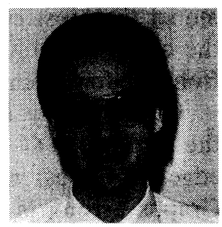
Random Intervals

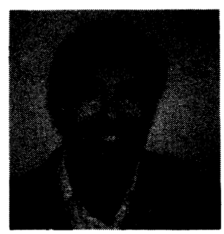
Joyce Justicz, Edward R. Scheinerman, and Peter M. Winkler

Joyce Justicz is a graduate student in mathematics at Emory University, where she received her bachelor's degree (as valedictorian) in 1985.

Edward Scheinerman is an associate professor in the Department of Mathematical Sciences of the Johns Hopkins University. He received his Sc.B. from Brown University in 1980 and his Ph.D. from Princeton in 1984.



Peter Winkler is Professor of Mathematics and Computer Science at Emory University, and manager of the Research Group in Mathematics and Theoretical Computer Science at Bellcore. His primary mathematical interests are combinatorics, logic and probabilistic methods.



Abstract. Fix a large number n and let $\{x_1, y_1, \dots, x_n, y_n\}$ be $2n$ points chosen independently from some fixed continuous probability distribution on the real line. Each pair $\{x_i, y_i\}$ determines a *random interval* $[\min(x_i, y_i), \max(x_i, y_i)]$. We examine the structure of the resulting family of intervals, and in particular answer the following questions: how large a subcollection of pairwise disjoint intervals can one expect to find? And, what is the probability that there is an interval in the family which intersects all the others?

Prelude. Before beginning, we invite the reader to test his intuition on the following problem (it won't be any worse than ours was!). The numbers from 1 to $2n$, with n large, are paired at random, each pair being regarded as the endpoints of a real interval. What is the probability that among these n intervals there is one which meets all the others?

1. Introduction: Random Intervals. In studying any type of combinatorial structure it is useful, and sometimes quite illuminating, to have models for "random" structures of that type. The most famous example is the random graph model of

Erdös and Rényi [2] in which edges are chosen independently with fixed probability. Random graph theory now forms a substantial subject of study in itself.

The combinatorial structures with which we concern ourselves here are those based on intersection properties of families of intervals. The most obvious definition of a "random interval" is the interval between two random numbers x and y; for convenience we denote that interval by $[x, y]$ even when $y < x$, so that by definition $[x, y] = [y, x]$.

We create a family F of random intervals in the following way: fix a number n and a continuous probability distribution on the real line. Choose values $x_1, y_1, \ldots, x_n, y_n$ independently and let F consist of intervals $I_1, \ldots, I_n$ where $I_j = [x_j, y_j]$.

Since intersection properties of F depend only on the order of the points $x_1, \ldots, y_n$, it is clear that the model is insensitive to the choice of distribution. In fact, an equivalent discrete model is obtained by choosing at random one of the $(2n)!$ assignments of $x_1, \ldots, y_n$ to the integers from 1 to $2n$. In the continuous model the most convenient distribution seems to be the uniform distribution on the unit interval $[0, 1]$; in that case we may think of the intervals as having arisen from random points (x_i, y_i) in the unit square.

To any family F of intervals we may associate a graph $G = (V, E)$ as follows: the vertices of G correspond to the intervals of F, and two vertices constitute an edge of G just when their corresponding intervals have non-null intersection. Graphs arising in this fashion are called *interval graphs*; when F is a family of random intervals as described above, they are called *random interval graphs* [**8**].

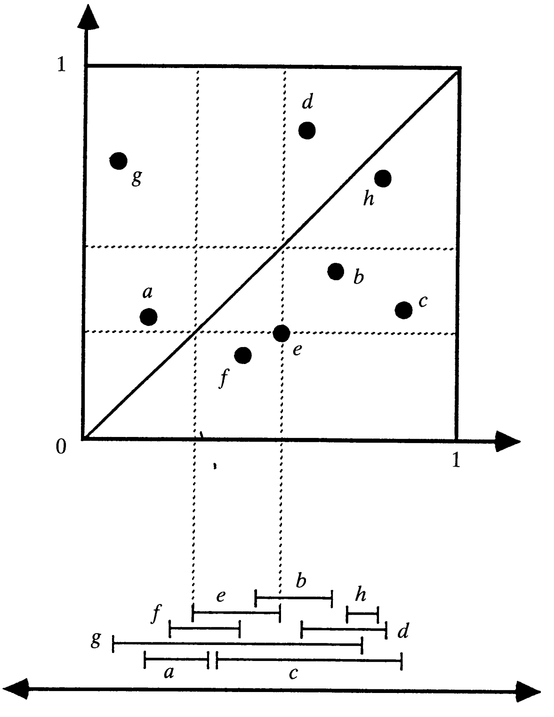


FIG. 1a. Eight random points and their intervals.

