involve
a journal of mathematics

# The sock matching problem

Sarah Gilliand, Charles Johnson, Sam Rush and Deborah Wood

msp

2014 vol. 7, no. 5

INVOLVE 7:5 (2014)
dx.doi.org/10.2140/involve.2014.7.691

# The sock matching problem

Sarah Gilliand, Charles Johnson, Sam Rush and Deborah Wood

(Communicated by Jim Haglund)

When matching socks after doing the laundry, how many unmatched socks can appear in the process of drawing one sock at a time from the basket? By connecting the problem of sock matching to the Catalan numbers, we give the probability that $k$ unmatched socks appear. We also show that, for each fixed $k$, this probability approaches 1 as the number of socks becomes large enough. The relation between the number of socks and the $k$ for which a given probability is first reached is also discussed, but a complete answer is open.

## 1. Introduction

In any load of clothes to be washed by a college student, there are inevitably a variety of socks tossed in with all the other garments. By the time the clothes come out of the dryer, the socks have been thoroughly mixed in, hiding underneath shirts or in pant legs. The game of matching then begins: does the sock you just picked randomly out of the pile match any of the others you've already removed from the pile? How big is your stack of unmatched socks going to get? This creates a scenario in which there can be $k$ unmatched socks out of $n$ pairs. We wish to determine the likelihood of obtaining a maximum of $k$ unmatched socks while folding a pile of laundry containing the $n$ pairs of socks. We assume that each pair of socks is complete and unique, and that socks are drawn randomly, one at a time.

## 2. Background

***Catalan numbers.*** The Catalan numbers are a sequence named after the Belgian mathematician Eugène Charles Catalan (1814–1894), who, in an 1838 paper, first defined them in their modern form [Larcombe 1999]. However, he was not the first to discover the numbers. In fact, according to J. J. Luo [1988], the Chinese mathematician Antu Ming (c. 1692–1763) discovered the numbers before anyone else [Koshy 2009; Larcombe 1999]. Leonard Euler (1707–1783) published a

*MSC2010:* primary 05A15, 05A16; secondary 03B48, 00A69.
*Keywords:* Catalan numbers, sock matching, Dyck paths.
This work supported, in part, by a QEP Mellon grant to the College of William & Mary.

recursive definition of the sequence in 1761 [Koshy 2009], almost eighty years *before* the man after whom the sequence was eventually named. He, like Catalan, discovered the sequence while investigating the problem of cutting polygons into triangles with diagonals that do not cross [Koshy 2009; Larcombe 1999]. Indeed, the Catalan numbers "have [a] delightful propensity for popping up unexpectedly, particularly in combinatorial problems" (Martin Gardner, as quoted in [Koshy 2009, p. vii]). Besides triangles within polygons, the Catalan numbers can be found in exponentiations, Pascal's triangle, binary trees, diagonals in frieze patterns, partitions, the ballot problem, folding paper, and even baseball [Conway and Guy 1996; Koshy 2009].

*Definition.* The Catalan numbers are defined by the sequence:

$$C_0 = 1, \quad C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}.$$

The generating function for the Catalan numbers is

$$\sum_{n=0}^{\infty} C_n x^n = \frac{2}{1+\sqrt{1-4x}},$$

from which we can determine that $C_n = \binom{2n}{n}/(n+1)$.

*Examples of problems in which Catalan numbers arise.* One example of the Catalan numbers is that $C_n$ is the number of paths in an $n \times n$ grid starting from the lower left corner and ending in the upper right corner using only moves up and to the right without moving across the diagonal (called *Dyck paths*). The recursive nature of this example arises from visits to locations along the diagonal.

Another example is the number of paths from $(0, 0)$ to $(2n, 0)$ on the Cartesian plane using only moves to the northeast and southeast that do not move below the $x$-axis. The recursive nature of this example arises from visits to the $x$-axis.

*Relation to our problem.* Take the situation that one has $n$ pairs of socks to match and none yet drawn and left unmatched. At this point, there are $C_n$ ways in which socks can be drawn one at a time and set aside as unmatched until they find a match (assuming that the order in which different pairs of socks are matched is not considered). Because of this connection to the Catalan numbers, we could use information already known about the integer series in our investigation of the sock matching problem.

