Inferring social network structure in ecological systems from spatio-temporal data streams

Ioannis Psorakis *[1,2], Stephen J. Roberts[1], Iead Rezek[1], and Ben C. Sheldon[2]

[1]Pattern Analysis and Machine Learning Research Group, University of Oxford
[2]Edward Grey Institute, University of Oxford

Abstract

We propose a methodology for extracting social network structure from spatio-temporal datasets that describe timestamped occurrences of individuals. Our approach identifies temporal regions of dense agent activity and links are drawn between individuals based on their co-occurrences across these "gathering events". The statistical significance of these connections is then tested against an appropriate null model. Such framework allows us to exploit the wealth of analytical and computational tools of network analysis in settings where the underlying connectivity pattern between interacting agents (commonly termed the *adjacency matrix*) is not given a priori. We perform experiments on two large scale datasets ($> 10^6$ points) of great tit *Parus major* wild bird foraging records and illustrate the use of this approach by examining the temporal dynamics of pairing behaviour, a process that was previously very hard to observe. We show that established pair bonds are maintained continuously, whereas new pair bonds form at variable times before breeding, but are characterised by a rapid development of network proximity. The method proposed here is general, and can be applied to any system with information about the temporal co-occurrence of interacting agents.

Keywords: network analysis, spatio-temporal data streams, animal social networks

1 Introduction

We use the terms *graph* or *network* to describe the simplified version of the pattern of interactions in a system, such as an animal population, where nodes are individual entities and edges represent some form of association, interaction, similarity or behavioural correlation between nodes. In the same way that a map is a simplified (though useful) version of a landscape, a network describes the *topology* of a real-world system by focusing on the connectivity patterns of its individual components [1].

The key motivation for employing network analysis tools is that the web of interconnections between individuals can provide us insights on the underlying mechanisms that govern the system under study [2]. For example in an ecological context, the position and role of animals in the network may have important fitness consequences [3] both for the individual and the population as a whole [4]. Additionally, the network paradigm gives us the flexibility to look at the system at various resolutions and model any type of interaction; sexual, cooperative, competitive, etc [4].

Despite the advantages of the network paradigm and the wealth of analytical and computational tools for network analysis [5–8], the problem of capturing any given system as a graph is not always trivial. Not all systems possess an obvious "web-like" structure (such as the Internet), where the interconnections between participating entities are apparent from direct observation (computers that are connected through physical cables). Additionally, collected data (from field studies, sensor observations, world-wide web etc) may not capture the associations between the observed agents, thus no relational structure can be directly defined. For example, in systems such as animal populations the underlying network of social affiliations needs to be inferred through proxies such as the behaviour (mobility patterns, foraging habits etc) of individual animals.

This work focuses on the problem of finding the *underlying social network structure* of a population that can only be observed through the spatial trajectories of its individual members. We use as a case study a setting where individual wild birds are marked with transponder devices and through appropriate logging hardware we are able to identify their position at various sites in their natural habitat. The observation data collected in this manner consists of a long stream of timestamped records where no obvious interaction or social affiliation is apparent. By assuming that *social structure* is a latent factor that affects the way birds visit locations (in the sense that socially affiliated individuals have similar mobility patterns), we propose a methodology that extracts a *social network* from such spatio-temporal data stream. Although we demonstrate our method in an ecological context, our approach can be generalised to any setting where agents perform timestamped "check-ins" at various locations.

---

[*] ioannis.psorakis@eng.ox.ac.uk

The paper is organised as follows; in Section 2 we outline our experiment settings and discuss our data format. In Section 3 we present our contribution, which is a methodology for extracting network structure from timestamped observation data. In Section 4 we apply our method to the wild bird dataset and show that the extracted networks reflect actual processes that take place in the population, by focusing on mating pair formation. We conclude this paper in Section 5 by discussing the next steps of our research, both in terms of method development and data collection extensions. The MATLAB code that implements the methods presented in the paper is made available[1] to the community.

2 Data collection

This work lies within the context of a large ongoing study of the great tit (GT) *Parus major* population at Wytham Woods near Oxford, United Kingdom. Thousands of individual birds are marked with transponders and a grid of sensor-enabled locations generates hundreds of thousands of records each winter. At each one of the sixty-seven locations in the forest, there is a feeder that acts as an attraction point for foraging individuals. By placing appropriate logging hardware at the feeder, we are able to record the presence of each individual bird. Due to equipment constraints, there were only 16 loggers available at any time, and these were thus rotated around the 67 locations following a structured randomised design, so that each of 8 approximately equally-sized sections of the site always had two active loggers in it. More details on our experiment set-up is provided in the Electronic Supplementary Material (ESM).

Table 1: Sample format of our data

Bird ID	timestamp	location ID
N199642	1/9/2007 10:02:15 (am)	1a
TE80535	1/9/2007 10:02:30 (am)	1a
V260952	1/9/2007 10:02:30 (am)	2b
V260952	1/9/2007 10:02:45 (am)	2b
N199642	1/9/2007 10:12:15 (am)	1c
...	...	...

The data generated from this scheme consists of a long stream of timestamped observations as shown in Table 1. Each row represents a single record that captures the ID of the bird along with the time and location where the foraging event took place. In this format, shown in Table 1, our data stream is only a transactions table in a relational database context, which restricts our analysis to a handful of relatively simple counting operations such as finding the total appearances of a given bird, total birds that visited a specific feeder, etc.

What we are interested in is to find an appropriate mapping of this spatio-temporal stream to a relational space, where social affiliations between individuals are revealed by the similarity of their feeder visitation patterns. We seek to characterise the overall social network of the population of marked birds, and explore the ability of this approach to recover relationships between mated pairs of individuals observed independently during breeding season data collection. We further wish to explore the temporal dynamics of the formation of mated pairs. In biological terms, the process by which pairs of individuals develop relationships that lead to mating is poorly understood in most natural populations, since the majority of work involves observations of pairs at the time of breeding, after pair formation has occurred. As a consequence, we have little knowledge of when such relationships form, and when they become distinguishable from other social relationships between individuals.

In the following section we introduce a method, based on the above goals, that extracts network structure given such spatio-temporal data. In Section 4 we present the application of this approach to the *Parus Major* data set.

3 Network inference from spatio-temporal data

3.1 The time-window problem

A typical approach for building a network from data such as the ones we presented in Section 2, would involve discretising the stream using a fixed *aggregation* or *time window* Δt and assuming that if two individuals are recorded within an interval Δt then there is a link between them in the network [9–12, 20]. The most obvious problem with this approach is that of finding the appropriate size for the time window. An inappropriately small Δt may lead to a network that does not capture important connections, while a very large Δt would overload the graph with "junk" links.