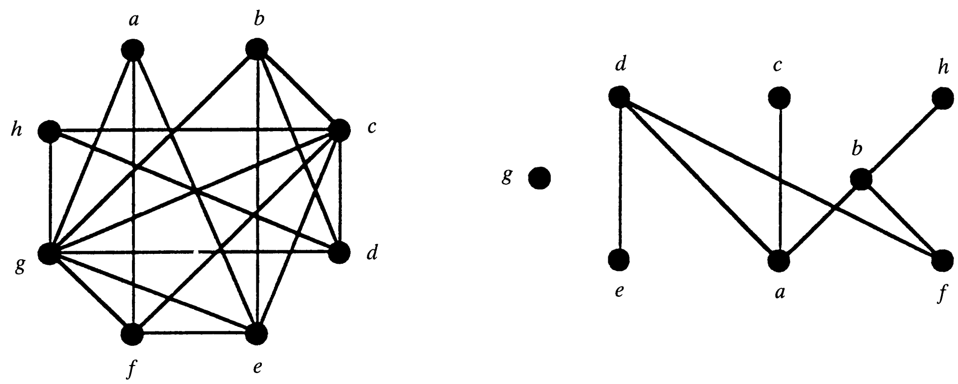


FIG. 1b. Random interval graph and interval order.

We preserve somewhat more of the properties of a family F of intervals by assigning to F a partially ordered set P; here the elements of P correspond to the intervals of F with $x \prec y$ just when the interval corresponding to x lies entirely to the left of the interval corresponding to y. Partially ordered sets arising in this fashion are called *interval orders*. Note that G may be obtained from P by declaring $\{x, y\}$ to be an edge of G whenever x and y are incomparable in P.

For further information about both interval graphs and interval orders, we refer the reader to Fishburn's excellent book [**4**].

FIGURE 1a depicts 8 random intervals (arising from 8 points chosen in the unit square) and FIGURE 1b depicts the corresponding interval graph and interval order.

2. Chains and Antichains.... or cliques and independent sets.

Given a collection of n random intervals, what is the size of a largest family of pairwise intersecting intervals? ...or pairwise disjoint intervals? In poset language we are asking for the sizes of the largest antichain (pairwise incomparable elements) and of the largest chain (pairwise comparable, and therefore totally ordered, elements). In graph language we seek the sizes of the largest clique (pairwise adjacent) and the largest independent set (pairwise nonadjacent).

The first question (pairwise intersecting) is easier to answer.

Intervals of reals satisfy the "Helly property," that is, every collection of pairwise-intersecting intervals has a nonempty intersection. Thus, when endpoints are uniformly distributed on $[0, 1]$, it is enough to find the point $x \in [0, 1]$ which is contained in the maximum number of intervals in the collection. The probability that an interval contains x is $1 - x^2 - (1 - x)^2 = 2x - 2x^2$. This is maximized when $x = 1/2$, and the expected number of intervals which contain $1/2$ is $n/2$. With a bit more work (the interested reader can consult [**8**]) one can obtain the following result:

THEOREM 1. *Let the random variable A_n denote the size of a largest set of pairwise intersecting intervals in a family of n random intervals. There exists a function $f(n)$ such that*

$$\lim_{n \to \infty} \frac{f(n)}{n} = 0$$

and

$$\lim_{n\to\infty} \Pr\left\{\frac{n}{2} - f(n) \leqslant A_n \leqslant \frac{n}{2} + f(n)\right\} = 1.$$

We turn now to the second question: What is the largest collection of pairwise non-intersecting intervals (longest chain) in a family of n random intervals? It turns out that this question is closely related to the following well-known problem, due to Stanislaw Ulam: What is the length of the longest increasing subsequence of a random permutation of $1,\dots,n$?

