

# A Model for Birdwatching and other Chronological Sampling Activities

Jesús A. De Loera<sup>1</sup>, Edgar Jaramillo-Rodriguez<sup>1</sup>,  
Deborah Oliveros<sup>2</sup>, and Antonio J. Torres<sup>2</sup>

<sup>1</sup>Department of Mathematics, University of California, Davis

<sup>2</sup> Instituto de Matemáticas, Universidad Nacional Autónoma de México

May 13, 2022

## Abstract

In many real life situations one has  $m$  types of random events happening in chronological order within a time interval and one wishes to predict various milestones about these events or their subsets. An example is birdwatching. Suppose we can observe up to  $m$  different types of birds during a season. At any moment a bird of type  $i$  is observed with some probability. There are many natural questions a birdwatcher may have: how many observations should one expect to perform before recording all types of birds? Is there a time interval where the researcher is most likely to observe all species? Or, what is the likelihood that several species of birds will be observed at overlapping time intervals? Our paper answers these questions using a new model based on random interval graphs. This model is a natural follow up to the famous coupon collector's problem.

## 1 Introduction.

Suppose you are an avid birdwatcher and you are interested in the migratory patterns of different birds passing through your area this winter. Each day you go out to your backyard and keep an eye on the skies; once you see a bird you make a note of the species, day, and time you observed it. You know from prior knowledge that there are  $m$  different species of birds that pass over your home every year and you would love to observe at least one representative of each species. Naturally, you begin to wonder: *after  $n$  observations, how*

*likely is it that I have seen every type of bird?* If we only care that all  $m$  types of birds are observed at least once after  $n$  observations, we recognize this situation as an example of the famous *coupon collector's problem* (for a comprehensive review of this problem see [7] and references therein). In this old problem a person is trying to collect  $m$  types of objects, the coupons, labeled  $1, 2, \dots, m$ . The coupons arrive one by one as an ordered sequence  $X_1, X_2, \dots$  of independent identically distributed (i.i.d.) random variables taking values in  $[m] = \{1, \dots, m\}$ .

But a professional birdwatcher is also interested in more nuanced information than the coupon collector. To properly understand interspecies interactions, one not only hopes to observe every bird, but also needs to know which species passed through the area at the same time(s). For example, the birdwatcher might also ask:

- *What are the chances that the visits of  $k$  types of birds do not overlap at all?*
- *What are the chances that a pair of birds is present on the same time interval?*
- *What are the chances of one bird type overlapping in time with  $k$  others?*
- *What are the chances that all the bird types overlap in a time interval?*

We note that very similar situations, where scientists collect or sample time-stamped data that comes in  $m$  types or classes and wish to predict overlaps, appear in applications as diverse as archeology, genetics, job scheduling, and paleontology [13, 8, 23, 14]. The purpose of this paper is to present a new *random graph model* to answer the four time-overlap questions above.

Our model is very general, but to avoid unnecessary formalism and technicalities, we present clear answers in some natural special cases that directly generalize the coupon collector problem. For the special cases we analyze, the only tools we use are a combination of elementary probability and combinatorial geometry.

## 1.1 Establishing a general random interval graph model.

In order to answer any of the questions above we need to deal with one key problem: how do we estimate which time(s) each species of bird might be present from a finite number of observations? To do so, we will make some modeling choices which we outline below.

The first modeling choice is that our observations are samples from a stochastic process indexed by a real interval  $[0, T]$  and taking values in  $[m]$ . We recall the definition of a stochastic process for the reader (see [20]): Let  $I$  be a set and let  $(\Omega, \mathcal{F}, P)$  be a probability space. Suppose that for each  $\alpha \in I$ , there is a random variable  $Y_\alpha : \Omega \rightarrow S \subset \mathbb{R}$  defined on  $(\Omega, \mathcal{F}, P)$ . Then the function  $Y : I \times \Omega \rightarrow S$  defined by  $Y(\alpha, \omega) = Y_\alpha(\omega)$  is called a *stochastic process* with *indexing set*  $I$  and *state space*  $S$ , and is written  $Y = \{Y_\alpha : \alpha \in I\}$ . When we conduct an observation at some time  $t_0 \in [0, T]$ , we are taking a sample of the random variable  $Y_{t_0}$ .

For each  $i \in [m]$ , the probabilities that  $Y_t = i$  give us a function from  $[0, T] \rightarrow [0, 1]$ , which we call the *rate function* of  $Y$  corresponding to  $i$ ; the name is inspired by the language of Poisson point processes where the density of points in an interval is determined by a *rate* parameter (see [24]).

**Definition 1** (Rate function). Let  $Y = \{Y_t : t \in [0, T]\}$  be a stochastic process with indexing set  $I = [0, T]$  and state space  $S = [m]$ . The *rate function* corresponding to label  $i \in S$  in this process is the function  $f_i : I \rightarrow [0, 1]$  given by

$$f_i(t) = P(Y_t = i) = P(\{\omega : Y(t, \omega) = i\}).$$

Figure 1 gives two examples of the rate functions of some hypothetical stochastic processes (we will clarify the meaning of stationary and non-stationary later in this section when we discuss a special case of our model). Observe that at a fixed time  $t_0$ , the values  $f_i(t_0)$  sum to 1 and thus determine the probability density function of  $Y_{t_0}$ . Therefore, the rate functions describe the change of the probability density functions of the variables  $Y_t$  with respect to the indexing variable  $t$ .

Next, note that the set of times where species  $i$  might be present is exactly the *support* of the rate function  $f_i$ . Recall, the support of a function is the subset of its domain for which the function is non-zero, in our case this will be a portion of  $[0, T]$ . Therefore, *our key problem is to estimate the support of the rate functions from finitely many samples*.

We note that the stochastic process  $Y$  is defined to take values in  $[m]$  due to a modeling choice on our part. Alternatively, one could have  $Y$  take values in the power set  $2^{[m]}$ , so as to allow for multiple species of birds to be observed at the same time. However, choosing  $[m]$  rather than  $2^{[m]}$  simplifies some calculations and, moreover, is quite reasonable. Rather than registering “three birds at 6 o’clock,” our birdwatcher can instead register three sightings: one bird at 6:00:00, a second at 6:00:01, and a third at 6:00:02, for example.

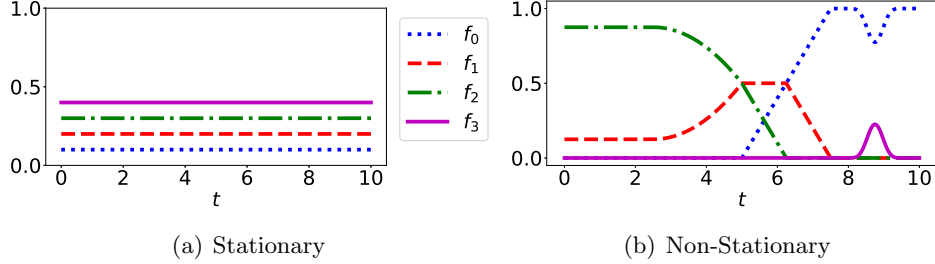


Figure 1: Two examples of hypothetical rate functions.

This brings us to our next modeling choice: all the rate functions  $f_i$  have connected support for each  $i \in [m]$ . This is reasonable for our motivation; after all, a bird species first seen on a Monday and last seen on a Friday is not likely to suddenly be out of town on Wednesday. The main benefit of this assumption is that now the support of the rate function  $f_i$ ,  $\text{supp}(f_i)$ , is a sub-interval of  $[0, T]$ . This fact provides a natural way of approximating the support of  $f_i$ : given a sequence of observations  $Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}$  with  $0 \leq t_1 < t_2 < \dots < t_n \leq T$ , let  $I_n(i)$  denote the sub-interval of  $[0, T]$  whose endpoints are the first and last times  $t_k$  for which  $Y_{t_k} = i$ . Note that it is possible for  $I_n(i)$  to be empty or a singleton. It follows that  $I_n(i) \subset \text{supp}(f_i)$  so we can use it to approximate  $\text{supp}(f_i)$ . We call the interval  $I_n(i)$  the *empirical support* of  $f_i$ , as it is an approximation of  $\text{supp}(f_i)$  taken from a random sample.