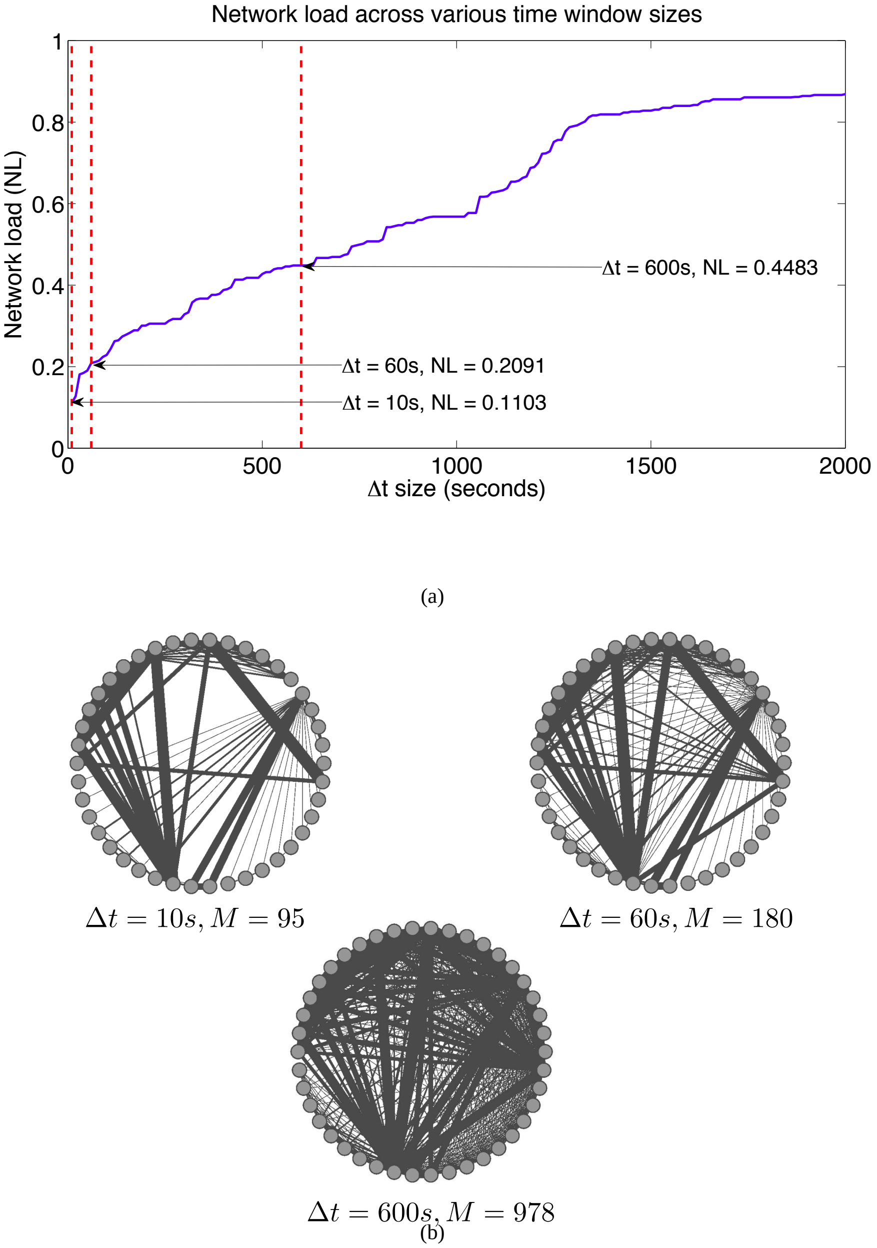
Using our wild-bird data as an example, we take a single day's worth of observations (in a format similar to the one shown in Table 1) and split that stream into time intervals of size Δt. We then place links between the N individual birds (nodes) based on the number of times they were recorded within a temporal distance of Δt. We seek to examine the changes that take place in the network as we vary the time window size by monitoring the *network load* (NL), which is the

---

[1] `http://www.robots.ox.ac.uk/~parg/software.html`

fraction of M links in the network over all possible pair combinations $\frac{1}{2}(N^2 - N)$ of N nodes[2]. We can see in Fig. 1(a) that NL increases along with the size of Δt, because more links are placed between nodes. An example of how network topology changes for various selections of time window size is is shown in Fig. 1(b), while Krings *et al.* in [20] have performed similar experimentation considering more network metrics such as average degree, average weight, clustering coefficients, etc.

Between all these different network topologies that result from varying Δt, there is no direct way of showing which one is the most appropriate. Additionally, even if we had some prior knowledge on the appropriate time window size or even a specific quality function for finding its optimum value, we have still made the strong assumption that Δt is *fixed* throughout the data stream. This corresponds to the belief that the "interaction radius" between individuals is *constant* across our observation period and is not affected by temporal changes in the overall system.



(a)

(b)

Figure 1: In Fig. 1(a) we plot the network load for various time window sizes, spanning from 10 seconds to half an hour. We can see that especially for early increases of Δt there is a large inclusion of links in the network. We also mark three cases of different time window sizes (dashed vertical line) and show in Fig. 1(b) how the graph topology changes based on the Δt value.

In the current work we shall pursue a different approach for building networks from spatio-temporal records, which

---

[2] In this example and throughout this paper we are considering networks that are *undirected* with nodes that have no self-edges.

exploits the inhomogeneous *density profile* of our data stream thus avoiding schemes such as multiple runs [20] in order to select an appropriate Δt. This methodology, which we will call GMMEvents (Gaussian mixture model for event streams) is complemented with an appropriate null model that allows us to distinguish between links that denote social tie and the ones that result from coincidence.

3.2 Identification of gathering events

Let our spatio-temporal data $\mathcal{D}$, a sample of which we showed in Table 1, to be represented in the form $\mathcal{D} = \{b_z, t_z, \ell_z\}_{z=1}^{Z}$ where Z is the total number of records or *tuples* in our database (e.g. the number of rows of Table 1). If we take a single tuple $\{b_z, t_z, \ell_z\}$, we read it as "the bird b_z appeared at time t_z at the feeding location ℓ_z". Note that $\{t_z\}_{z=1}^{Z}$ denotes *event time*, therefore for every timestamp t_z there is a corresponding bird appearance b_z. Additionally, given a specific bird i out of total N birds, there can be many records z for which $b_z = i$, as a single individual may appear many times in the data. Our goal is to find an appropriate mapping from the stream $\mathcal{D}$ to an *adjacency matrix* $\mathbf{A} \in \mathbb{R}^{N \times N}$, where $a_{ij} \neq 0$ denotes a link between birds i and j. To keep the notation uncluttered, from now on we will focus on the case of a single location and show later that results can be easily generalised to the multi-site case.

Consider the plot of Fig. 2, which illustrates how bird arrivals at a particular feeding location are spread throughout a small sample of our observation timeline. Each stem represents an actual sensor capture of a specific bird b_z at time t_z. We can see that the records are not uniformly spread across time, but they are "packed" in small observation-dense regions. Indeed, if we take the whole data stream and extract the histogram of the time differences $\delta(t_z) = t_z - t_{z-1}$ between every pair of consecutive observations, as seen in Fig. 3, we find a broad power-law tail with exponent $\simeq 2.5$ for $\delta(t_z) > 10^3$. This non-Poissonian decay of inter-record timestamps, along with the fact that most $\delta(t_z)$ take small values, implies that the observation profile is comprised of temporally-focused bursts of recording activity, which can be seen as *flocks of foraging individuals*.