Ulam's problem, posed in **[10]**, was tackled by Hammersley in a famous paper **[6]** in which the answer was shown to be asymptotic to $c\sqrt{n}$ for some constant c. Later, a combination of the efforts of Schensted **[9]**, Logan and Shepp **[7]** and Veršik and Kerov **[11]** determined that the constant is, in fact, 2.

Ulam's "random permutation" may be constructed by choosing n random points from the uniform distribution in the unit square, and numbering the points from left to right (i.e., according to their x-coordinates). The permutation arises in reading the label-numbers from bottom to top.

The points determine a partially ordered set Q under the product ordering, where $(x_i, y_i) \preceq_Q (x_j, y_j)$ just when $x_i \leqslant x_j$ and $y_i \leqslant y_j$. An increasing subsequence of the permutation corresponds exactly to a chain in Q; geometrically, a subset of the points extending from southwest to northeast in the square. Let us denote by X_n the random variable whose value is the size of the largest such subset.

As we have seen, the same set of points in the unit square determines a family of random intervals, but in the interval order P, we have the stronger condition $(x_i, y_i) \prec_P (x_j, y_j)$ iff $\max(x_i, y_i) < \min(x_j, y_j)$. Let us denote by Y_n the size of a largest chain in P; it is the asymptotic behavior of Y_n which we wish to determine. Since every chain of P is a chain of Q, we have $Y_n \leqslant X_n$.

For example, consider the point **e** in FIGURE 1a. The points greater than **e** in Q are those to the northeast (namely, **b**, **c**, **d** and **h**). We can also describe geometrically the points which lie above **e** in P. Draw a vertical and horizontal line through the point **e**. Also, draw a vertical and horizontal line through **e**'s reflection across the positively sloped diagonal of the square. These four lines divide the unit square into 9 rectangles. (See FIGURE 1a.) The upper right-hand square contains the points which correspond to intervals to the right of **e** (namely, **d** and **h**).

Hammersley **[6]** used the method of subadditivity to prove that $X_n/\sqrt{n}$ approaches a constant in probability, and in fact a nearly identical argument can be made to achieve the same result for $Y_n/\sqrt{n}$. (See, for example, Bolobás and Winkler **[1]** where Hammersley's results are extended to higher dimension.)

In the case of Ulam's problem the actual value of c was obtained only with great difficulty, and in fact the values of the constants in higher dimensions remain unknown. Luckily, a special feature of the interval case allows us both to prove the existence of the constant and determine its value with relative ease. The difference is that in our interval problem a maximum-length chain can be built in "greedy" fashion from the bottom up, whereas the equivalent process in Ulam's problem fails by a constant factor, in the limit, to attain maximal length.

THEOREM 2. *Let Y_n denote the maximum number of pairwise disjoint intervals in a family of n random intervals. Then*

$$\lim_{n\to\infty} \frac{Y_n}{\sqrt{n}} = \frac{2}{\sqrt{\pi}}$$

in probability.

This means for every $\varepsilon > 0$, there exists an n_0 so that for all $n > n_0$,

$$\Pr\left\{\left|\frac{Y_n}{\sqrt{n}} - \frac{2}{\sqrt{\pi}}\right| \leqslant \varepsilon\right\} > 1 - \varepsilon.$$

Proof. Let us establish a Poisson process on the plane of density 1. We select an infinite chain $C = \{(u_1, v_1), (u_2, v_2), \dots\}$ of points of the process in the following manner: (u_1, v_1) is the point in the positive quadrant which minimizes $\max(u_1, v_1)$, and thereafter, (u_k, v_k) is the point above (u_{k-1}, v_{k-1}) which minimizes $\max(u_k, v_k)$. In terms of intervals, C represents the chain built from the bottom by always selecting the interval with least possible upper endpoint; it is easily seen by induction that in any finite collection of intervals such a chain has maximum possible length.