In summary, our model is actually quite simple: given a sequence of observations  $Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}$  we construct  $m$  random intervals  $I_n(1), \dots, I_n(m)$  whose endpoints are the first and last times we see its corresponding species. These intervals, known as the *empirical supports*, are approximations of the supports of the rate functions,  $f_i$ , and satisfy  $\text{supp}(f_i) \supset I_n(i)$ .

The four birdwatching questions above may be expressed in terms of the empirical supports as follows:

- What are the chances that none of the empirical supports  $I_n(i)$  intersect?
- What are the chances that a particular pair of empirical supports  $I_n(i)$  and  $I_n(j)$  intersect?
- What are the chances that one empirical support,  $I_n(i)$ , intersects with  $k$ -many others?

- What are the chances that the collection of empirical supports has a non-empty intersection?

To make these questions even easier to analyze, we will present a combinatorial object: an *interval graph* that records the intersections of the intervals  $I_n(i)$  in its edge set.

**Definition 2.** Given a finite collection of  $m$  intervals on the real line, its corresponding interval graph,  $G(V, E)$ , is the simple graph with  $m$  vertices, each associated to an interval, such that an edge  $\{i, j\}$  is in  $E$  if and only if the associated intervals have a nonempty intersection, i.e., they overlap.

Figure 2 demonstrates how we construct the desired interval graph from some observations. Figure 2(a) shows a sequence of  $n = 11$  points on the real line, which corresponds to the indexing set  $I$  of our random process  $Y$ . Above each point we have a label, representing a sample from  $Y$  at that time. Displayed above the data are the empirical supports  $I_n(i)$  for each  $i \in [m] = [4]$ . Finally, Figure 2(b) shows the interval graph constructed from these four intervals where each vertex is labeled with the interval it corresponds to. In this example there are no times shared by the species  $\{1, 2\}$  and the species  $\{4\}$ , so there are no edges drawn between those nodes.

We emphasize that the interval graph constructed in this way will contain up to  $m$ -many vertices, but may contain fewer if some of the intervals  $I_n(i)$  are empty, i.e., if we never see species  $i$  in our observations.

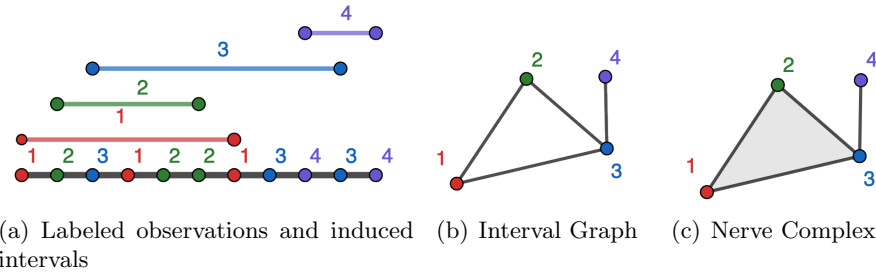


Figure 2: Example observations with their corresponding graph and nerve.

Although the interval graph  $G(V, E)$  is constructed using only pairwise intersections, we can further encode all  $k$ -wise intersections for  $k = 2, \dots, m$  in a *simplicial complex*, which is a way to construct a topological space by gluing *simplices* (generalizations of triangles, tetrahedra, etc). A simplicial complex  $K$  must satisfy the two requirements that every face of a simplex in  $K$  is also in  $K$  and that the non-empty intersection of any two simplices

in  $K$  is a face of both. (for an introduction to basic topology and simplicial complexes see [11, 16]). The construction we need is known as the *nerve complex* (see [19], [29], [22, p. 197] and [11, p. 31]).

**Definition 3.** Let  $\mathcal{F} = \{F_1, \dots, F_m\}$  be a family of convex sets in  $\mathbb{R}^d$ . The *nerve complex*  $\mathcal{N}(\mathcal{F})$  is the abstract simplicial complex whose  $k$ -facets are the  $(k+1)$ -subsets  $I \subset [m]$  such that  $\bigcap_{i \in I} F_i \neq \emptyset$ .

Figure 2(c) shows the nerve complex constructed from the intervals  $I_n(i)$  in Figure 2(a). Note the presence of a 2-simplex (triangle) with vertices  $\{1, 2, 3\}$  because the corresponding intervals mutually intersect.

By construction, the interval graph  $G$  is exactly the 1-skeleton of the nerve complex  $\mathcal{N}$  generated by the intervals. In fact, because our intervals lie in a 1-dimensional space,  $\mathcal{N}$  is completely determined by  $G$ . To see this, suppose we have a collection of intervals  $(x_1, y_1), \dots, (x_k, y_k)$  such that all intervals intersect pairwise. It follows that  $y_i \geq x_j$  for all  $i, j \in [k]$ , and so  $(\max\{x_1, \dots, x_k\}, \min\{y_1, \dots, y_k\}) \subseteq \bigcap_{i=1}^k (x_i, y_i)$ . Hence the whole collection has non-empty intersection (this is a special case of Helly's theorem [1], which is necessary in higher dimensional investigations). Thus, the  $k$ -dimensional faces of the nerve complex are precisely  $k$ -cliques of the interval graph.

Therefore, going forward we will refer to the nerve complex  $\mathcal{N}$  and the graph  $G$  interchangeably depending on the context, but the reader should understand that these are fundamentally the same object as long as the family of convex sets  $\mathcal{F}$  lies in a 1-dimensional space. We stress that in higher dimensions the intersection graph of convex sets *does not* determine the nerve complex (we demonstrate this by an example in the Conclusion).

We can now present our random interval graph model in its entirety:

**Definition 4** (The Random Interval Graph Model). We let  $Y = \{Y_t : t \in [0, T]\}$  be a stochastic process as above and let  $\mathcal{P} = \{t_1, t_2, \dots, t_n\}$  be a set of  $n$  distinct observation times or sample points in  $[0, T]$  with  $t_1 < t_2 < \dots < t_n$ . Then let  $Y = (Y_1, Y_2, \dots, Y_n)$  be a random vector whose components  $Y_i$  are samples from  $Y$  where  $Y_i = Y_{t_i}$ , so each  $Y_i$  takes values  $\{1, \dots, m\}$ . For each label  $i$  we define the (possibly empty) interval  $I_n(i)$  as the convex hull of the points  $t_j$  for which  $Y_j = i$ , i.e., the interval defined by points colored  $i$ . Explicitly  $I_n(i) = \text{Conv}(\{t_j \in \mathcal{P} : Y_j = i\})$ , and we refer to  $I_n(i)$  as the *empirical support* of label  $i$ . Furthermore, because it comes from the  $n$  observations or samples, we call the nerve complex,  $\mathcal{N}(\{I_n(i) : i = 1, \dots, m\})$ , the *empirical nerve* of  $Y$  and denote it  $\mathcal{N}_n(Y)$ .

Under this random interval graph model our four questions can be rephrased in terms of the random graph  $\mathcal{N}_n(Y)$ :

- *What is the likelihood that  $\mathcal{N}_n(Y)$  has no edges?*
- *What is the likelihood that a particular edge is present in  $\mathcal{N}_n(Y)$ ?*
- *What is the likelihood of having a vertex of degree at least  $k$  in  $\mathcal{N}_n(Y)$ ?*
- *What is the likelihood that  $\mathcal{N}_n(Y)$  is the complete graph  $K_m$ ?*

Our original questions have become questions about random graphs!

## 1.2 The special case this paper analyzes.

We presented a very general model because it best captures the nuances and subtleties of our motivating problem. However, without additional assumptions on the distribution  $Y$ , the prevalence of pathological cases makes answering the motivating questions above become very technical and unintuitive. Therefore, our analysis will focus on a special case of this problem where we make two additional assumptions on  $Y$  so that our analysis only requires basic combinatorial probability.

