives all have their own ecological interpretation, providing an objective way to compare networks (e.g. in terms of efficiency and stability). Also nestedness and modularity proof to be meaningful measures. The nested structure of mutualistic (plant-pollinator) networks gives them a natural buffer capacity against external disruptions, which can be of great importance in times of climate change.

Next to single metrics, formulas can be used to built models. Models (hence machine learning) can be of use in ecological context because of the way interaction datasets are constructed. A finetuning step for missing values is recommended. Both the linear filter model as the two-step kernel ridge regression have shown to be able to make reasonable predictions for species combinations. Performance metrics (AUC's) of over 80% can be reached. In this way models are able to guide future field research, by targeting interactions with more chance of happening. Prioritization of interactions was the main goal in this work.

Overall, being able to simulate pollination networks and their properties can be helpful for other applications too. Pollination is an extremely valuable phenomenon, at the root of lots of biological interactions and all our feeding habits. Models can, for example, predict the interaction behaviour of invasive species and stability measures can then predict potential effects for the network. Numerous utilizations of prediction models can be examined, but as stated, the focus was on detecting missing values.

BIBLIOGRAPHY

[1] Angiosperm Phylogeny Group. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: Apg ii. *Botanical Journal of the Linnean Society 141*, 4 (2003), 399–436.

[2] Api:Cultural. Types of pollinators. URL = http://www.apicultural.co.uk/types-of-pollinators, consulted on 06-11-2017.

[3] Armbruster, W. S. Patterns of character divergence and the evolution of reproductive ecotypes of Dalechampia scandens (Euphorbiaceae). *Evolution 39*, 4 (1985), 733–752.

[4] Avraham, S., Jiang, S., Ota, S., Fu, Y., Deng, B., Dowler, L., White, R., and Avraham, H. Structural and functional studies of the intracellular tyrosine kinase MATK gene and its translated product. *The Journal of Biological Chemistry 270* (1995), 1833–1842.

[5] Bangerth, K. Floral induction in mature, perennial angiosperm fruit trees: Similarities and discrepancies with annual/biennial plants and the involvement of plant hormones. *Scientia Horticulturae 122*, 2 (2009), 153–163.

[6] Barber, D. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[7] Bastolla, U., Fortuna, M. A., Pascual-Garcia, A., Ferrera, A., Luque, B., and Bascompte, J. The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature 458*, 7241 (2009), 1018.

[8] Beach, J. H., and Bawa, K. S. Role of pollinators in the evolution of dioecy from distyly. *Evolution 34*, 6 (11 1980), 1138–1142.

[9] BOLDSystems. Barcode of life data system. URL = http://www.boldsystems.org, consulted on 21-10-2017.

[10] BugGuide. Identification, images & information. For insects, spiders & their kin. URL = https://bugguide.net/, consulted on 21-08-2017.

[11] Canard, E., Mouquet, N., Marescot, L., Gaston, K. J., Gravel, D., and Mouillot, D. Emergence of structural patterns in neutral trophic networks. *PLoS One 7*, 8 (2012), e38295.

[12] Cook, J. M., and Rasplus, J.-Y. Mutualists with attitude: coevolving fig wasps and figs. *Trends in Ecology & Evolution 18*, 5 (2003), 241 – 248.

[13] Cooper, G. *The Cell: A Molecular Approach. Chloroplasts and Other Plastids*, 2 ed. Sunderland (MA): Sinauer Associates, 2000. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9905/.

[14] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems 26* (06 2013).

[15] Demirel, İ. *The Monge-Kantorovich mass transportation problem*. PhD thesis, Bilkent University, 2017.

[16] Desjardins-Proulx, P., Laigle, I., Poisot, T., and Gravel, D. Ecological interactions and the Netflix problem. *PeerJ 5* (2017), e3644.

[17] Dey, S., North, J. A., Sriram, J., Evans, B. S., and Tabita, F. R. In vivo studies in Rhodospirillum rubrum indicate that ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) catalyzes two obligatorily required and physiologically significant reactions for distinct carbon and sulfur metabolic pathways. *The Journal of Biological Chemistry 290* (2015), 30658–30668.

[18] Eggelte, H., Lid, D. T., and Ebregt, A. *Veldgids Nederlandse Flora*. KNNV, 2014.

[19] Evans, D. M., Kitson, J. J., Lunt, D. H., Straw, N. A., and Pocock, M. J. Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Functional Ecology 30*, 12 (2016), 1904–1916.