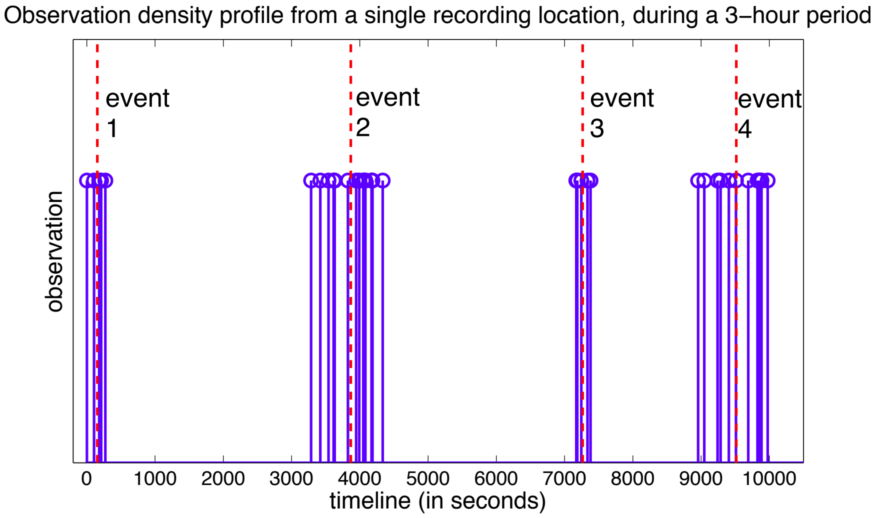


Figure 2: We plot bird arrivals as recorded at a specific location over the course of 3-hour period. We can see that the visitation profile is temporally focused, consisting of bursts of bird activity. Our goal is to identify such regions of increased observation density and examine which individuals participate in these gathering events.

Our main hypothesis is that birds not only visit the feeder as part of such small flocks but also have a *preference* to the members of the flock they choose to forage with. Such regions of increased observation density, can be viewed as K *gathering events* of socially affiliated birds. We seek to *cluster* our Z observations in a way such that closely-appearing individuals, based on their arrival timestamp t_z, are assigned to the same gathering event k.

We perform this clustering scheme using a Gaussian mixture model, with an appropriate configuration that allows us to automatically infer the effective number K of events/clusters (see ESM). The result is described by an observation-to-cluster *responsibility matrix* $\mathbf{\Gamma} \in \mathbb{R}^{Z \times K}$, where Z is the total number of bird observations, K the number of clusters and the elements $\{\gamma_{z1}, \gamma_{z2}, ..., \gamma_{zK}\}$ of each row denote a membership score of a single observation z to an event k.

As there is a one-to-many correspondence between a given bird i and timestamps t_z, a single bird can be recorded many times in the observation stream or, in other words, there are many tuples $\{t_z, b_z\}$ for which $b_z = i$. Therefore, we seek to map the observation-to-cluster matrix $\mathbf{\Gamma} \in \mathbb{R}^{Z \times K}$ to a bird-to-cluster matrix $\mathbf{B} \in \mathbb{R}^{N \times K}$. We start by taking each row $\boldsymbol{\gamma}_z = \{\gamma_{zk}\}_{k=1}^{K}$ of $\mathbf{\Gamma}$ and set the largest element to 1 and all the others to zero. This allows us to describe participation scores γ_{zk}, and all the other measures we derive from them, as integer valued *occurrences*. For each individual bird $i \in \{1, ..., N\}$, we identify the subset $\mathcal{Z}_i$ of rows $\boldsymbol{\gamma}_z$ of $\mathbf{\Gamma}$ that correspond to observations regarding i. We thus set each

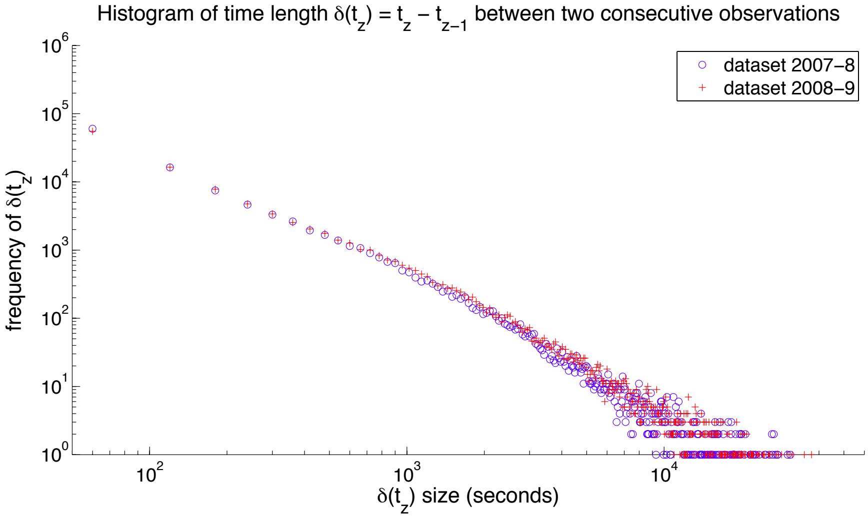
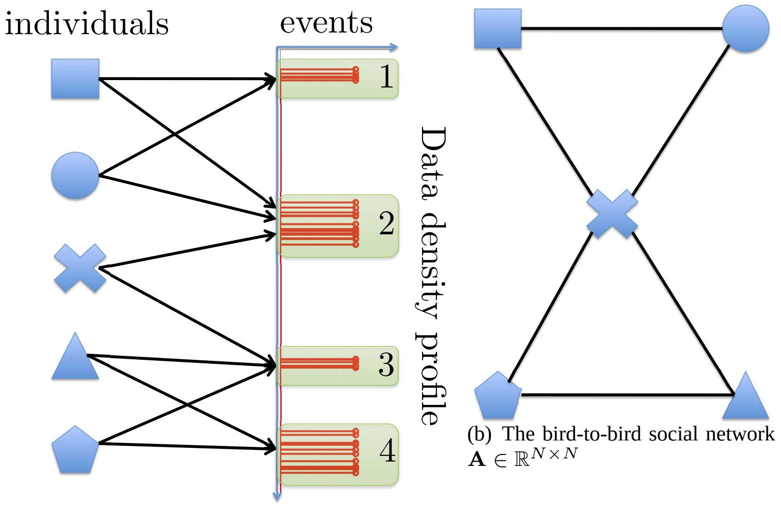


Figure 3: We calculate the time difference $\delta(t_z) = t_z - t_{z-1}$ between every pair of consecutive observations at each location in our two data streams (seasons 2007-8 and 2008-9) and plot the histogram of those values on a logarithmic scale. The $\delta(t_z)$ that refer to pairs where $z-1$ is the last observation of day $d-1$ and z the first observation of day d have been ommitted, in order avoid bias in the results (there is no bird feeding activity during night time).

row $\mathbf{b}_i$ of $\mathbf{B}$ as the sum $\mathbf{b}_i = \sum_{z \in \mathcal{Z}_i} \gamma_z$. The resulting matrix $\mathbf{B} \in \mathbb{R}^{N \times K}$ can be seen as a representation of a *bipartite* or *two-mode network*, that is a graph with two types of nodes; N birds and K events, as shown in Fig. 4(a). Each element b_{ik} denotes the number of times each bird was observed at a specific foraging group.



(a) The bird-to-event bipartite network $\mathbf{B} \in \mathbb{R}^{N \times K}$

(b) The bird-to-bird social network $\mathbf{A} \in \mathbb{R}^{N \times N}$

Figure 4: Our method identifies gathering events from the bursts in our observation stream as seen in Fig. 4(a). Then individuals are assigned to such events creating a bipartite network. In Fig. 4(b) we recover the bird-to-bird social network, via an appropriate *one-mode projection*, based on the co-participation of individuals to these events.