The first assumption we make is that our observations  $Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}$  are mutually independent random variables. Note, we do not claim that all pairs of random variables  $Y_s, Y_t$  for  $s, t \in [0, T]$  are independent. We only claim this holds for all  $s, t \in \{t_1, t_2, \dots, t_n\}$ . The second assumption we make is that the rate functions  $f_i$  be constant throughout the interval  $[0, T]$ . In this case, there exist constants  $p_1, p_2, \dots, p_m \geq 0$  such that  $\sum_{i=1}^m p_i = 1$  and  $f_i(t) = p_i$  for all  $t \in [0, T]$  and all  $i \in [m]$ . We call the special case of our model where both of these assumptions are satisfied the *stationary case* and all other cases as *non-stationary*. Figure 1 shows examples of a stationary case, 1(a), and a non-stationary case, 1(b). We will also refer to the *uniform case*, which is the extra-special situation where  $p_i = \frac{1}{m}$  for all  $i \in [m]$ . Note Figure 1(a) is stationary but not uniform.

Of course, the stationary case is less realistic and applicable in many situations. For example, it is not unreasonable to suppose that the presence of a dove at 10 o'clock should influence the presence of another at 10:01, or that the presence of doves might fluctuate according to the season and time of day. However, the stationary case is still rich in content and, importantly, simplifies things so that this analysis requires only college-level tools from

probability and combinatorics. Moreover, as we discuss below, the stationary case has a strong connection to the famed coupon collector problem and is of interest as a novel method for generating random interval graphs.

The stationary case assumptions directly lead to two important consequences that greatly simplify our analysis. The first is that now the random variables  $Y_{t_1}, \dots, Y_{t_n}$  are independent and identically distributed (i.i.d.) such that  $P(Y_{t_k} = i) = p_i > 0$ . Note that this is true for any set of distinct observation times  $\mathcal{P} = \{t_1, \dots, t_n\}$ . The second consequence simplifies things further still: though the points  $\mathcal{P}$  corresponding to our sampling times have thus far been treated as arbitrary, one can assume without loss of generality that  $\mathcal{P} = [n] = \{1, 2, \dots, n\}$  since all sets of  $n$  points in  $\mathbb{R}$  are combinatorially equivalent, as explained in the following lemma.

**Lemma 5.** *Let  $\mathcal{P} = \{x_1, \dots, x_n\}$  and  $\mathcal{P}' = \{x'_1, \dots, x'_n\}$  be two sets of  $n$  distinct points in  $\mathbb{R}$  with  $x_1 < \dots < x_n$  and  $x'_1 < \dots < x'_n$ . Let  $Y = (Y_1, \dots, Y_n)$  and  $Y' = (Y'_1, \dots, Y'_n)$  be i.i.d. random vectors whose components are i.i.d. random variables taking values in  $[m]$  with  $P(Y_j = i) = p_i > 0$  and  $P(Y'_j = i) = p_i > 0$ . Then for any abstract simplicial complex  $\mathcal{K}$  we have that  $P(\mathcal{N}_n(\mathcal{P}, Y) = \mathcal{K}) = P(\mathcal{N}_n(\mathcal{P}', Y') = \mathcal{K})$ .*

*Proof.* Let  $c_1, c_2, \dots, c_n$  be an arbitrary sequence of labels, so  $c_i \in [m]$  for each  $i$ . Because  $Y, Y'$  are i.i.d. we have that  $P(\cap_{i=1}^n \{Y_i = c_i\}) = P(\cap_{i=1}^n \{Y'_i = c_i\})$ . Therefore if both sequences of colors  $Y_i = Y'_i = c_i$  have the same order for all  $i = 1, \dots, n$ , then it is sufficient to show that the two empirical nerves are the same. Consider two empirical supports  $I_n(j)$  and  $I_n(k)$  of labels  $j, k$ , and observe that if they do (do not) intersect on  $Y_i$ , then the two empirical supports  $I'_n(j)$  and  $I'_n(k)$  of labels  $j, k$  do (do not) intersect, then the two corresponding empirical nerves do (do not) contain the edge  $\{j, k\}$ . This implies that the two nerves have the same edge set. Furthermore, as we observed before, due to Helly's theorem in the line the empirical nerve is completely determined by its 1-skeleton. Then both empirical nerves are the same.  $\square$

We now summarize the key assumptions of our model in the stationary case.

**Key assumptions for our analysis:** *In all results that follow let  $Y = (Y_1, \dots, Y_n)$  be a random vector whose components are i.i.d. random variables such that  $P(Y_j = i) = p_i > 0$  for all  $i \in [m]$ . As a consequence the support functions of the underlying stochastic process are constant and each has support on the entire domain. We denote by  $\mathcal{N}_n = \mathcal{N}_n([n], Y)$  the empirical nerve of the random coloring induced by  $Y$ . We also denote the*



graph or 1-skeleton of  $\mathcal{N}_n$  by the same symbol. When we refer to the uniform case this means the special situation when  $p_i = \frac{1}{m}$  for all  $i = 1, \dots, m$ .

### 1.3 Context and prior work.

We want to make a few comments to put our work in context and mention prior work:

The famous coupon collector problem that inspired us dates back to 1708 when it first appeared in De Moivre's *De Mensura Sortis (On the Measurement of Chance)* [7]. The answer for the coupon collector problem depends on the assumptions we make about the distributions of the  $X_i$ . Euler and Laplace proved several results when the coupons are equally likely, that is when  $P(X_i = k) = \frac{1}{m}$  for every  $k \in [m]$ . The problem lay dormant until 1954 when H. Von Schelling obtained the expected waiting time when the coupons are not equally likely [27]. More recently, Flajolet et. al. introduced a unified framework relating the coupon collector problem to many other random allocation processes [9]. We note that the stationary case of our model has the same assumptions as this famous problem: an observer receives a sequence of i.i.d. random variables taking values in  $[m]$ . In the language of our model, the coupon collector problem could be posed as, *What is the likelihood that the nerve  $\mathcal{N}_n(Y)$  will contain exactly  $m$  vertices?* Thus, we can consider this model a generalization of the coupon collector problem which seeks to answer more nuanced questions about the arrival of different coupons.

Interval graphs have been studied extensively due to their wide applicability in areas as diverse as archeology, genetics, job scheduling, and paleontology [13, 8, 23, 14]. These graphs have the power to model the overlap of spacial or chronological events and allow for some inference of structure. There are also a number of nice characterizations of interval graphs that have been obtained [21, 10, 12, 15]. For example, a graph  $G$  is an interval graph if and only if the maximal cliques of  $G$  can be linearly ordered in such a way that, for every vertex  $x$  of  $G$ , the maximal cliques containing  $x$  occur consecutively in the list. Another remarkable fact of interval graphs is that they are *perfect* and thus the weighted clique and coloring problems are polynomial time solvable [13]. Nevertheless, sometimes it may not be immediately clear whether a graph is an interval graph or not. For example, of the graphs in Figure 3 only 3(a) is an interval graph.

The most popular model for generating random graphs is the Erdős-Renyi model [6], but it is insufficient for studying random interval graphs. The reason is that, as pointed out in [2], an Erdős-Renyi graph is almost

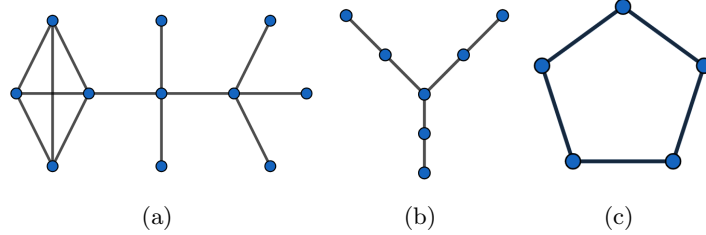


Figure 3: It is not obvious which of these graphs are interval.

certainly *not* an interval graph as the number of vertices goes to infinity.

Several other authors have studied various models for generating random *interval graphs* (see [4, 25, 26, 18, 17, 23] and the many references therein). Perhaps most famously Scheinerman introduced [25, 26], and others investigated [4, 18, 17], a method of generating random interval graphs with a fixed number of intervals  $m$ : the extremes of the intervals  $\{(x_1, y_1), \dots, (x_m, y_m)\}$  are  $2m$  points chosen independently from some fixed continuous probability distribution on the real line. Each pair  $(x_i, y_i)$  determines a random interval. This is a very natural simple random process, but it is different from our random process (see the Appendix).