## 3. Initial observations

Every time a sock is drawn from the laundry pile, there are two possible outcomes: it could either match a sock that has already been drawn, or it is temporarily a lone

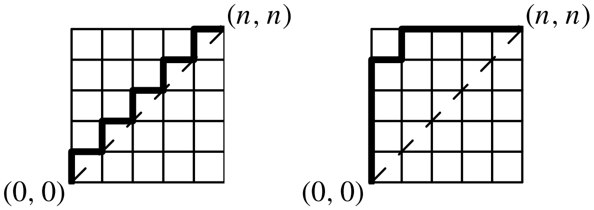

**Figure 1.** Left: $k = 1$; right: $k = n - 1$.

sock whose match has not yet been encountered. It is this aspect of the problem that allows us to use the grid visualization (Dyck paths) of the Catalan numbers as a model.

Every time a move is added to a path on the grid, there are also two options: it either moves one unit up or one unit to the right. Therefore, a move up can be taken to represent drawing a sock that has no match as of yet. And a move to the right represents drawing a match to some sock that has already been obtained from the laundry pile.

For example, Figure 1 (left) would result from an instance in which you continually pick a sock, and then pick its match. And Figure 1 (right) is the grid that shows the scenario in which you pick $n - 1$ socks without a single match, get a match, pick the last nonmatch, and then necessarily match the rest.

As can be seen from these grids, all of the paths that apply to this problem will begin at the origin (where no socks have yet been drawn). Because we are assuming that every sock has a match, all of the paths will also terminate at $(n, n)$, because for every move up on the grid, there is guaranteed to be a corresponding move to the right. Whenever the path hits the line $y = x$, we can see that all socks that have been drawn thus far have a match. None of the possible paths will ever cross below the line $y = x$, because this would indicate that there have been more matches than there have been previous unmatched socks, which is impossible. Also, a grid makes it easy to examine different values of $k$, because whenever the path hits or crosses the line $y = x + k$, we know that at least $k$ unmatched socks have been attained.

As opposed to using a grid, we could instead use a graph to model the paths created by drawing and matching the socks. Every time a sock without a match is pulled out, the path would move diagonally up and to the right one unit. Every time a match is obtained, it would move diagonally down and to the right. So, every return to the $x$-axis indicates an instance in which all socks that have been drawn have also been matched, and any time it hits or passes above the line $y = k$ indicates an instance in which at least $k$ unmatched socks have been reached. Any path would still begin at the origin, but must terminate at the point $(2n, 0)$.

***Expected value for maximum k.*** Drawing randomly from $n$ pairs of unmatched socks, how many socks may one expect to find drawn but unmatched at any point in the pairing process? Mathematically, this question asks for the average maximum value $k$ may reach, and in terms of Dyck paths of order $n$, this is the average distance from the diagonal to the path.

Here, the expected value equals the number of ways in which $n$ socks can be matched weighted by the maximum $k$ reached on that path divided by $C_n$. That is,

$$E(n) = \frac{\sum_{i=1}^{n} (\text{\# of ways to match socks such that at most } i \text{ are unmatched at any time} \times i)}{C_n}.$$

The numerator is the sum of heights of all Dyck paths of order $n$, sequence A136439 in the Online Encyclopedia of Integer Sequences (OEIS) [Finch 2008], and it must only be divided by $C_n$, the number of all those paths, in order to find the average. From OEIS, as well as Bruijn, Knuth and Rice [de Bruijn et al. 1972], we have as an equation $E(n)$ for the expected maximum number of socks to be left unmatched while matching $n$ pairs of socks:

$$E(n)=(n+1)\Biggl[\sum_{j=1}^{n+1}\Bigl(\sum_{i\mid j} i^0\Bigr)\frac{n!}{(n+a+j)!\,(j-a)!}-2\sum_{j=1}^{n}\Bigl(\sum_{i\mid j} i^0\Bigr)\frac{n!}{(n+a+j)!\,(j-a)!}\\ +\sum_{j=1}^{n-1}\Bigl(\sum_{i\mid j} i^0\Bigr)\frac{n!}{(n+a+j)!\,(j-a)!}\Biggr]-1.$$

***Recurrence formula.*** In this section, we focus upon the grid model for our problem. We define $B_{n,k}$ as the number of ways to get from (0, 0) to $(n, n)$ without crossing the diagonal but reaching the line $y = x + k$. This can be thought of as the total number of ways to get at least $k$ unmatched socks at least once during the matching process.