[20] Falk, S. J. *Field Guide to the Bees of Great Britain and Ireland*. British Wildlife Publishing, 2015. 'Veldgids Bijen voor Nederland en Vlaanderen' is mostly a translation of this book, with a few adaptations for The Netherlands and Flanders.

[21] Featherly, H. I. *Taxonomic Terminology of the Higher Plants*. Lowa State College Press; Ames, Lowa, 1954.

[22] Fenster, C. B., Armbruster, W. S., Wilson, P., Dudash, M. R., and Thomson, J. D. Pollination syndromes and floral specialization. *Annu. Rev. Ecol. Evol. Syst. 35* (2004), 375–403.

[23] Fortuna, M. A., Stouffer, D. B., Olesen, J. M., Jordano, P., Mouillot, D., Krasnov, B. R., Poulin, R., and Bascompte, J. Nestedness versus modularity in ecological networks: two sides of the same coin? *Journal of Animal Ecology 79*, 4 (2010), 811–817.

[24] García-Horsman, J. A., Barquera, B., Rumbley, J., Ma, J., and Gennis, R. B. The superfamily of heme-copper respiratory oxidases. *Journal of Bacteriology 176*, 18 (1994), 5587.

[25] GeneCards; Human Gene Database. Megakaryocyte-associated Tyrosine Kinase. URL = http://www.genecards.org/cgi-bin/carddisp.pl?gene=MATK, consulted on 15-10-2017.

[26] Gombault, C., Morison, N., Guilbaud, L., and Vaissière, B. E. FlorAbeilles: Base de données en ligne sur les interactions plantes-abeilles en France métropolitaine. *INRA, Unité Abeilles et Environnement and Laboratoire de Pollinisation et Ecologie des Abeilles, Avignon, France* (2018).

[27] Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics* (1971), 857–871.

[28] Granek, E. F. An analysis of Pteropus livingstonii roost habitat: Indicators for forest conservation on Anjouan and Moheli. *Tri News* (2000).

[29] Hall, B. G. Building phylogenetic trees from molecular data with MEGA. *Molecular Biology & Evolution 30*, 5 (2013), 1229–1235.

[30] Hegland, S. J., Nielsen, A., Lázaro, A., Bjerknes, A.-L., and Totland, Ø. How does climate warming affect plant-pollinator interactions? *Ecology Letters 12*, 2 (2009), 184–195.

[31] Heimans, E., Heinsius, H. W., and Thijsse, J. P. *Geïllustreerde Flora van Nederland*. 22. Versluys-Amsterdam, 1994.

[32] Hofmann, T. Probabilistic latent semantic indexing. *SIGIR Forum 51*, 2 (2017), 211–218.

[33] Iwasa, Y., Ezoe, H., and Yamauchi, A. Evolutionarily stable seasonal timing of univoltine and bivoltine insects. *Series Entomologica 52* (1994), 69–89.

[34] James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning*, vol. 112. Springer, 2013.

[35] Klein, A. What is an intuitive explanation of the concept of entropy in information theory? URL = https://www.quora.com/What-is-an-intuitive-explanation-of-the-concept-of-entropy-in-information-theory, consulted on 03-04-2018.

[36] Kullback, S. *Information theory and statistics*. Courier Corporation, 1997.

[37] Levy, B., and Schwindt, E. Notions of optimal transport theory and how to implement them on a computer. *arXiv preprint arXiv:1710.02634* (2017).

[38] Liu, C., Berry, P. M., Dawson, T. P., and Pearson, R. G. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography 28*, 3 (2005), 385–393.

[39] Magrach, A., González-Varo, J. P., Boiffier, M., Vilà, M., and Bartomeus, I. Honeybee spillover reshuffles pollinator diets and affects plant reproductive success. *Nature Ecology & Evolution 1*, 9 (2017), 1299.

[40] Morales-Castilla, I., Matias, M. G., Gravel, D., and Araújo, M. B. Inferring biotic interactions from proxies. *Trends in Ecology & Evolution 30*, 6 (2015), 347–356.

[41] Newman, M. E. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103*, 23 (2006), 8577–8582.

[42] Olmstead, R. G., Michaels, H. J., Scott, K. M., and Palmer, J. D. Monophyly of the asteridae and identification of their major lineages inferred from DNA sequences of rbcL. *Annals of the Missouri Botanical Garden 79*, 2 (1992), 249–265.

[43] Olsen, K. M. Pollination effectiveness and pollinator importance in a population of Heterotheca subaxillaris (Asteraceae). *Oecologia 109*, 1 (1996), 114–121.