We noted earlier that because our intervals lie in a 1-dimensional space, the nerve complex is completely determined by the interval graph because the  $k$ -facets of the nerve complex are exactly the  $k$ -cliques of the interval graph. In other words, the nerve complex is precisely the *clique complex* of the interval graph. We also remark that the complement graph of the interval graph  $G$  is the graph  $H$  of non-overlapping intervals. The graph  $H$  is in fact a partially ordered set, called the *interval order* where one interval is less than the other if the first one is completely to the left of the second one. We can associate to each *independent set* of  $k$  non-intersecting intervals, a  $(k - 1)$ -dimensional simplex, this yields a simplicial complex, the *independence complex* of the corresponding interval order graph  $H$ . Observe that this independence complex is the same as the nerve  $\mathcal{N}$  we just defined above. This is all well-known since the independence complex of any graph equals the clique complex of its complement graph, and vice versa (see Chapter 9 in [19]).

#### 1.4 Outline of our contributions.

In this paper we answer the four birdwatching questions using the random interval graphs and complexes generated by the stochastic process described

above. Here are our results section by section:

Section 2 presents various results about the expected structure of the random interval graph  $\mathcal{N}_n$ , including the expected number of edges and the likelihood that the graph has an empty edge set.

Section 3 presents results regarding the distribution of maximum degree and clique number of the graph  $\mathcal{N}_n$ , and our results show that the random interval graph asymptotically approximates the complete graph,  $K_m$ , as the number of samples  $n$  grows large. This means the nerve complex is asymptotically an  $(m - 1)$ -dimensional simplex. From the results of Section 3 one can see that as we sample more and more bird observations it becomes increasingly unlikely that we see any graph other than  $K_m$ . We investigate the number of samples needed to find  $K_m$  with high probability.

Section 4 closes the paper outlining three natural open questions. We also include an Appendix that contains computer experiments to evaluate the quality of various bounds proved throughout the paper and to show our model is different from earlier models of random interval graphs.

## 2 Random Interval Graphs and Behavior in Expectation.

In this section we explore what type of nerve complexes one might expect to find for a fixed number of observations  $n$  when the likelihood of observing each label  $i$  is a constant  $p_i > 0$ .

**Proposition 6.** *Under the key assumptions in Section 1, the probability that the random graph  $\mathcal{N}_n$  is the empty graph with  $0 \leq k \leq m$  vertices but no edges,  $K_k^c$ , is given by*

$$P(\mathcal{N}_n = K_k^c) \geq p_*^n k! \binom{m}{k} \binom{n-1}{k-1},$$

where  $p_* = \min\{p_1, p_2, \dots, p_m\}$ . Moreover, if  $p_i = \frac{1}{m}$  for all  $i \in [m]$ , then

$$P(\mathcal{N}_n = K_k^c) = \frac{k!}{m^n} \binom{m}{k} \binom{n-1}{k-1}.$$

*Proof.* Note that for  $\mathcal{N}_n$  to form a disjoint collection of  $k$  points, the intervals induced by the coloring must also be disjoint. This occurs if and only if all points of the same color are grouped together. Given  $k$  fixed colors it is well known that the disjoint groupings are counted by the number of compositions of  $n$  into  $k$  parts,  $\binom{n-1}{k-1}$ . Each composition occurs with

probability at least  $p_*^n$ . Finally, considering the  $\binom{m}{k}$  different ways to choose these  $k$  colors and the  $k!$  ways to order them, we have that,

$$P(\mathcal{N}_n = K_k^c) \geq p_*^n k! \binom{m}{k} \binom{n-1}{k-1}.$$

The last statement follows the same idea but here every  $k$ -coloring of the  $n$  points happens with probability  $\frac{1}{m}$ .  $\square$

Next we bound the probability that a particular edge is present in the random interval graph.

**Theorem 7.** *Under the key assumptions in Section 1 and for any pair  $\{i, j\}$ ,  $1 \leq i < j \leq m$ , the probability of event  $A_{ij} = \{\{i, j\} \in \mathcal{N}_n\}$ , i.e., that the edge  $\{i, j\}$  is present in the graph  $\mathcal{N}_n$  equals*

$$P(A_{ij}) = 1 - q_{ij}^n - \sum_{k=1}^n \binom{n}{k} \left[ \left( 2 \sum_{r=1}^{k-1} p_i^r p_j^{k-r} \right) + p_i^k + p_j^k \right] q_{ij}^{n-k},$$

where  $q_{ij} = 1 - (p_i + p_j)$ .

When  $p_i = \frac{1}{m}$  for all  $i \in [m]$ , then  $P(A_{ij}) = 1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n}$ .

*Proof.* We will find the probability of the complement,  $A_{ij}^c$ , which is the event where the two empirical supports do not intersect, i.e.,  $A_{ij}^c = \{I_n(i) \cap I_n(j)\} = \emptyset$ . Let  $C_i = \{\ell : Y_\ell = i, 1 \leq \ell \leq n\}$  and define  $C_j$  analogously. Note that  $A_{ij}^c$  can be expressed as the disjoint union of three events:

1.  $\{C_i \text{ and } C_j \text{ are both empty}\}$ ,
2.  $\{\text{Exactly one of } C_i \text{ or } C_j \text{ is empty}\}$ ,
3.  $\{C_i \text{ and } C_j \text{ are both non-empty but } I_n(i) \text{ and } I_n(j) \text{ do not intersect}\}$ .

The probability of the first event is simply  $q_{ij}^n$ . For the second event, assume for now that  $C_i$  will be the non-empty set and let  $k \in [n]$  be the desired size of  $C_i$ . There are  $\binom{n}{k}$  ways of choosing the locations of the  $k$  points in  $C_i$ . Once these points are chosen, the probability that these points receive label  $i$  and no others receive label  $i$  nor label  $j$  is exactly  $p_i^k q_{ij}^{n-k}$ . Summing over all values of  $k$  and noting that the argument where  $C_j$  is non-empty is analogous, we get that the probability of the second event is exactly  $\sum_{k=1}^n \binom{n}{k} (p_i^k + p_j^k) q_{ij}^{n-k}$ .

Now, note that the third event only occurs if all the points in  $C_i$  are to the left of all points in  $C_j$  or vice versa; for now assume  $C_i$  is to the left. Let  $k \in [n]$  be the desired size of  $C_i \cup C_j$  and let  $r \in [k-1]$  be the desired size of  $C_i$ . As before there are  $\binom{n}{k}$  ways of choosing the locations of the  $k$  points in  $C_i \cup C_j$ . Once these points are fixed, we know  $C_i$  has to be the first  $r$  many points,  $C_j$  has to be the remaining  $k-r$  points, and all other points cannot have label  $i$  nor label  $j$ . This occurs with probability  $p_i^r p_j^{k-r} q_{ij}^{n-k}$ . Finally, summing over all values of  $k$  and  $r$  and adding a factor of 2 to account for flipping the sides of  $C_i$  and  $C_j$  we get that the third event occurs with probability  $2 \sum_{k=1}^n \binom{n}{k} \sum_{r=1}^{k-1} p_i^r p_j^{k-r} q_{ij}^{n-k}$ .

Since  $A_{ij}^c$  is the disjoint union of these three events,  $P(A_{ij}^c)$  is equal to the sum of these three probabilities, which gives the desired result. For the uniform case, simply set  $p_i = p_j = 1/m$  in the general formula and note,

$$\begin{aligned} P(A_{ij}) &= 1 - \left(\frac{m-2}{m}\right)^n - \sum_{k=1}^n \binom{n}{k} \left[ \left(2 \sum_{r=1}^{k-1} \frac{1}{m^k}\right) + \frac{2}{m^k} \right] \left(\frac{m-2}{m}\right)^{n-k} \\ &= 1 - \left(\frac{m-2}{m}\right)^n - \frac{1}{m^n} \sum_{k=1}^n \binom{n}{k} 2k(m-2)^{n-k} \\ &= 1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n}. \end{aligned}$$

□

From this we can compute the expected number of edges in the random interval graph, which is the 1-skeleton of  $\mathcal{N}_n$ . The proof follows immediately from the above by the linearity of expectation.