To do this, we must hit at least one point on the diagonal $y = x$ after (0, 0), since at the very least we must hit $(n, n)$. Let us consider the point $(i, i)$, which is the first point on the line $y = x$ that the path visits after (0, 0). For analysis, we consider three possibilities: the line hits $y = x + k$ before $(i, i)$, the line hits $y = x + k$ after $(i, i)$, and the line hits $y = x + k$ both before and after $(i, i)$ (which is counted twice so we want to subtract case 3 from the other two).

*Case 1*: The number of ways to hit $y = x + k$ between (0, 0) and $(i, i)$ is just $B_{i,k}$. The number of ways to get from $(i, i)$ to $(n, n)$ without hitting $y = x$ is the same as the number of ways to pass from $(i, i + 1)$ to $(n - 1, n)$ without crossing $y = x + 1$, which is $C_{n-i-1}$. Therefore the number of paths for this case is $B_{i,k}C_{n-i-1}$.

*Case 2*: The number of ways to get from (0, 0) to $(i, i)$ without crossing $y = x$ is $C_i$. Then, the number of ways to get from $(i, i)$ to $(n, n)$ hitting $y = x + k$ but not $y = x$ is the same as the number of ways to get from $(i, i + 1)$ to $(n - 1, n)$ hitting

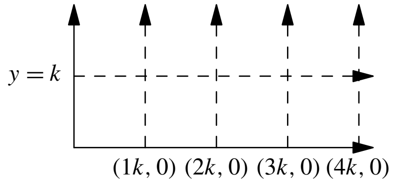

**Figure 2.** $\lim_{n\to\infty} P_{n,k}$.

$y=x+k$ but not crossing $y=x+1$, which is $B_{n-i-1,k-1}$. Therefore the number of paths in this case is $C_i B_{n-i-1,k-1}$.

*Case 3*: In this case, we want to count trajectories that hit $y=x+k$ before and after. This is a combination of our previous cases, $B_{i,k}B_{n-i-1,k-1}$. Now, we just need to add up the total paths for all of our total cases, so our final recurrence is:

$$B_{n,k}=\sum_{i=1}^{n}\bigl(B_{i,k}C_{n-i-1}+C_i B_{n-i-1,k-1}-B_{i,k}B_{n-i-1,k-1}\bigr). \tag{1}$$

We refer to the above as the *Sock Matching Theorem*.

## 4. Asymptotic behavior

The question that arises from the recurrence formula is whether or not the basic patterns we see for small values of $n$ and $k$ hold true for all values of $n$ and $k$.

***Large n.*** In particular, we first ask if the probability of reaching a given, fixed $k$ approaches 1 as $n$ approaches infinity. In this section we show that it does.

Let us define $P_{n,k}$ as the probability that we reach $k$ unmatched socks at least once in a draw of $n$ pairs. Notice that $P_{n,k}=B_{n,k}/C_n$. We must prove that $\lim_{n\to\infty} P_{n,k}=1$. To do this, we return to the graph model.

First, split up the graph into sections of length $k$ as shown in Figure 2. Call the probability that the path reaches $y=k$ in the first section $p_1$. Since moving up at every step reaches $y=k$ in $k$ steps, $p_1$ is positive. The probability that the path reaches $y=k$ in any subsequent section is dependent on where the path terminated in the prior section. However, the probability of reaching $y=k$ in any section is at least $p_1$ *no matter what happened in prior sections*. That is, $p_i \ge p_1$ for section $i$, or $1-p_i \le 1-p_1$. This allows us to say that the probability we never reach $y=k$, which is $1-P_{n,k}$, is at most $\prod_{i=1}^{2n/k}(1-p_1)$. Therefore,

$$\lim_{n\to\infty} 1-P_{n,k}=\lim_{n\to\infty}\prod_{i=1}^{2n/k}(1-p_1)=\lim_{n\to\infty}(1-p_1)^{2n/k}=0.$$