3.3 Building the social network

The bipartite network we extracted in Section 3.2 and shown in Fig. 4(a) describes the event participation structure of the bird population, that is the weighted allocations of N birds to K foraging events, encoded by $\mathbf{B} \in \mathbb{R}^{N \times K}$. Although this finding is important by itself, as it allows us to quantify the structure of such small foraging groups in terms of the number, individual characteristics, relatedness of their members etc, we seek to move one step further and extract the bird-to-bird *social network* based on the mutual participation of individuals to such events.

Therefore, we seek to define an appropriate *one-mode projection* $\mathbf{B} \in \mathbb{R}^{N \times K} \rightarrow \mathbf{A} \in \mathbb{R}^{N \times N}$, shown in Fig. 4, so that a link a_{ij} between a pair i, j in the resulting network will express how strongly the two birds forage together. We start by defining *co-occurrence* of individuals i and j as the number of times they were recorded in the same foraging group.

Thus, given the event membership profiles $\mathbf{b}_i$ and $\mathbf{b}_j$ for i and j respectively, we define the total co-occurrences a_{ij} as $a_{ij} = \sum_{k=1}^{K} \min(b_{ik}, b_{jk})$, where K is the number of foraging groups and a_{ij} is effectively the *link weight* between i and j in the resulting social network described by the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. Note that other association indices, such as the ones presented in [13] can be used depending on the problem context.

3.4 Co-occurrences: social tie versus coincidence

The next issue we seek to address is the statistical significance of the extracted link weights. Building the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ in the manner described in Section 3.3 makes the very strong assumption that if two individuals participate in the same gathering event, they have some form of social affiliation. This assumption, known in the animal social network literature as the *Gambit of the Group* (GoG) [14], may lead us to adjacency matrices encumbered with "junk" links, produced by co-occurrences that happened by chance. Such coincidences are also frequent in settings where there are natural peak-hours in the data collection period and also when the sensor hardware act as attraction points, as for example the bird feeders in our study. Hence, we seek to define an appropriate *null model* that describes how "statistically surprising" a given link weight would be, if there was no underlying social preference in the foraging habits of the bird population. From previous sections we have discussed that observations occur in bursts (as shown in Fig. 2) that denote small foraging groups of birds that arrive together at the feeders. This is captured by the bird-to-event matrix $\mathbf{B} \in \mathbb{R}^{N \times K}$, where each element b_{ik} in the row vector $\mathbf{b}_i$ denotes the number of times bird i appeared at the gathering event k.

Consider each row vector $\mathbf{b}_i$ as a draw from a multinomial distribution $\mathcal{M}(n_i, \mathbf{p}_i)$, with parameters $n_i = \sum_{k=1}^{K} b_{ik}$ and $p_{ik} = b_{ik}/n_i$. The values of the parameter vector $\{p_{ik}\}_{k=1}^{K}$ can be viewed as a *preference profile* of a bird i to each foraging event k. If our hypothesis that social affiliation between birds affects event membership holds, then closely interacting birds i, j will have similar preference profiles $\mathbf{p}_i$ and $\mathbf{p}_j$.

Let us now propose an element shuffling σ of $\mathbf{p}_i$ so that $\mathbf{p}_i \to \sigma(\mathbf{p}_i)$ and draw a new event occurrence vector $\mathbf{b}_i^{(0)}$ from the multinomial distribution $\mathcal{M}(n_i, \sigma(\mathbf{p}_i))$. Performing this permutation and sampling scheme independently for all birds $i \in \{1, ..., N\}$ leads to a new bird-to-event bipartite network described by $\mathbf{B}^{(0)} \in \mathbb{R}^{N \times K}$. This new matrix $\mathbf{B}^{(0)}$ preserves many key characteristics of the original data, among them the event membership structure, because bird appearances remain concentrated in K regions of increased observation density. Quantities such as the number of individuals N, and the total records n_i, of bird i in the data are also retained.

The key difference introduced in $\mathbf{B}^{(0)}$ is that, although a bird's uneven participation preference $\mathbf{p}_i$ across foraging groups is preserved (as the permuted $\sigma(\mathbf{p}_i)$ has the same entropy as $\mathbf{p}_i$), the shuffling σ "breaks" all correlations between $\mathbf{b}_i$ and $\mathbf{b}_j$ induced by latent social affiliation between individuals i and j. In other words, under our null model birds still forage in small groups, but with *no social preference to which other members of the group they will forage with.* We repeat this process R-times and for each generated bird-to-event matrix $\mathbf{B}^{(0)}$ we extract the bird-to-bird matrix $\mathbf{A}^{(0)}$ using the same one-mode projection presented in Section 3.3. By generating multiple instances of $\mathbf{A}^{(0)}$ in this manner, we are effectively drawing samples from the *ensemble* or family of graphs $\mathcal{G}^{(0)}$ that contains all possible network configurations generated by the null model. Our goal is to examine if our observed network $\mathbf{A}$ is an unlikely case of $\mathcal{G}^{(0)}$.

The randomisation process generates R values of the weight of each link between i and j. From the histogram we get the empirical distribution $P(a_{ij}|H_0)$ that denotes the probability of having a link of weight a_{ij} given that the null hypothesis H_0 holds. We examine how statistically surprising is each observed link a_{ij} by performing a hypothesis test, given an appropriate significance level α, by examining the likelihood $p = P(x \geq a_{ij}|H_0)$ of co-occurrences as large as a_{ij}. Note that the key point of a null model is that co-occurrences happen between individuals, but not as a result of an underlying social structure. In other words, the links in $\mathbf{A}^{(0)}$ are *independent* under H_0, hence $P(\mathbf{A}|H_0) = \prod_{ij} P(a_{ij}|H_0)$. Thus our significance test lies in examining how well this independence assumption can explain the observed co-occurrences encoded in each link of $\mathbf{A}$.

3.5 Integrating information from multiple locations

We briefly expand our graph inference scheme to the multi-location setting. For each record $\{t_z, b_z, \ell_z\}$ in our data stream, we now have an additional term $\ell_z \in \{1, ..., L\}$ that denotes the index of the location where observation z took place.

We start by segmenting our data $\mathcal{D} = \{t_z, b_z, \ell_z\}_{z=1}^{Z}$ into L streams, so that each $\mathcal{D}^{(\ell)}$ contains records referring only to location ℓ. For each $\mathcal{D}^{(\ell)}$ we perform the network extraction process as presented in Sections 3.2 and 3.3 leading to L adjacency matrices $\mathbf{A}^{(\ell)} \in \mathbb{R}^{N_\ell \times N_\ell}$, where $N_\ell \leq N$ the subset of birds recorded at location ℓ. Significance tests, as described in Section 3.4, are performed independently for each ℓ, in order to preserve the unique visitation and location load statistics of each site.

Each matrix $\mathbf{A}^{(\ell)} \in \mathbb{R}^{N_\ell \times N_\ell}$ generated in this scheme captures a subset of the overall connectivity profile in the population. As the interpretation of link weight is the number of co-occurrences between two individuals, the overall a_{ij} is simply the summation $a_{ij} = \sum_{\ell=1}^{L} a_{ij}^{(\ell)}$ over multiple sites.