**Corollary 8.** *Let  $X$  be the random variable equal to the number of edges in  $\mathcal{N}_n$ , the random interval graph. Under the key assumptions in Section 1,*

$$\mathbb{E}X = \sum_{1 \leq i < j \leq m} 1 - q_{ij}^n - \sum_{k=1}^n \left[ \binom{n}{k} \left( 2 \sum_{r=1}^{k-1} p_i^r p_j^{k-r} \right) + p_i^k + p_j^k \right] q_{ij}^{n-k},$$

where  $q_{ij} = 1 - (p_i + p_j)$ . In the uniform case where  $p_i = \frac{1}{m}$  for all  $i \in [m]$ , this expectation equals

$$\binom{m}{2} \left( 1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n} \right).$$

### 3 Maximum Degree, Cliques, and Behavior in the limit.

In this section we investigate the connectivity of the empirical nerve. First we study the maximum degree and clique number of  $\mathcal{N}_n$ . Then we show that as the number of samples  $n$  goes to infinity, then  $\mathcal{N}_n$  is almost surely the  $(m-1)$ -simplex.

#### 3.1 Maximum Degree.

The following result is a lower bound on the probability of finding an interval intersecting all others, i.e., that the maximum degree  $\text{Deg}(\mathcal{N}_n)$  of  $\mathcal{N}_n$  is  $m-1$ . In our birdwatching story this can be interpreted as the probability of finding a species which overlaps in time with all others.

In the following theorem we let  $\mathcal{X}_{m,k}^n$  denote the set of weak-compositions of  $n$  with length  $m$  containing exactly  $k$ -many non-zero parts [28, p. 25]. Formally,  $\mathcal{X}_{m,k}^n = \{(x_1, \dots, x_m) \in \mathbb{Z}_{\geq 0}^m : \sum_{i=1}^m x_i = n, |\{x_i : x_i \neq 0\}| = k\}$ . Also let  $M(x) = \frac{(x_1+x_2+\dots+x_m)!}{x_1!x_2!\dots x_m!} \prod_{i=1}^m p_i^{x_i}$  denote the multinomial distribution applied to the vector  $x \in \mathcal{X}_{m,k}^n$  considering the associated probabilities  $p_1, p_2, \dots, p_m$ . Finally, let  $S_n^k$  denotes the *Stirling numbers* of the second kind [28, p. 81].

**Theorem 9.** *Under the key assumptions in Section 1, the maximum degree of  $\mathcal{N}_n$  satisfies*

$$P(\text{Deg}(\mathcal{N}_n) = m-1) \geq$$

$$\max_r \left\{ \left[ 1 - \sum_{k=1}^{m-1} \frac{k^r}{m^r} \binom{m}{k} \sum_{x \in \mathcal{X}_{m,k}^r} M(x) (m-k)^r p_*^r \right] \left[ \sum_{x \in \mathcal{X}_{m,m}^{n-2r}} M(x) + \sum_{x \in \mathcal{X}_{m,m-1}^{n-2r}} M(x) \right] \right\}.$$

Moreover, in the uniform case where  $p_i = \frac{1}{m}$  for all  $i \in [m]$ , we have that

$$P(\text{Deg}(\mathcal{N}_n) = m-1) \geq$$

$$\max_r \left\{ \left[ 1 - \frac{m!}{m^{2r}} \sum_{k=1}^{m-1} \frac{(m-k)^r}{(m-k)!} S_r^k \right] \left[ \frac{m!}{m^{n-2r}} S_{n-2r}^m + \frac{(m-1)!}{(m-1)^{n-2r}} S_{n-2r}^{m-1} \right] \right\}.$$

*Proof.* Fix some  $1 \leq r \leq \frac{n-m}{2}$  and consider the sets  $L = \{1, 2, \dots, r\}$ ,  $C = \{r, r+1, \dots, n-(r+1)\}$  and  $R = \{n-r, n-(r-1), \dots, n\}$ . If the following events hold, we can guarantee that  $\text{Deg}(\mathcal{N}_n) = m-1$ .

$A = \{\text{There exists a chromatic class with points in } L \text{ and } R\},$   
 $B = \{\text{There exists at least one point with each of the remaining } m-1 \text{ colors in } C\}.$

In order to calculate  $P(A)$ , we will compute the probability of its complement  $A^c$ , i.e., the event where no color appears in both  $L$  and  $R$ . First we calculate the probability of  $L$  being colored with exactly  $k$  colors with  $1 \leq k \leq m-1$ . Observe that there are  $\binom{m}{k}$  ways to choose these colors and  $k^r \sum_{x \in \mathcal{X}_{m,k}^r} M(x)$  ways to color  $L$  with them. As there exist  $m^r$  different colorings with all the  $m$  colors, we have that for a fixed  $k$  the probability is

$$\frac{1}{m^r} k^r \binom{m}{k} \sum_{x \in \mathcal{X}_{m,k}^r} M(x).$$

In order for  $A^c$  to occur, we need that  $R$  be colored with only the  $(m-k)$  remaining colors. Note that this event is independent from the coloring of  $L$  as the two sets are disjoint. There are  $(m-k)^r$  different ways of coloring  $R$ , and each occurs with probability at most  $p_*^r$ , where  $p_* = \max\{p_i : 1 \leq i \leq m\}$ . Thus, for a fixed  $k$  we have that the probability that no color appears in both  $L$  and  $R$  is at most

$$\left[ \frac{1}{m^r} k^r \binom{m}{k} \sum_{x \in \mathcal{X}_{m,k}^r} M(x) \right] [(m-k)^r p_*^r].$$

Then, by summing over all  $k$  we have that

$$P(A^c) \leq \sum_{k=1}^{m-1} \left[ \frac{1}{m^r} k^r \binom{m}{k} \sum_{x \in \mathcal{X}_{m,k}^r} M(x) \right] [(m-k)^r p_*^r],$$

which implies that

$$P(A) \geq 1 - \sum_{k=1}^{m-1} \frac{1}{m^r} k^r \binom{m}{k} \sum_{x \in \mathcal{X}_{m,k}^r} M(x) (m-k)^r p_*^r.$$

To compute  $P(B)$ , note that the probability of coloring  $C$  with  $m$  or  $m-1$  colors is exactly

$$\sum_{x \in \mathcal{X}_{m,m}^{n-2r}} M(x) + \sum_{x \in \mathcal{X}_{m,m-1}^{n-2r}} M(x).$$

Finally, as  $A$  and  $B$  are independent events, we have  $P(\text{Deg}(\mathcal{N}_n) = m-1)$  is greater than

$$\left[1 - \sum_{k=1}^{m-1} \frac{1}{m^r} k^r \binom{m}{k} \sum_{x \in \mathcal{X}_{m,k}^r} M(x) (m-k)^r p_*^r\right] \left[ \sum_{x \in \mathcal{X}_{m,m}^{n-2r}} M(x) + \sum_{x \in \mathcal{X}_{m,m-1}^{n-2r}} M(x) \right].$$

Maximizing over  $r$  gives the desired result. For the case uniform, we just apply  $p_* = 1/m$  and use the former equality together with the fact that  $k!/k^n S_n^k = \sum_{x \in \mathcal{X}_{m,k}^n} M(x)$ .  $\square$

### 3.2 Cliques.

The expected clique number of  $\mathcal{N}_n$  is of special interest to us. In our bird-watching story this corresponds to the maximal subset of birds whose time intervals all intersect. While we do not compute the expected clique number exactly, the next theorem provides a lower bound on the expected clique number which performs very well in simulations (see the Appendix).