[44] Pérez-Mellado, V., and Casas, J. L. Pollination by a lizard on a Mediterranean island. *Copeia 1997*, 3 (1997), 593–595.

[45] Petanidou, T. *Pollination Ecology in a Phryganic Ecosystem*. PhD. Thesis, Aristotelian University, Thessaloniki, 1991.

[46] Poisot, T., Cirtwill, A. R., Cazelles, K., Gravel, D., Fortin, M.-J., and Stouffer, D. B. The structure of probabilistic networks. *Methods in Ecology & Evolution 7*, 3 (2016), 303–312.

[47] Provost, F., and Kohavi, R. Guest editors' introduction: On applied research in machine learning. *Machine Learning 30*, 2 (1998), 127–132.

[48] Pulliam, H. R. On the theory of optimal diets. *The American Naturalist 108*, 959 (1974), 59–74.

[49] Rafferty, N. E., and Ives, A. R. Phylogenetic trait-based analyses of ecological networks. *Ecology 94*, 10 (2013), 2321–2333.

[50] Refaeilzadeh, P., Tang, L., and Liu, H. Cross-validation. In *Encyclopedia of Database Systems*. Springer, 2009, pp. 532–538.

[51] Rutledge, R. W., Basore, B. L., and Mulholland, R. J. Ecological stability: an information theory viewpoint. *Journal of Theoretical Biology 57*, 2 (1976), 355–371.

[52] Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development 3*, 3 (1959), 210–229.

[53] Santamaría, L., and Rodríguez-Gironés, M. A. Linkage rules for plant–pollinator networks: trait complementarity or exploitation barriers? *PLoS Biology 5*, 2 (2007), e31.

[54] Schmid-Hempel, P., Kacelnik, A., and Houston, A. Honeybees maximize efficiency by not filling their crop. *Behavioral Ecology and Sociobiology 17* (1985), 61.

[55] Schölkopf, B., and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. MIT press, 2002.

[56] Schrynemackers, M., Küffner, R., and Geurts, P. On protocols and measures for the validation of supervised methods for the inference of biological networks. *Frontiers in Genetics 4* (2013), 262.

[57] Shawe-Taylor, J., and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[58] Stock, M., De Baets, B., and Waegeman, W. *Exact and Efficient Algorithms for Pairwise Learning*. Ghent University PhD, 2017.

[59] Stock, M., Pahikkala, T., Airola, A., Waegeman, W., and De Baets, B. Algebraic shortcuts for leave-one-out cross-validation in supervised network inference. *bioRxiv* (2018), 242321.

[60] Su, X., and Khoshgoftaar, T. M. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence 2009* (2009), 4.

[61] USDA (United States Departement of Agriculture) and NRCS (Natural Resources Conservation Service). Plants database. URL = https://plants.usda.gov/java/, consulted on 15-10-2017.

[62] Van Peer, G., De Paepe, A., Stock, M., Anckaert, J., Volders, P.-J., Vandesompele, J., De Baets, B., and Waegeman, W. miSTAR: miRNA target prediction through modeling quantitative and qualitative miRNA binding site information in a stacked model structure. *Nucleic Acids Research 45*, 7 (2017), e51–e51.

[63] Vázquez, D. P., Chacoff, N. P., and Cagnolo, L. Evaluating multiple determinants of the structure of plant–animal mutualistic networks. *Ecology 90*, 8 (2009), 2039–2046.

[64] Waegeman, W., Stock, M., Dhiedt, E., Hoebeke, L., Puynen, S., and Vanwyck, T. *Hoeveel informatie zit er in een ecologisch netwerk?* Ghent University Bachelorproject, 2016.

[65] Wall, M. E., Rechtsteiner, A., and Rocha, L. M. Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*. Springer, 2003, pp. 91–109.

[66] WEB OF LIFE. Ecological networks database. URL = http://www.web-of-life.es/, consulted on 11-03-2018.

[67] Welling, M. Kernel ridge regression. *Max Welling's Classnotes in Machine Learning* (2013), 1–3.

[68] Whitfield, C. W., Behura, S. K., Berlocher, S. H., Clark, A. G., Johnston, J. S., Sheppard, W. S., Smith, D. R., Suarez, A. V., Weaver, D., and Tsutsui, N. D. Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera. *Science 314*, 5799 (2006), 642–645.

[69] Wilson, A. G. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy 3*, 1 (1969), 108–126.