Now, for any positive real s, the region $\{(x, y): 0 \leqslant x, y \leqslant s\} = [0, s]^2$ is with probability exactly $\exp\{-s^2\}$ unoccupied by a point of the Poisson process. It follows that if S is the random variable whose value is $\max(x_1, y_1)$, then the mass density of S is given by the function

$$f(s) = \frac{d}{ds}(1 - e^{-s^2}) = 2se^{-s^2}$$

whose expected value is

$$\int_0^\infty 2s^2 e^{-s^2}\, ds = \int_0^\infty t^{1/2} e^{-t}\, dt = \Gamma(3/2) = \frac{\sqrt{\pi}}{2}.$$

The differences

$$\max(u_1, v_1) - 0, \max(u_2, v_2) - \max(u_1, v_1), \max(u_3, v_3) - \max(u_2, v_2), \dots$$

are independent and identically distributed with mean $\sqrt{\pi}/2$. It follows from the law of large numbers that for any $\varepsilon > 0$,

$$(1 - \varepsilon)\frac{\sqrt{\pi}}{2} < \frac{\max(x_m, y_m)}{m} < (1 + \varepsilon)\frac{\sqrt{\pi}}{2}$$

with probability at least $1 - \varepsilon$, for every sufficiently large m.

Now let $r(n)$ be the least r such that $[0, r]^2$ contains exactly n points of the Poisson process; these points then determine a family of n random intervals, as described above, and we may therefore identify Y_n with the largest m such that (u_m, v_m) lies in the square $[0, r(n)]^2$. Since the Poisson process has density 1, we will have

$$(1 - \varepsilon)\sqrt{n} < u(n) < (1 + \varepsilon)\sqrt{n}$$

with probability at least $1 - \varepsilon$, for sufficiently large n.

Let $m_1 = \lfloor (1 - \varepsilon)(2/\sqrt{\pi})\sqrt{n} \rfloor$ and $m_2 = \lfloor (1 + \varepsilon)(2/\sqrt{\pi})\sqrt{n} \rfloor$; then for large enough n, we have that (u_{m_1}, v_{m_1}) will lie inside the square $[0, r(n)]^2$ and (u_{m_2}, v_{m_2}) outside the square, with probability at least $1 - \varepsilon$. We conclude that

$$(1 - \varepsilon)\frac{2}{\sqrt{\pi}} < \frac{Y_n}{\sqrt{n}} < (1 + \varepsilon)\frac{2}{\sqrt{\pi}}$$

with probability at least $1 - \varepsilon$, proving the theorem.

Note: The Poisson process enables us to place the random variables Y_n all in the same sample space; thus, with the help of the strong law of large numbers, we may obtain the slightly stronger result that

$$\lim_{n\to\infty} \frac{Y_n}{\sqrt{n}} = \frac{2}{\sqrt{\pi}}$$

with probability 1.

3. A Matter of Degree. In a graph, the degree of a vertex x, denoted $d(x)$, is the number of edges incident with x. The minimum and maximum degrees of a graph are denoted by δ and Δ respectively. What can we say about the minimum and maximum degree of a random interval graph?

The minimum degree question is answered in [**8**]. We simply repeat the result here:

THEOREM 3. *If δ_n denotes the minimum degree of a vertex in the interval graph generated by n random intervals, then for fixed $k > 0$,*

$$\lim_{n\to\infty} \Pr\left(\delta_n \leqslant k\sqrt{n}\right) = 1 - \exp\{-k^2/2\}.$$

It follows that the average minimum degree approaches $\sqrt{n\pi/2}$.

The maximum degree question is more interesting, partly on account of the role played by a point of degree $n-1$. Recall that the *diameter* of a graph is the maximum over all pairs (x, y) of vertices of the least number of edges in a path from x to y. It is easy to show that with probability approaching 1, the diameter of our random interval graph is 2 or 3; and, for an interval graph G, $\mathrm{diam}(G) \leqslant 2$ iff there is a vertex adjacent to all others. In fact, the following are equivalent for any collection of n intervals:

1. the interval graph G has a vertex of degree $n-1$;
2. G has diameter at most 2;
3. the interval order P has an isolated point;
4. there is an interval in the family which meets all the others; and
5. there is an interval in the family which meets both $[u, v]$ and $[p, q]$, where $[u,v]$ is the interval with leftmost right endpoint and $[p, q]$ the interval with rightmost left endpoint.

We are thus moved to ask: What is the limiting probability that in a family of random intervals, there is an interval which intersects all the others?

One normally expects such limits to be either 0 or 1, and in fact in many classes of structures (such as the random graphs of Erdös and Rényi) there is a "0-1 Law" —in that case proved by Fagin [**3**]—which guarantees that first-order statements, such as "there is a vertex which is adjacent to all others," must have trivial limiting probability. Exceptions to this sort of behavior are sometimes quite startling, as in the famous "probleme de rencontres" in which the probability that a random permutation has no fixed point approaches $1/e$.