**Lemma 10.** *Under the key assumptions in Section 1, the probability that an arbitrary point  $x \in [n]$  lies inside the interval of color  $i$ ,  $I_n(i)$ , is exactly  $1 - q_i^x - q_i^{n-x+1} + q_i^n$ , where  $q_i = 1 - p_i$ .*

*Proof.* Fix an arbitrary  $x \in [n]$  and define the event  $A = \{x \in I_n(i)\}$ . Note that in order for  $A$  to occur either  $x$  lies between two points with label  $i$  or  $x$  itself is labeled  $i$ . Now consider the complementary probability event,  $A^c = \{x \notin I_n(i)\}$ . Next define the events  $L, R$  where  $L = \{\text{none of the points less than or equal to } x \text{ have label } i\}$  and  $R = \{\text{none of the points greater than or equal to } x \text{ have label } i\}$ . Note  $A^c = L \cup R$  and  $P(L) = q_i^x$ ,  $P(R) = q_i^{n-x+1}$  and  $P(L \cap R) = q_i^n$ . Therefore, by the inclusion-exclusion principle we have,

$$P(A^c) = P(L) + P(R) - P(L \cap R) = q_i^x + q_i^{n-x+1} - q_i^n,$$

and hence  $P(A) = 1 - q_i^x - q_i^{n-x+1} + q_i^n$ .  $\square$

**Theorem 11.** *Let  $\omega$  be the random variable equal to the clique number of  $\mathcal{N}_n$ , i.e., the size of the largest clique in the 1-skeleton of  $\mathcal{N}_n$ . Under the key assumptions in Section 1 we have*

$$\mathbb{E} \omega \geq \sum_{i=1}^m (1 - q_i^{\lceil \frac{n}{2} \rceil} - q_i^{n - \lceil \frac{n}{2} \rceil + 1} + q_i^n)$$



where  $q_i = 1 - p_i$ . Moreover, in the uniform case where  $p_i = \frac{1}{m}$  for all  $i \in [m]$ , we have that

$$\mathbb{E} \omega \geq m - \left(\frac{m-1}{m}\right)^{\lceil \frac{n}{2} \rceil} - \left(\frac{m-1}{m}\right)^{n - \lceil \frac{n}{2} \rceil + 1} + \left(\frac{m-1}{m}\right)^n.$$

*Proof.* By the preceding lemma we know that the probability that  $x \in I_n(i)$  for some  $x \in [n]$  is exactly  $1 - q_i^x - q_i^{n-x+1} + q_i^n$ . To maximize this quantity over  $x \in [n]$  we will first minimize  $f(x) = q_i^x + q_i^{n-x+1} - q_i^n$  over all  $x$ . Note  $f$  is convex so a simple calculus exercise shows that  $f$  is minimized at  $x^* = \frac{n+1}{2}$ . This can also be seen directly from the fact that  $f$  is convex and symmetric about  $\frac{n+1}{2}$ . When  $n$  is odd the minimizer  $x^*$  is an integer and lies in  $[n]$ . To handle the case when  $n$  is even, note that  $f$  is symmetric about the minimizer  $x^*$ . Therefore, when  $n$  is even,  $f$  is minimized over  $[n]$  at the integers closest to  $x^*$ , which are  $\frac{n}{2}$  and  $\frac{n}{2} + 1$ . We see then that  $f$  is minimized over  $[n]$  at the point  $x = \lceil \frac{n}{2} \rceil$ , which holds whether  $n$  is even or odd.

Now, for  $i = 1, \dots, m$  let  $X_i$  be the indicator random variable which equals 1 if  $\lceil \frac{n}{2} \rceil \in I_n(i)$  and is 0 otherwise and set  $X = \sum_{i=1}^m X_i$ , so  $X$  counts the number of intervals containing the point  $\lceil \frac{n}{2} \rceil$ . Note that the clique number  $\omega \geq X$ , so

$$\mathbb{E} \omega \geq \mathbb{E} X = \sum_{i=1}^m \mathbb{E} X_i = \sum_{i=1}^m P(X_i) = \sum_{i=1}^m (1 - q_i^{\lceil \frac{n}{2} \rceil} - q_i^{n - \lceil \frac{n}{2} \rceil + 1} + q_i^n).$$

The result for the uniform case follows directly by setting  $p_i = \frac{1}{m}$  for every  $i$ .  $\square$

### 3.3 Behavior of the nerve complex as the number of samples goes to infinity.

Note that as the number of samples  $n$  grows large, Theorem 11 implies that the expected clique number  $\mathbb{E} \omega \rightarrow m$ . Since  $\omega$  only takes values in  $\{1, \dots, m\}$  it follows that the clique number also converges to  $m$  in probability. Thus, as  $n$  goes to infinity, the probability that the nerve of the observations is the  $(m-1)$ -simplex denoted by  $\Delta_{m-1}$ , i.e., a complete graph, goes to 1. In our birdwatcher analogy, this implies that with sufficiently many observations one is almost sure to find a common interval of time where all  $m$  species can be observed. The following theorem provides a lower bound on this convergence.

**Theorem 12.** *Under the key assumptions in section 1, the probability that  $\mathcal{N}_n$  is an  $(m-1)$ -simplex (or equivalently the graph is a complete graph  $K_m$ ) satisfies*

$$P(\mathcal{N}_n = \Delta_{m-1}) \geq \left( \sum_{x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}} M(x) \right)^2$$

where  $\mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor} = \{(x_1, x_2, \dots, x_m) \in \mathbb{N}^m : \sum_{i=1}^m x_i = \lfloor \frac{n}{2} \rfloor\}$ .

In the uniform case where  $p_i = \frac{1}{m}$  for every  $i \in [m]$ , this gives that

$$P(\mathcal{N}_n = \Delta_{m-1}) \geq \left( \frac{m!}{m^{\lfloor \frac{n}{2} \rfloor}} S_{\lfloor \frac{n}{2} \rfloor}^m \right)^2$$

where, again,  $S_n^k$  denotes the Stirling numbers of the second kind.

*Proof.* For each vector  $x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}$  the multinomial  $M(x)$  computes the probability that there exist exactly  $x_i$  points with color  $i$  for every  $1 \leq i \leq m$ . Therefore, the sum over all the vectors of  $\mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}$  gives us the probability of having at least one point of each color.

Now, we consider the events  $L = \{\text{the first } \lfloor \frac{n}{2} \rfloor \text{ points are colored with exactly } m \text{ colors}\}$  and  $R = \{\text{the last } \lfloor \frac{n}{2} \rfloor \text{ points are colored with exactly } m \text{ colors}\}$ . We have

$$P(L) = P(R) = \sum_{x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}} M(x).$$

Then  $P(\mathcal{N}_n = \Delta_{m-1}) \geq P(L \cap R)$  and as  $L$  and  $R$  are independent events, we conclude

$$P(\mathcal{N}_n = \Delta_{m-1}) \geq P(L \cap R) = P(L)P(R) = \left( \sum_{x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}} M(x) \right)^2.$$

The result for the uniform case follows because  $k!/k^n S_n^k = \sum_{x \in \mathcal{X}_{m,k}^n} M(x)$ .  $\square$

Theorem 12 tells us how likely it is for the empirical nerve of  $n$  samples to form the  $(m-1)$ -simplex for fixed  $n$ . A related question asks what is the *first* observation  $n$  for which this occurs, i.e., if we have a sequence of observations  $Y_1, Y_2, \dots$  what is the least  $n$  such that  $\mathcal{N}_n((Y_1, \dots, Y_n)) = \Delta_{m-1}$ ? We call this quantity the *waiting time* to form the  $(m-1)$ -simplex and provide a lower bound on its expectation below.