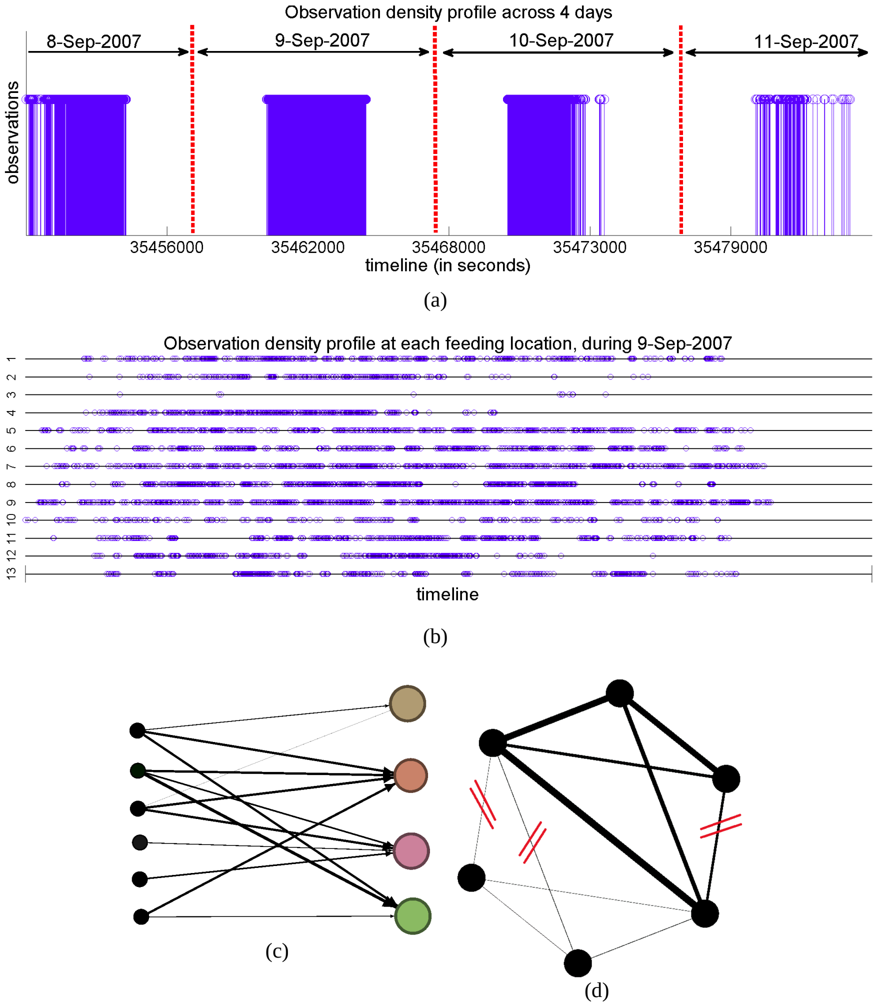
In the next section we will demonstrate how these methodologies are applied to the wild-bird dataset described in Section 2.

4 Results

4.1 Application on the wild-bird dataset

We apply GMMEvents on the dataset of wild-bird foraging records presented in Section 2. Our observations consist of two main streams; $\mathcal{D}_{7,8}$ that covers the activity of $N_{7,8} = 770$ birds from August 2007 to March 2008 and $\mathcal{D}_{8,9}$ that spans from August 2008 to March 2009 and contains $N_{8,9} = 753$ birds.

Instead of applying our method on the whole 2-season data stream directly, we start by breaking it down into 24-hour segments. Our aim is to produce a collection of network snapshots that would allow us to study the day-by-day changes in the population's sociality. An example of the observation data is shown in Fig. 5(a), where we can see the isolated observation-rich regions (blue stem lines) that refer to each particular day. Note that the night period (no-observation zones in between days) acts as a natural separator in our data stream, as no bird foraging activity takes place during that time.



(a)

(b)

(c)

(d)

Figure 5: In Fig. 5(a), we show a segment of our data stream profile for a duration of four days. We pick a single day "data-chunk" of observations and break it down into separate streams that refer to bird records at each particular location, as shown in Fig. 5(b). For each location-specific stream, we use our method to identify gathering events, as shown in coloured nodes on the right of the bipartite graph in Fig. 5(c). We assign birds (black nodes on the left of the graph) into such events based on their participation strength. We project the bird-to-event bipartite graph of Fig. 5(c) into an one-mode network based on co-occurrences in gathering events, as shown in Fig. 5(d). We remove any links (marked with double lines) that can be explained away by the null model.

We proceed by breaking down each daily segment of our data into sub-streams that correspond to L different feeding locations, shown in Fig. 5(b) for the 9^{th} of September 2007. We then apply GMMEvents at each location ℓ *separately*, as co-occurrences need to be defined both in terms of temporal and spatial proximity. On each one of those feeder-specific streams for that day, our method identifies bursts in the observation density profile and builds a bipartite network $\mathbf{B}^{(\ell)}$ between birds and gathering events, as shown in Fig. 5(c). The weight of each link $b_{ik}^{(\ell)}$ denotes the number of times bird i appeared in the gathering event k. Based on Section 3.3, we then perform one-mode projection of this bipartite network into a bird-to-bird social network, shown in Fig. 5(d), described by the adjacency matrix $\mathbf{A}^{(\ell)}$. The weight of each link $a_{ij}^{(\ell)} = \sum_k^K \min(b_{ik}^{(\ell)}, b_{jk}^{(\ell)})$ denotes the total number of co-occurrences between bird i and j across all K gathering events that took place at location ℓ. The statistical importance of each $a_{ij}^{(\ell)}$ is then tested against the null model we formulated in Section 3.4, where all links below the significance threshold (marked with double lines in Fig. 5(d)) are removed. For our significance test we used $R = 10^4$ samples of the null ensemble along with a standard $\alpha = 0.05$ importance threshold.

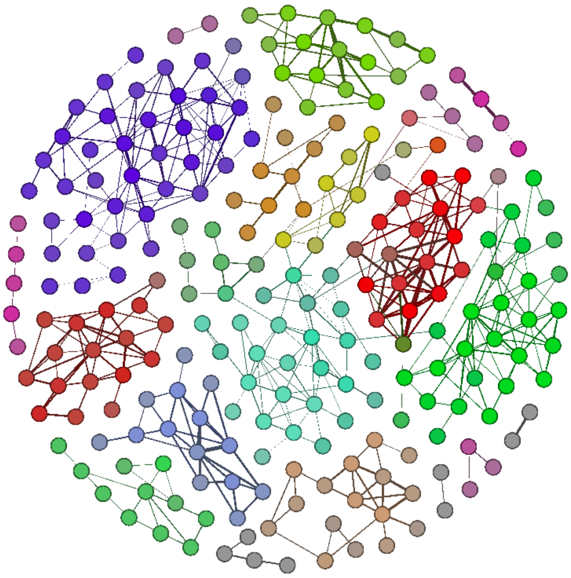


Figure 6: The Wytham woods Parus Major wild bird social network at the 9^{th} of September 2007, with $N = 240$ nodes, $M = 491$ edges, created by integrating all location-specific subgraphs shown in Fig. 5(d). Note that not all 770 birds of the 2007-8 season have been recorded during that day and also individuals no connections have been removed from the network.

We repeat this process for all L locations and based on Section 3.5, we combine all site-specific adjacency matrices $\mathbf{A}^{(\ell)}$ to a single one $\mathbf{A}_t$ that captures the population-wide social structure on the given day t. An example is shown in Fig. 6, where we have summarised the subgraphs (such as the one shown in Fig. 5(c)) from all $L = 13$ locations shown in Fig. 5(b) into a single, global network that describes wild-bird social organisation on the 9th of September 2007. We repeat the process for all T 24-hour segments of our data stream, we get a stack of adjacency matrices $\{\mathbf{A}_t\}_{t=1}^T$ that represent daily snapshots of the wild-bird social network.