Here we shall obtain an answer which not only ruins a possible 0-1 Law for random interval graphs, but adds insult to injury in a surprising way.

Let us consider first the Poisson model, except that this time for convenience, let the process be of density n on $[0, 1]^2$. Let S be the minimum, over all points (x, y) of the process, of $\max\{x, y\}$; thus $S = \max\{u, v\}$ where $[u, v]$ is the interval with leftmost right endpoint mentioned above. Similarly let R be the position of the rightmost left endpoint. Then the "big interval," i.e., the interval which intersects all others, will exist just when there is a point of the process in the union of the rectangles $[0, S] \times [R, 1]$ and $[R, 1] \times [0, X]$. This will occur with limiting probability

$$\lim_{n\to\infty}\left[\int_0^\infty\int_0^\infty e^{-2nxy}\frac{\partial^2}{\partial x\,\partial y}\left(1-e^{-nx^2}-e^{-ny^2}+e^{-n(x^2+y^2)}\right)dx\,dy\right]=\frac{2}{3}.$$

It was when the authors asked themselves how fast the probability converged to $2/3$ that the startling truth emerged:

THEOREM 4. *For all $n > 1$, the probability that in a collection of n random intervals there is one which intersects all the others is exactly* $2/3$.

Proof. One can show, using the uniform model on $[0, 1]^2$, that the above probability is

$$1-4n(n-1)\int_0^1\int_0^{1-y} xy\left(1-x^2-y^2-2xy\right)^{n-2}-1\,dx\,dy,$$

which a patient reader will reduce to the constant $2/3$. Fortunately, there is a relatively painless combinatorial proof, given below, which is both more intuitive and more powerful.

In this proof we employ the "discrete" model, in which integers between 1 and $2n$ are paired at random. Once the intervals have been selected, we label the endpoints $A(1), B(1), \ldots, A(n-2), B(n-2)$ recursively as follows:

Refer to the endpoints $\{1, \ldots, n\}$ as the *left side*, and $\{n+1, \ldots, 2n\}$ as the *right side*. Let $A(1) = n$ and let $B(1)$ be its mate. Suppose we have assigned through $A(j), B(j)$. We attach the labels $A(j+1)$ and $B(j+1)$ by the following rules:

- If $B(j)$ is on the left side: Let $A(j+1)$ be the leftmost point on the right side which has not yet been labeled. Let $B(j+1)$ be its mate.
- If $B(j)$ is on the right side: Let $A(j+1)$ be the rightmost point on the left side which has not yet been labeled. Let $B(j+1)$ be its mate.

If $A(j) < B(j)$ we say that this interval "went to the right"; otherwise, it "went to the left". Note that we are labeling endpoints of intervals from the center outwards, starting from the left when the last interval went to the right, and vice-versa. Endpoints marked $A(\cdot)$ are called *inner* points and those marked $B(\cdot)$ are called *outer*. (See FIGURE 2.)

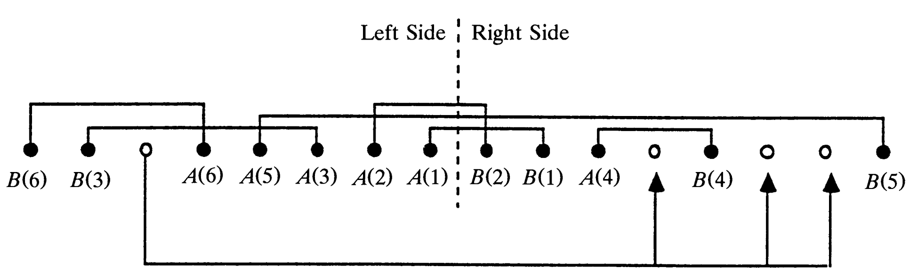


FIG. 2. Labeling intervals.

It is easy to prove by induction that immediately after the labels $A(j)$ and $B(j)$ have been assigned either:

1. an equal number of points have been assigned from the left and right sides (in case $A(j) < B(j)$), or
2. two more points have been labeled on the left than on the right (in case $A(j) > B(j)$).

Now when the labels $A(n-2)$ and $B(n-2)$ have been assigned, there remain four endpoints $a < b < c < d$ which are unlabeled. Note that the three ways of pairing them are equiprobable. Our claim is that if a is paired with either c or d, then that interval intersects all others, but if a is paired with b, then *no* interval intersects all others. This will prove the result.