**Theorem 13.** *Let  $X$  be the random variable for the waiting time until  $\mathcal{N}_n = \Delta_{m-1}$ , explicitly  $X = \inf\{n \in \mathbb{N} : \mathcal{N}_n = \Delta_{m-1}\}$ . Then, under the key assumptions in Section 1, we have  $\mathbb{E}X \leq 2 \int_0^\infty \left(1 - \prod_{i=1}^m (1 - e^{-p_i x})\right) dx$ . Moreover, in the uniform case, where  $p_i = \frac{1}{m}$  for all  $i \in [m]$ , we have that  $\mathbb{E}X \leq 2m \sum_{i=1}^m \frac{1}{i}$ .*

*Proof.* The results follow directly from the expected waiting time of the classical coupon collector problem. Let  $Z$  denote the waiting time until we have observed every label, i.e.,  $Z$  is the waiting time until we have completed a collection of coupons if each coupon is an i.i.d. random variable that takes value  $i$  with probability  $p_i$ . It is known that  $\mathbb{E}Z = 2 \int_0^\infty \left(1 - \prod_{i=1}^m (1 - e^{-p_i x})\right) dx$ , and in the uniform case where  $p_i = \frac{1}{m}$  for all  $i \in [m]$ ,  $\mathbb{E}Z = m \sum_{i=1}^m \frac{1}{i}$  (see [7] for several detailed proofs). Now, note that  $\mathcal{N}_n = \Delta_{m-1}$  if we complete a collection, then complete a second collection, disjoint from the first. Let  $Z_1$  denote the waiting time to complete the first collection, and let  $Z_2$  be the additional waiting time to complete a second collection. Then  $X \leq Z_1 + Z_2$  and  $Z_1, Z_2$  are equal in distribution to  $Z$ , so  $\mathbb{E}X \leq \mathbb{E}(Z_1 + Z_2) = 2\mathbb{E}Z$ .  $\square$

## 4 Conclusion.

In this paper we introduced a novel random interval graph model. It is well-suited for applications involving the overlap patterns of chronological observations. There are a number of natural fascinating questions for the curious reader. First, obviously the distribution of birds varies in time as seasonal changes (or other factors such as predators or climate change) affect the species, thus the non-stationary case is better for applications. We ask ourselves, which of the results can be extended to the non-stationary case when the key assumptions made here are no longer valid?

Second, Hanlon presented in [15] a characterization of all interval graphs using a unique interval representation. He used this to enumerate all interval graphs. The analysis we presented in Theorem 12 indicates that, when we use our stochastic process to generate random intervals on the line, the probability of getting an interval graph other than the complete graph goes to 0 as the number of samples  $n$  goes to infinity. A natural challenge is to understand the decay of probabilities for different classes of graphs, for instance, random *interval trees* (see [5]).

Finally, the story we presented is about data samples indexed by a single parameter, say time. But what happens when geographical coordinates,

temperature, humidity, or other parameters are considered to model the distribution of birds? Extending the model to higher-dimensions produces new challenges. For example, the random interval graphs are no longer sufficient to capture all the information. Instead, one needs to investigate random simplicial complexes (see [3]) because we lose the natural order for the points that we have in the line. This implies that an equivalent result to Lemma 1 is no longer possible. For instance, continuing with our birdwatcher’s analogy, suppose that colored points in Figure 4 represent the geographic coordinates of three different types of birds that have been studied. If our birdwatcher is trying to determine the usual habitat and the territorial interactions between them he/she will face the problem that two very similar data sets will induce different simplicial complexes.

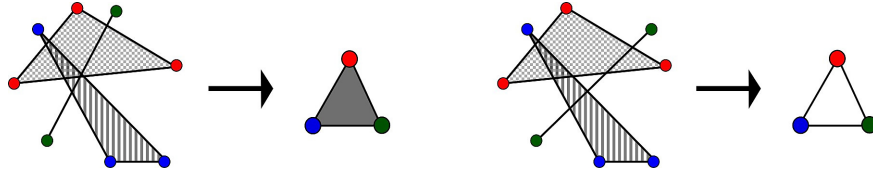


Figure 4: Two data sets of 3 different bird species with the same order type inducing different simplicial complexes.

**Acknowledgements.** The first and second authors gratefully acknowledge partial support from NSF DMS-grant 1818969. The second author also acknowledges support from the NSF-AGEP supplement. Finally, the third and fourth authors gratefully acknowledge partial support from PAPIIT IG100721 and CONACyT 282280.

## References

- [1] Barvinok, A. (2002). *A Course in Convexity*. Graduate Studies in Mathematics, Vol. 54. Providence, RI: American Mathematical Society.
- [2] Cohen, J. E., Komlós, J., Mueller, T. (1979). The probability of an interval graph, and why it matters. In: Ray-Chaudhuri, D. K., ed. *Proceedings of the Symposia in Pure Mathematics*, Vol. 34. Providence, RI: American Mathematical Society, pp. 97–115.
- [3] De Loera, J. A., Hogan, T. (2020). Stochastic Tverberg theorems with applications in multiclass logistic regression, separability, and center-

- points of data. *SIAM Jour. on Math. of Data Science*. 2:1151–1166. doi.org/10.1137/19M1277102.
- [4] Diaconis, P., Holmes, S., Janson, S. (2013). Interval graph limits. *Ann. of Comb.* 17(1):27–52. doi.org/10.1007/s00026-012-0175-0.
  - [5] Eckhoff, J. (1993). Extremal interval graphs. *Jour. Graph Theory*. 17(1):117–127. doi.org/10.1002/jgt.3190170112.
  - [6] Erdős, P., Renyi, A. (1959). On random graphs i. *Publ. Math. Debrecen*. 6(18):290–297.
  - [7] Ferrante, M., Saltalamacchia, M. (2014). The coupon collector’s problem. *MATerials MATematics*. 2014(2):35.
  - [8] Fishburn, P. C. (1985). *Interval orders and interval graphs: a study of partially ordered sets*. New York: Wiley.
  - [9] Flajolet, P., Gardy, D., Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Appl. Math.* 39(3):207–229. doi.org/10.1016/0166-218X(92)90177-C.
  - [10] Fulkerson, D. R., Gross, O. A. (1965). Incidence matrices and interval graphs. *Pacific Jour. Math.* 15(3):835–855. doi.org/10.2140/PJM.1965.15.835.
  - [11] Ghrist, R. (2014). *Elementary Applied Topology*. CreateSpace.
  - [12] Gilmore, P. C., Hoffman, A. J. (1964). A characterization of comparability graphs and of interval graphs. *Canadian Jour. of Math.* 16:539–548. doi.org/10.4153/CJM-1964-055-5.
  - [13] Golumbic, M. C. (2004). Interval graphs. In: Golumbic, M. C., ed. *Algorithmic Graph Theory and Perfect Graphs*. Annals of Discrete Mathematics, Vol. 57. Amsterdam, Netherlands: North Holland, pp. 171 – 202.
  - [14] Hammer, O., Harper, D. A. T. (2007). *Paleontological Data Analysis*. Oxford, UK: Blackwell Publishing.
  - [15] Hanlon, P. (1982). Counting interval graphs. *Trans. of the Amer. Math. Soc.* 272(2):383–426. doi.org/10.1090/S0002-9947-1982-0662044-8.

- [16] Hatcher, A. (2002). *Algebraic Topology*. Cambridge, UK: Cambridge University Press.
- [17] Iliopoulos, V. (2017). A study on properties of random interval graphs and Erdős Renyi graph  $\mathcal{G}(n, 2/3)$ . *Jour. of Discrete Math. Sci. and Cryptography*. 20(8):1697–1720. doi.org/10.1080/09720529.2016.1184453.
- [18] Justicz, J., Scheinerman, E. R., Winkler, P. (1990). Random intervals. *Amer. Math. Monthly*. 97(10):881–889. doi.org/10.1080/00029890.1990.11995679.
- [19] Kozlov, D. N. (2008). *Combinatorial Algebraic Topology*. Algorithms and Computation in Mathematics, Vol. 21. Berlin, Germany: Springer.
- [20] Krylov, N. (2000). *Introduction to the Theory of Random Processes*. Graduate Studies in Mathematics, Vol. 43. Providence, RI: American Mathematical Society.
- [21] Lekkerkerker, C., Boland, J. (1962). Representation of a finite graph by a set of intervals on the real line. *Fundam. Math.* 51(1):45–64.
- [22] Matousek, J. (2002). *Lectures on Discrete Geometry*. Graduate Texts in Mathematics, Vol. 212. New York: Springer.
- [23] Pippenger, N. (1998). Random interval graphs. *Random Struct. Algorithms*. 12(4):361–380. doi.org/10.1002/(SICI)1098-2418(199807)12:4<361::AID-RSA4>3.0.CO;2-R.
- [24] Ross, S. (1996). *Stochastic processes*, 2nd ed. New York: Wiley.
- [25] Scheinerman, E.R. (1988). Random interval graphs. *Combinatorica*, 8(4):357–371. doi.org/10.1007/BF02189092.
- [26] Scheinerman, E.R. (1990). An evolution of interval graphs. *Discrete Math.* 82(3):287–302. doi.org/10.1016/0012-365X(90)90206-W.
- [27] Schelling, H.V. (1954). Coupon collecting for unequal probabilities. *Amer. Math. Monthly*. 61(5):306–311. doi.org/10.1080/00029890.1954.11988466.
- [28] Stanley, R. (2011). *Enumerative Combinatorics*, Vol. 1, 2nd ed. Cambridge, UK: Cambridge University Press.