From an implementation perspective, GMMEvents runs L times for each day-segment of the data stream. For each location ℓ, R randomisations of the bird-to-event incidence matrix $\mathbf{B}$ are generated and for each one we perform one-mode projection in order to sample the weight distributions for each link pair i, j. Although it may appear computationally prohibitive for large datasets, our method is able to analyse 2-years worth of data that correspond to about 1 million observations in approximately 6 hours, run on a modern 8-core machine under a MATLAB implementation. This is due to the fact that our method itself is executed on multiple small sub-streams (that refer to different locations per day) and can be directly parallelised. Our R randomisation/sampling schemes are also independent by definition, so they can run concurrently on different processing units. More details on computational issues are discussed in the Supplementary Material and our source code documentation.

4.2 Using GMMEvents to study the dynamics of mating pair formation

In this section we examine the validity of the graphs we extracted in Section 4.1 using GMMEvents . As the ground truth network is not available to us in such settings, we can not directly compare our inferred topologies with some form of given solution. Although tests on simulated data streams have been performed (see ESM), our aim is to examine how well our dynamic network reflects meaningful quantities from our application-domain perspective.

We make use of an additional dataset, compiled from an *independent* field study at Wytham woods, which provides wild-bird mating records for each season. Such *pedigree* dataset logs the IDs of individuals that formed a breeding pair each year. Some bird pairs persist over several seasons while others only one year due to either divorce or fatalities. We assume that if the extracted network structure is valid, then breeding individuals will be closely connected, either in terms of a direct link or being in the same *social circle*. Although looking for direct links between mated individuals is an obvious choice, it is a very strict case and thus very sensitive to missing data and noise. Therefore our approach is to examine if breeding pairs belong more frequently and consistently than random into social circles that denote birds with similar foraging patterns.

Our first objective is to identify such social circles in our population. In Fig. 6, where we have visualised the network structure of the wild-bird population for a specific day, we can see certain regions in the graph (shown in different colour) where nodes are more *densely connected* with their immediate neighbours than the rest of the population. Such "hot-spots" of increased link presence are called modules or *communities* in the network analysis jargon [7,15]. For each daily network described by $\mathbf{A}_t$ we extract such communities using a Non-negative Matrix Factorisation (NMF) approach [16].

We find that the majority of mated pairs in network communities are connected through a direct link in 77.26% of cases for the 2007-8 data and 71.57% of cases for the 2008-9 data. Reachability through a path of two links is reported for the 14.74% of cases in 2007-8 and 17.06% of cases in 2008-9. The average path length between two members, for the cases where both of them are observed in the data, is 1.33 (2007-8) and 1.46 (2008-9) with median value of 1 in both datasets. Finally, there are still cases (8% in 2007-8 and 11.37% in 2008-9) of pairs where their geodesic distance spans from 3 to 6 edges but still belong to the same community.

We monitor bird membership into these groups using a binary matrix $\mathbf{C}_t$, where each element $c_{ijt} = 1$ denotes that birds i, j appeared in the same community at day t. This leads us to a new collection of *co-membership* matrices $\{\mathbf{C}_t\}_{t=1}^{T}$ that encode temporal changes in the way birds participate with each other in communities. From a summation across t we get a matrix $\mathbf{C}^{(s)} \in \mathbb{R}^{N\times N}$ where each element $c_{ij}^{(s)}$ denotes the total number of days in the season where the pair i, j participated in the same community. In Fig. 7 we plot a histogram of all co-membership values (y-axis on a logarithmic scale) based on two matrices $\mathbf{C}^{(s)}$ that refer to bird co-membership values in field seasons 2007-8 and 2008-9 respectively. We can see that for both seasons, the vast majority of pairs have never participated in the same group and the distribution is heavily skewed. This implies a strong preferential mechanism in the population, where random individuals rarely belong to the same social circle.

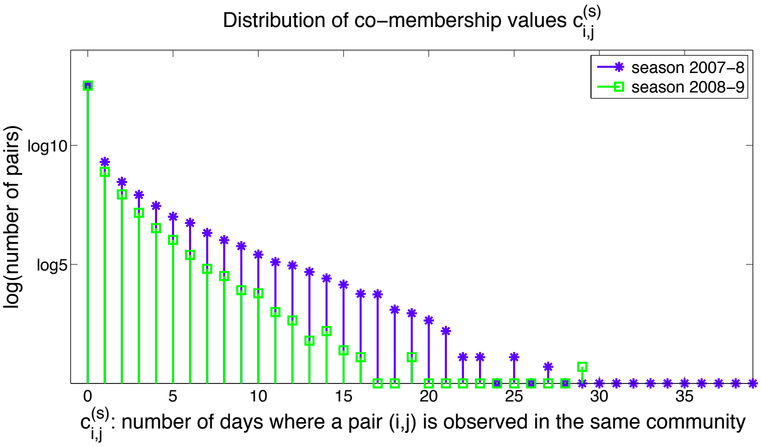
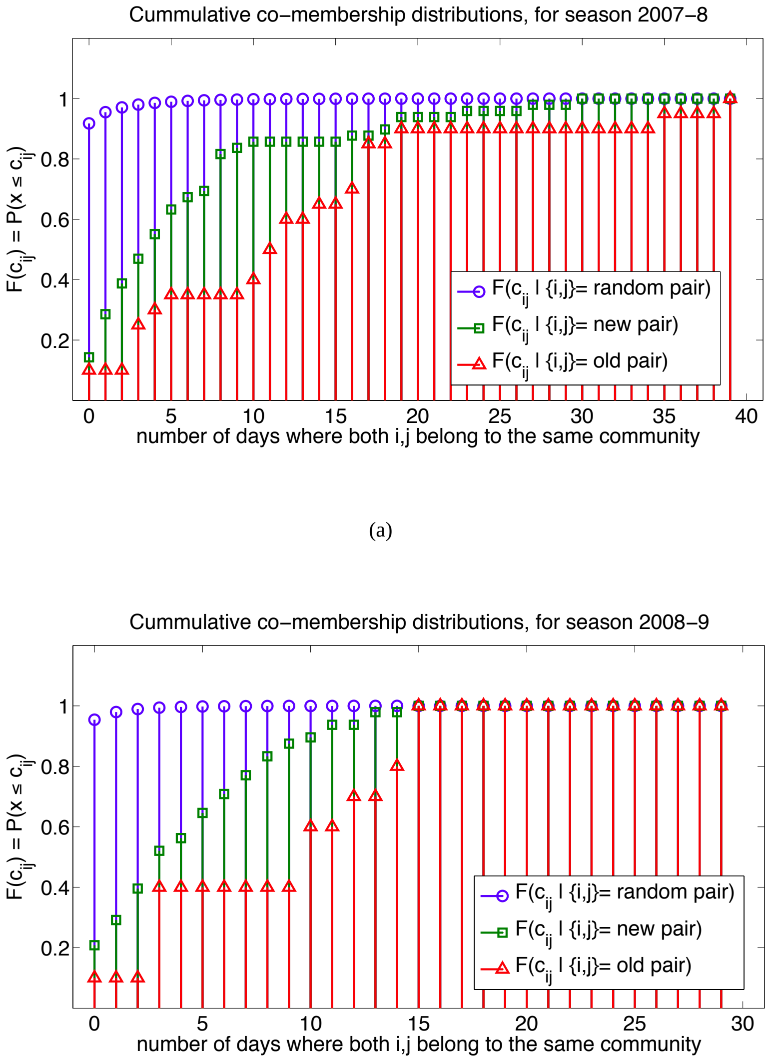


Figure 7: We plot the co-membership values of $\mathbf{C}^{(s)}$ on a logarithmic scale. Each value (x-axis) denotes the total number of days a random pair is observed in the same community. We can see that $\mathbf{C}^{(s)}$ is sparse and the vast majority co-membership values are zero. This shows that if we pick a random dyad in the population, it will most likely be never seen in the same social circle.