We consider the only two possible cases:

1. a and b are on the left and c and d are on the right, and
2. only a is on the left.

In either case, all points labeled with $A(\cdot)$ are between a and c for otherwise one of a or c would have been labeled. It follows that either $[a, c]$ or $[a, d]$ meets all other intervals.

On the other hand, suppose $[a, b]$ and $[c, d]$ are intervals in the collection. Neither is a candidate as an interval which intersects all others since $[a, b] \cap [c, d] = \varnothing$. Suppose some interval $[e, f]$ (where $e < f$) intersects all others. Suppose e and f have received the labels $A(j)$ and $B(j)$.

In case (1), where a and b are on the left, we know that the endpoint $[e, f]$ labeled $A(j)$ is between b and c. Thus $[e, f]$ cannot intersect both $[a, b]$ and $[c, d]$.

Now consider case (2), where just a is on the left. Since $[e, f]$ meets $[c, d]$ we have $f > c$, hence f is an outer point ($f = B(j)$). Further, e is an inner point and therefore $[e, f]$ went to the right. However, the last labeled pair, $\{A(n-2), B(n-2)\}$ must have gone to the left since (in this case) we assigned more labels on the left than on the right. Thus, for some k, with $j < k \leqslant n-2$ we have that $[A(k), B(k)]$ went to the left, but $[A(k-1), B(k-1)]$ went right. Thus $A(k) < n$ and $A(k) < A(j)$ since $A(k)$ is a later-assigned, left-side inner point. It now follows that $[B(k), A(k)]$ is disjoint from $[A(j), B(j)] = [e, f]$, a contradiction.

With slightly more care one may use this construction to show that for any $k < n$, the probability that in a family of n random intervals there are at least k which intersect all others is

$$\frac{2^k}{\binom{2k+1}{k}}$$

independent, again, of n.

Acknowledgements. Edward Scheinerman is supported, in part, by ONR contract N00014-85-K-0622. Peter Winkler is supported, in part, by ONR contract N00014-85-K-0769.

REFERENCES

1. Béla Bollobás and Peter Winkler, On the longest chain among random points in Euclidean space, *Proceedings of the American Mathematical Society*, 103 (1988) 347–353.

2. P. Erdös and A. Rényi, On random graphs I, *Publ. Math. Debrecen*, 6 (1959) 290–297.
3. R. Fagin, Probabilities on finite models, *J. Symbolic Logic*, 41 (1976) 50–58.
4. Peter C. Fishburn, *Interval Orders and Interval Graphs*, Wiley, 1985.
5. Martin C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, 1980.
6. J. M. Hammersley, A few seedlings of research, *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, U. of California Press, 1972, pp. 345–394.
7. B. F. Logan and L. A. Shepp, A variational problem for random Young tableaux, *Adv. in Math.*, 26 (1977) 206–222.
8. Edward R. Scheinerman, Random interval graphs, *Combinatorica*, 8 (1988) 357–371.
9. C. Schensted, Longest increasing and decreasing subsequences, *Canad. J. Math.*, 13 (1961) 179–191.
10. S. M. Ulam, Monte Carlo calculations in problems of mathematical physics, *Modern Mathematics for the Engineer*, E. F. Beckenbach, ed., McGraw Hill, New York, 1961.
11. A. M. Veršik and S. V. Kerov, Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tableaux, *Dokl. Akad. Nauk SSSR*, 233 (1977) 1024–1028.

Playing the Numbers

You can play mathematical sequences as musical notes on your computer. If the sequence is periodic, the melody might be quite interesting, and it certainly puts a new dimension into the subject.

For example, fix a modulus m. The sequence of Fibonacci numbers

$$0, 1, 1, 2, 3, 5, \ldots \quad (\text{modulo } m)$$

is periodic. You can listen to the Fibonacci numbers (mod m) by running the following little Basic program (enter your favorite modulus m at the prompt).

```
10  CLS: INPUT "MODULUS"; M : A=0: B=1
20  C = A + B: C = C - M*INT(C / M)
30  SOUND 130*2 ^(C / 12), 2 : A = B: B = C: GOTO 20
```

The tune '$m = 51$,' for instance, is rather appealing. Try other periodic mathematical sequences, scoring several of them together, varying the frequencies and durations, etc. See problem E3410 on p. 916 of this issue for information about the periods.

—U. Phonious