- [29] Tancer, M. (2013). Intersection patterns of convex sets via simplicial complexes: a survey. In: Pach, J., ed. *Thirty essays on geometric graph theory*. New York: Springer, pp. 521–540.

**J. A. De Loera** is a professor of Mathematics at the University of California, Davis. His main mathematical themes are discrete geometry and combinatorial optimization. He enjoys walking with his dog Bolo and watching coyotes roam the fields near his house.

Department of Mathematics, University of California, Davis  
deloera@math.ucdavis.edu

**E. Jaramillo-Rodriguez** is a PhD candidate in Mathematics at the University of California, Davis. Edgar is writing their thesis on stochastic combinatorial geometry applied to data science and machine learning. Edgar likes spending time outdoors with good friends or good books.

Department of Mathematics, University of California, Davis  
ejaramillo@ucdavis.edu

**D. Oliveros** is a professor at the Institute of Mathematics at the National Autonomous University of Mexico UNAM (Campus Juriquilla). Her areas of interest in mathematics are discrete and computational geometry and convexity. She enjoys dancing, gardening and playing with her dogs.

Instituto de Matemáticas, Universidad Nacional Autónoma de México  
doliveros@im.unam.mx

**A. J. Torres** is a doctoral student in Mathematics at the National Autonomous University of Mexico UNAM. His main areas of interest include discrete geometry, data analysis, and combinatorics. He enjoys jogging around the city and hiking the trails near his hometown in Querétaro.

Instituto de Matemáticas, Universidad Nacional Autónoma de México  
antonio.torres@im.unam.mx

## 5 Appendix.

### 5.1 Experimental Results.

In Theorems 9, 11, and 12 we provided lower bounds on the likelihood of various events occurring given  $n$  points and  $m$  labels. To study the usefulness of these bounds we ran simulations. For each pair  $(m, n)$  we randomly colored  $n$  points on the real line using  $m$  colors with uniform probability (each color was equally likely) then constructed the induced interval graph. We repeated this process 100 times for each pair  $(m, n)$  and plotted the percentage of the simulations where the desired event occurred. We also plotted our lower bounds from the theorems above and found that, in general, our bounds perform well for most values of  $m$  and  $n$ .

Figure 5 compares the bound on the maximum degree obtained in Theorem 9 and the empirical approximation generated by our simulations. Figure 6 compares the bound on the expected clique number obtained in Theorem 11 and the empirical approximation generated by our simulations. Figure 7 compares the bound on the probability of the nerve being the  $(m - 1)$  simplex obtained in Theorem 12 and the empirical approximation generated by our simulations.

Finally, we also compared the probability of obtaining  $m - 1$  as a maximum degree  $\text{Deg}(\mathcal{N}_n)$  in the Scheinerman model with our model. In [18], the authors prove in a clever way, that this probability is exactly  $2/3$  and their result does not depend on the number of intervals. On the other hand, in our model this probability depends on both, the number of points and the number of colors that we use. We generated 1000 random  $m$ -colorings, for Scheinerman's model (the solid line)  $1 \leq m \leq 50$ . For our model we use several values of  $n$  with  $1 \leq m \leq n$ . The results are displayed in Figure 8.



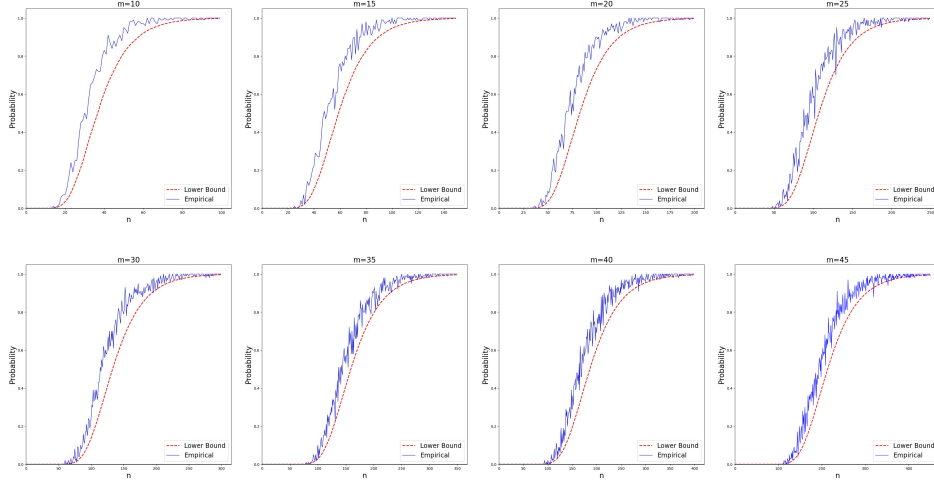


Figure 5: Probability of  $\text{Deg}(\mathcal{N}_n) = m - 1$ , simulations compared to bound from Theorem 9.

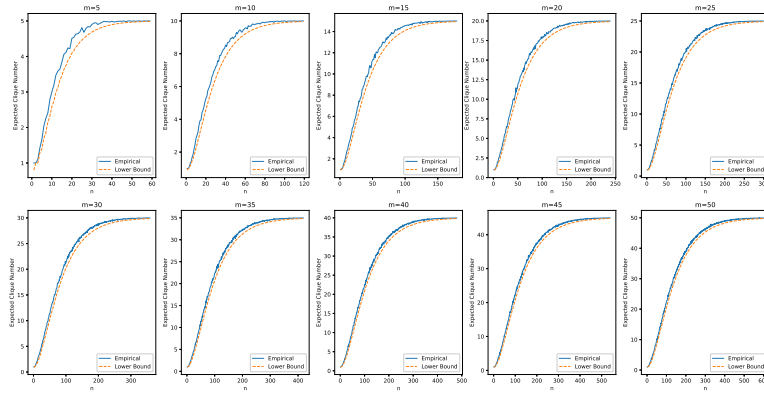


Figure 6: Expected clique number of  $\mathcal{N}_n$  with uniform coloring as a function of  $n$ .

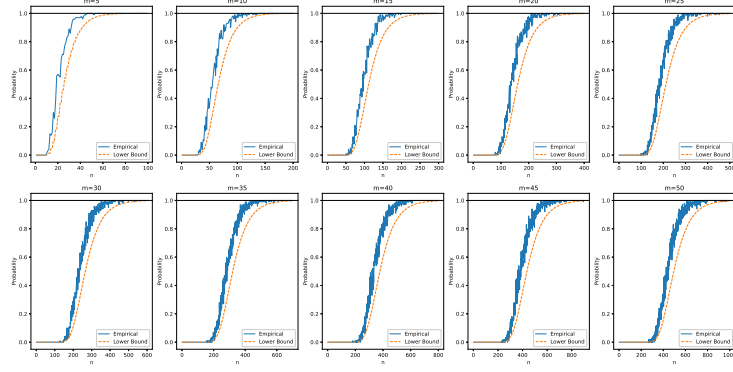


Figure 7: Probability that  $\mathcal{N}_n = \Delta_{m-1}$ .

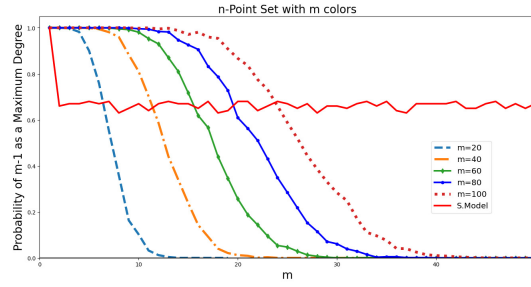


Figure 8: Comparison between Scheinerman's model and ours with the probability that  $\text{Deg}(\mathcal{N}_n) = m - 1$ .