We now examine if the above distribution holds for certain sub-category of pairs in the network, which we know a priori that are connected with actual social ties. This prior information is provided by the pedigree dataset we mentioned previously, which gives a list of node dyads i, j that denote breeding individuals. In this list we also distinguish between mated pairs that were formed *during* our observation season, called *new pairs*, and others that already existed before, called *old pairs*. In Fig. 8 we plot the cumulative distributions $F(c_{ij})$, where c_{ij} are values co-membership matrix $\mathbf{C}^{(s)}$ and i, j can be a) any node pair (blue ○-stem), b) a new pair (green □-stem) and c) old pair (red △-stem). In Fig. 8(a) we plot the distributions that refer to the 2007-08 season, with $N = 217$ individuals, from which we have 49 new pairs and 20 old pairs. For season 2008-9, shown in Fig. 8(b), we have $N = 203$ individuals that include 48 new pairs and 10 old pairs.

We can see that for both seasons presented in Fig. 8 the distributions that refer to mated pairs differ significantly from the one for random ones, with p-values $< 10^{-15}$ under a Kolmogorov-Smirnov test [17] with 5% precision level for both seasons. In contrast to the random case where values c_{ij} are mostly zero, co-membership for mated pairs achieves larger values thus denoting stronger and consistent graph proximity. The differences between old and new pairs are also revealed between their respective cumulative distributions (green □-stem and red △-stem), where old pairs achieve higher co-membership values due to the fact that they existed before new pairs where formed, thus they had more opportunities during the season to participate in the same foraging flocks.

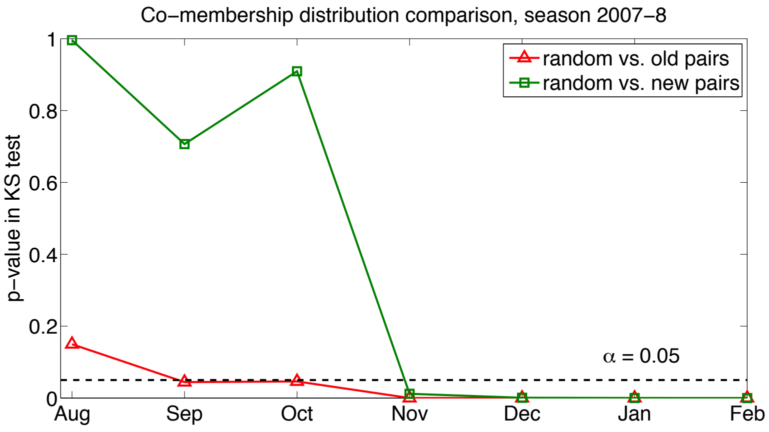


(a)

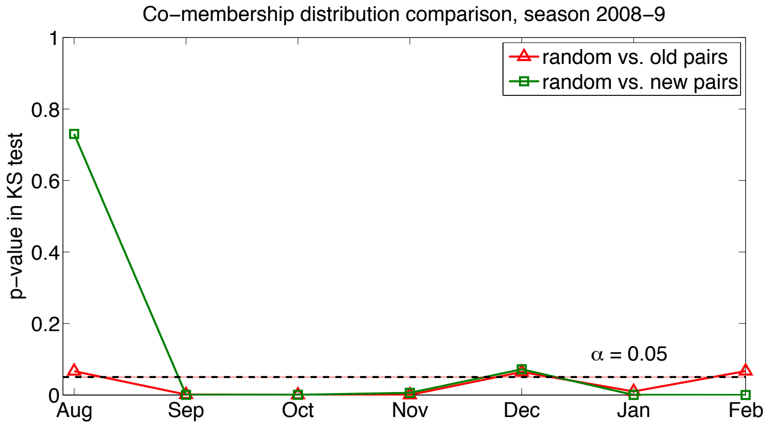
(b)

Figure 8: We plot the cumulative co-membership distributions for three different dyad types; random pairs, mating pairs formed in previous seasons and pairs that formed in the current season. Although for the majority of random bird pairs in the network co-membership values are concentrated around zero, breeding individuals tend to participate much more frequently into the same flocks.

We have already seen that co-membership distributions differ between various pair types. We will now examine when that differentiation takes place during the observation season. We start by breaking down break-down the observation period into 8 months. For each month, we used the respective daily networks in order to find the three co-membership distributions of interest. We then compared $P(c_{ij}|\{i,j\} = \text{random pair})$ versus $P(c_{ij}|\{i,j\} = \text{old pair})$ and $P(c_{ij}|\{i,j\} = \text{random pair})$ versus $P(c_{ij}|\{i,j\} = \text{new pair})$, by calculating the p-value under a Kolmogorov-Smirnov test with a proposed significance level 0.05. In Fig. 9 can see that at the beginning of the season, new pairs have similar co-membership patterns to random ones, as they have not been formed at such early point. But as we move through the year, this similarity drops and from the “cloud” of random associations, breeding relationships emerge. On the other hand, old pairs that have been already formed from previous seasons have a consistent non-random co-membership pattern, even from very early points in the season.



(a)



(b)

Figure 9: We compare the co-membership distributions $P(c_{ij}|\{i,j\} = \text{random pair})$ versus $P(c_{ij}|\{i,j\} = \text{old pair})$ (red △-line) and $P(c_{ij}|\{i,j\} = \text{random pair})$ versus $P(c_{ij}|\{i,j\} = \text{new pair})$ (green □-line) in a month-by-month basis, using a Kolmogorov-Smirnov test. Values above the proposed $\alpha = 0.05$ significance threshold imply that the two distributions under comparison are similar. We can see that from very early in the year old pairs differentiate themselves from random, by starting to participate frequently in the same communities. On the other hand, members of new pairs in the beginning of the year treat each other as random, while preferential mechanism that makes them flock together, starts to build-up during early winter.

5 Discussion and future work

The network paradigm is a powerful tool for studying real-world complex systems. As there is an extensive toolset of methods and algorithms for network analysis, in this work we have focused on the problem of constructing the network in the first place. In many applications, the collected data capture the behaviour of the system in some manner, like the spatial trajectories of participating agents, but not the underlying relations between them. We address this issue by assuming that mobility patterns of individuals may be correlated based on some form of underlying social connection. By identifying observation-dense regions in the data stream, which can be seen as *gathering events* of affiliated individuals, we propose a methodology of drawing links between agents based on their co-participation into those events.

Traditional approaches [9–12] in constructing social networks from spatio-temporal data involve discretising the observation stream based on some fixed time window Δt and drawing links between individuals when they lie within such “interaction-radius”. Our method overcomes the practical difficulties of such time-slicing approach in cases when we have no prior knowledge of how big or small the time window size should be, thus we need to perform multiple runs across various Δt and select the appropriate one based on some ad-hoc quality function. Additionally, we have proposed an appropriate null model, which allows us to examine if the co-occurrence of individuals into gathering events are a result of a latent social tie, or coincidence. Our null model retains the "bursty" nature of the data stream but breaks all correlations between the individuals' appearance patterns through an appropriate randomisation.