Therefore, $\lim_{n\to\infty} P_{n,k}=1-\lim_{n\to\infty}(1-P_{n,k})=1-0=1$.

| $k$ | $P_{n,k} \geq 0.99$ | $P_{n,k} \geq 0.999$ | $P_{n,k} \geq 0.9999$ |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 6 | 8 | 10 |
| 3 | 12 | 16 | 20 |
| 4 | 20 | 27 | 33 |
| 5 | 30 | 39 | 49 |
| 6 | 41 | 54 | 67 |
| 7 | 55 | 72 | 88 |
| 8 | 70 | 91 | >93 |
| 9 | 86 | >93 | >93 |

**Table 1.** First value of $n$ at which $P_{n,k}$ has reached a certain threshold, for various values of $k$.

*A quadratic relationship?* In the previous section, we considered the asymptotic behavior of the model as $n$ approaches infinity given a fixed $k$. For our next step, we instead fixed the probability, $P_{n,k}$ in order to discover the behavior of $k$ as $n$ again approaches infinity. We started by setting the probability at 0.99, and from the tables of data generated by a computer program we acquired the necessary data to speculate.

This investigation proved intriguing but unsatisfying. For $1 < k < 6$ when $P_{n,k} = 0.99$, the relationship between $k$ and the first $n$ for which the probability of reaching $k$ is greater than or equal to $P_{n,k}$ can be described by the quadratic equation $n = k^2 + k$. This, however, fails for all other values of $k$ and all other probabilities. When the constant probability is 0.999, $n$ increases more rapidly as $k$ increases, and for the constant probability 0.9999, the rate of increase for $n$ rises even more. Our data, moreover, end at $k = 8$ for 0.999 and at $k = 7$ for 0.9999. Although the patterns in the Table 1 suggest a quadratic relationship exists in this context, a specific equation is not sustained by high values of $k$ or the given probability.

## References

[de Bruijn et al. 1972] N. G. de Bruijn, D. E. Knuth, and S. O. Rice, "The average height of planted plane trees", pp. 15–22 in *Graph theory and computing*, edited by R. C. Read, Academic Press, New York, 1972. MR 58 #21737 Zbl 0247.05106

[Conway and Guy 1996] J. H. Conway and R. K. Guy, *The book of numbers*, Springer, New York, 1996. MR 98g:00004 Zbl 0866.00001

[Finch 2008] S. Finch, "Sum of heights of all 1-watermelons with wall of length $2n$", in *The online encyclopedia for integer sequences*, 2008.

[Koshy 2009] T. Koshy, *Catalan numbers with applications*, Oxford University Press, Oxford, 2009. MR 2010g:05008 Zbl 1159.05001

[Larcombe 1999] P. J. Larcombe, “The 18th century Chinese discovery of the Catalan numbers”, *Mathematical Spectrum* (1999), 5–7.

[Luo 1988] J. J. Luo, “Antu Ming, the first inventor of Catalan numbers in the world”, *Neimenggu Daxue Xuebao* **19** (1988), 239–245. In Chinese.

Received: 2013-08-21 Accepted: 2013-11-25

scgilliand@email.wm.edu
*Department of Biology, The College of William & Mary,*
*College Station Unit 3011, P.O. Box 8793,*
*Williamsburg, VA 23187, United States*

crjohn@wm.edu
*Department of Mathematics, The College of William & Mary,*
*P.O. Box 8795, Williamsburg, VA 23187, United States*

samuel.j.rush@gmail.com
*Department of Computer Science, California Institute of*
*Technology, 1200 East California Boulevard, MS 305-16,*
*Pasadena, CA 91125, United States*

dwood@email.wm.edu
*Department of Mathematics, The College of William & Mary,*
*College Station Unit 4085, P.O. Box 8793,*
*Williamsburg, VA 23187, United States*

# involve

msp.org/involve

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US $120/year for the electronic version, and $165/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve

2014 vol. 7 no. 5

1944-4176(2014)7:5;1-4