We applied GMMEvents into two large-scale datasets that provide wild-bird foraging records. We showed that the inferred network topologies reflect mating pair formation events that take place in the population, where breeding individuals tend to belong into the same foraging groups more often that random dyads. We also showed that the dynamics of community structure in the system reveal how newly formed pairs initially have a random-like behaviour, while as we approach the mating season they start to participate more often than random into the same communities.

The communities identified here are based on temporal occurrence at feeding stations, and while the data analysed here are extensive, they are incomplete, as observations are made for only a proportion of the time, and only for feeding-related activity. While more complete data would be expected to result in more completely connected communities (both in terms of link number and connection strength), it is not necessarily the case that all communities would ultimately be fully connected. For example, communities might be comprised of pairs of individuals that avoided each other (e.g. territorial males, competing females) relative to the other members of the community, even thought they had links via other individuals. As expected for individuals linked via a network, there is a variety of direct and indirect ways that individuals within and between communities might influence each other. In the case of the present network, we might expect that an important source of direct effects lie in the flow of information between community members about the presence of food, but such information will also spread indirectly to other individuals via network links between communities [19]. Numerous other effects might also be considered. For example, like many animals, small passerine birds give alarm calls that alert other individuals about the presence of predators [18]. While the individuals in the same community may be expected to be nearest to a focal individual, other linked communities may also be influenced directly by this sort of behaviour, and the overall inter-community network may serve as a hypothesis for the likelihood of such effects being transmitted between individuals. So far all feeding sites have been analysed isolation until the last stage. Site-specific network adjacency matrices are extracted and tested for significance whence they are all combined to one single adjacency matrix. An alternative to this spatial aggregation over sites would be temporal aggregation. In this approach temporal data could be aggregated and behaviour and feeding sites analysed directly. While such an analysis may account for popular feeders it would not achieve the high temporal resolution of the existing approach. For instance, using a temporal aggregation strategy, a group of birds feeding in the morning and one in the evening would all be treated as one single group when the times of their feeding site visitation clearly suggest otherwise. A proper resolution of this conflict may require a full spatio-temporal clustering stage and another bespoke hypothesis test to detect both spatially and temporally insignificant events. Such a multivariate approach would alleviate the necessity to account for spatial correlations during hypothesis testing which otherwise would be extremely hard to extract from data. Thus, in our future work we will focus on a full spatio-temporal analysis of bird behaviour and the development of clustering models that combine data of different characteristics, such as the bursty behavioural and the continuous spatial data.

The next stages of our research consist of two main modules. From the perspective of the model, we seek to extend the way we define the link a_{ijt} between two individuals at time t so that we take into account *prior knowledge* from previous observations. This has the advantage of capturing the uncertainty over the link weight, detect abrupt changes in the network topology and handling missing observations in a principled manner. From an ecological point of view, we currently run an improved scheme of our data collection, where we have sensors at each feeding location. This gives us the advantage of looking at the data at much greater resolutions thus having a more accurate view of the overall bird population's foraging patterns.

Although the methodology we presented is applied to animal observation records, it can be extended to any system where agents perform check-ins at certain locations and such observations are not uniformly spread in data stream, but *temporally focused*. We believe methodologies and theoretical results derived from the study of animal social networks will benefit the wider field of network analysis, as individuals can be monitored from the beginning to the end of their life span, there are no privacy issues associated with data collection and understanding the dynamics of animal interactions provides an insight into the behaviour and evolution of complex systems.

6 Acknowledgements

The authors are grateful to Simon Evans, Ada Grabowska and particularly Teddy Wilkin for data collection, and to the National Environment Research Council (NE/D011744/1) and the European Research Council (AdG 250164) for funding aspects of the work. Ioannis Psorakis is funded by a Microsoft Research European PhD scholarship, for which we are most grateful. Additionally, we gratefully acknowledge funding from the UK Research Council for project 'Orchid', grant EP/I011587/1. Finally, we warmly thank the reviewers for their constructive comments and evaluation.

References

[1] Rosvall M. 2006 *Information Horizons in a Complex World*, ISBN 91-7264-117-7.

[2] Newman M.E.J. 2010 *Networks: an Introduction*, Oxford University Press, ISBN 978-0-19-920665-0

[3] Wey T., Blumstein D.T., Shen W., Jordn F. 2008 Social network analysis of animal behaviour: a promising tool for the study of sociality, *Animal Behaviour.* **75** pp. 333 – 344 DOI:10.1016/j.anbehav.2007.06.020

[4] Krause J., Lusseau D., James R. 2009 Animal social networks: an introduction, *Behavioral Ecology and Sociobiology* **63** pp. 967–973.

[5] Buchanan M., Caldarelli G. 2010 A networked world, *Physics World* **23** pp. 22–24.

[6] Barrat A., Barthélemy M., Vespignani A. 2004 Modeling the evolution of weighted networks *Phys Rev E* **70** pp. 066149.

[7] Fortunato S. 2010 Community Detection in Graphs *Physics Reports* **486** pp. 75–174.

[8] Newman M.E.J. The structure and function of complex networks *SIAM Rev.*, **45** pp. 167–256.

[9] Lauw H.W., Lim E.P., Pang H., Tan T.T. 2005 Social network discovery by mining spatio-temporal events *Computational & Mathematical Organization Theory* **11** pp. 97–118.

[10] Whitehead H. 2008 Analyzing Animal Societies: *Quantitative Methods for Vertebrate Social Analysis*, Chicago University Press ISBN: 978-0226895239.

[11] Oh K.P., Badyaev A.V. 2010 Structure of social networks in a passerine bird: consequences for sexual selection and the evolution of mating strategies. *The American Naturalist*, **176** pp. E80–E89.

[12] Gero S., Engelhaupt D., Rendell L., Whitehead H. 2009 Who Cares? Between-group variation in alloparental caregiving in sperm whales *Behavioral Ecology*, **20** pp. 838.

[13] Ginsberg J.R., Young T.P. 1992 Measuring association between individuals or groups in behavioural studies *Animal Behaviour*.

[14] Whitehead H., Dufault S. 1992 Techniques for analyzing vertebrate social structure using identified individuals: review and recommendations *Advances in the Study of Behavior* **28** pp. 33–74.

[15] Porter M.A., Onnela J.P., Mucha P.J. 2009 Communities in Networks *Notices of the American Mathematical Society*, **56** pp. 1082–1097 and 1164–1166.

[16] Psorakis I., Roberts S., Ebden M., Sheldon B. Overlapping community detection using Bayesian non-negative matrix factorization *Phys Rev E*, **83** pp. 066114.

[17] Lilliefors H.W. 1967 On the Kolmogorov-Smirnov test for normality with mean and variance unknown *Journal of the American Statistical Association* 1967 p. 399–402.

[18] Zuberbühler K. 2009 Survivor signals: the biology and psychology of animal alarm calling *Advances in the Study of Behavior* 40:277–322.

[19] Franz M. and Nunn C.L. 2009 Network-based diffusion analysis: a new method for detecting social learning *Proceedings of the Royal Society B: Biological Sciences* 276(1663):1829–1836.

[20] Krings, G., Karsai, M., Bernharsson, S., Blondel, V. D. and Saramäki, J. 2012 Effects of time window size and placement on the structure of aggregated networks *arXiv:1202.1145v1*.