RESEARCH STATEMENT

Craig Lennon
United States Military Academy

1. Introduction

Probability theory has frequently been used to deepen our understanding of computer algorithms and combinatorial objects. In the analysis of deterministic computer algorithms, one may assume that the algorithm receives input that is, in some sense, random, and then consider the likely behavior of the algorithm. With combinatorial objects, one can sometimes turn problems involving enumeration into probabilistic questions, borrowing techniques from probability theory to solve these problems. For non-deterministic computer algorithms, such as randomized algorithms and genetic or evolutionary algorithms, randomness is inherent in the nature of the algorithm, and most analysis will be of a probabilistic nature.

The author's research has largely been in the veins described above. In joint work with Boris Pittel, the author has applied probability to the examination of the stable marriage problem. In his own research, the author conducted an analysis into the effectiveness of the Prüfer code as a tree representation in genetic algorithms. These efforts, and possible extensions of them, are described in the two sections below, with the first section addressing the stable marriage problem and the second section describing the Prüfer code research.

2. The Stable Marriage problem

Background

An instance of a size n stable marriage problem involves n men and n women, each ranking all members of the opposite sex in order of individual preference as a marriage partner. A complete matching (a set of n marriages) is called stable if no unmatched man and woman prefer each other to their actual partners in the matching. For a given set of marriage partner preferences, let $\mathcal{S}_n$ denote the total number of stable matchings. In 1962, Gale and Shapely [GS], who introduced the problem, showed that at least one stable matching always exists, i.e. $\mathcal{S}_n \geq 1$, by developing a "proposal" algorithm which always finds a stable matching.

In 1976, Knuth [K1] produced a problem instance with $2^{n/2}$ stable matchings, and many years later, Irving and Leather [IL] demonstrated that, for n a power of 2, there exist problem instances with at least 2^{n-1} stable matchings. They also found a sequence of problem instances for $n = 2^k$, $k \geq 1$, and an associated recurrence for the number of solutions that seemed to indicate that this number grew faster than 2^n. Knuth [Pi1] was able to show rigorously that, indeed, the number of solutions in Irving-Leather's problem instance is on the order of 2.28^n at least. As the number of stable matchings can vary widely, it is natural to ask what the likely number of stable matchings is.

In 1976, Knuth [K1] discovered a $2n$-dimensional integral formula for $\mathrm{E}[\mathcal{S}_n]$, the expected value of the random number $\mathcal{S}_n$ of stable matchings. He conjectured that this formula could be used to find an asymptotic estimate for $\mathrm{E}[\mathcal{S}_n]$. Confirming Knuth's conjecture, Pittel [Pi2] used the integral formula to show that

$$\mathrm{E}[\mathcal{S}_n] = (1 + o(1))e^{-1} n \ln n, \quad n \to \infty. \tag{1}$$

This formula suggests, but does not prove, that a likely value of $\mathcal{S}_n$ is of order $n \ln n$ as well.

Meanwhile, in 1971-1972, McVitie and Wilson [MW] found a sequential version of Gale-Shapley's proposal algorithm, and Wilson [Wi] studied its expected running time for problem instances chosen uniformly at random from among all $(n!)^{2n}$ instances. In [KMP], Knuth, Motwani, and Pittel found an extension of the McVitie-Wilson algorithm that delivered all the stable husbands for any given woman, and used it to show that with high probability, the number of those stable husbands is between $(1/2 - \epsilon) \ln n$ and $(1/2 + \epsilon) \ln n$. Thus a likely number of stable matchings must grow as $\ln n$ at least. Later, Pittel [Pi5] proved that with high probability $\mathcal{S}_n$ is much larger, namely $n^{1/2-o(1)}$, but a lower bound $cn \ln n$ appeared out of reach.

Our results

In [LPi], we prove that

$$\mathrm{E}\left[\mathcal{S}_n^2\right] = (1+o(1))\left(\frac{1}{e^2}+\frac{1}{2e^3}\right)n^2\ln^2 n. \tag{2}$$

The proof builds on techniques developed in [Pi2] and [Pi3]. Preference lists are determined by the values of matrices of uniform random variables. Then the condition of stability can be described as an intersection of events involving the values of these variables. This allows us to express the probability that two matchings are simultaneously stable as an integral, which can then be asymptotically estimated. Summing over all pairs of distinct matchings gives us the second factorial moment, which is asymptotically equal to (2).

Combined with Cantelli's inequality, relations (1) and (2) imply that

$$\mathrm{P}\left(\mathcal{S}_n \geq \epsilon \mathrm{E}[\mathcal{S}_n]\right) \geq \frac{(1-\epsilon)^2}{(1-\epsilon)^2+(2e)^{-1}} > .84$$

for ϵ sufficiently small. Hence the fraction of problem instances with roughly $cn\ln n$ solutions is 0.84, at least.

Future research

The fact that $\lim \mathrm{E}^2[\mathcal{S}_n]/\mathrm{E}[\mathcal{S}_n] < 1$ is a strong indication (though not a proof) that the distribution of $\mathcal{S}_n$ is not sharply concentrated around $\mathrm{E}[\mathcal{S}_n]$. Still, $\mathcal{S}_n$ may well be of order $n\ln n$ in probability; formally if $\epsilon_n \downarrow 0$, however slowly, then

$$\mathrm{P}\left(\mathcal{S}_n \geq \epsilon_n \mathrm{E}[\mathcal{S}_n]\right) \to 1, \quad n\to\infty. \tag{3}$$

Alternatively, in light of [Pi5], there may exist a function $\omega(n) >> n^{1/2}$, but $\omega(n) = o(n\ln n)$, such that

$$\lim_{n\to\infty} \mathrm{P}\left(\mathcal{S}_n \leq \omega(n)\right) > 0.$$

By our result this limit (if it exists) is 0.16 at most. The question of which (if either) of these results hold remains open.

We believe that there are two possible avenues for a better estimate of the likely number of stable marriages, but explaining them requires a brief discussion of r-regular graphs and the random k-SAT problem. A random r-regular graph is a graph chosen uniformly at random from the set of all graphs in which every vertex has degree r. We will denote such a randomly chosen graph (on n vertices) by $G_{r,n}$. We will let $H(G)$ be the number of hamilton cycles in the graph G. In [RW3] Robinson and Wormald sharply estimated the average and variance of $H(G_{3,n})$, and found that the variance was a constant multiple of the square of the first moment, as is true in our estimates. Thus they could not use Chebyshev's inequality to show that

$$\lim_{n\to\infty} \mathrm{P}\left(H(G_{3,n}) > 0\right) = 1.$$

Much later, in [RW1], they found a way to prove the above statement by conditioning on the number of small cycles in the $G_{3,n}$, and then using Chebyshev's inequality on the variable $\mathrm{E}[G_{3,n} \mid \#\text{ small cycles}]$. Finally, in [RW2] they extended this result, showing that with high probability all r-regular graphs ($r \geq 3$) have a hamilton cycle.

Robinson and Wormald were able to prove their result because much of the variance in the number of hamilton cycles (and of matchings) in r-regular graphs is between groups of graphs with different numbers of small cycles. Within each group, the variance is smaller. It may be that in the stable marriage problem we could condition on some property $\mathcal{P}$ of the preference lists, and that the variance of $\mathrm{E}[\mathcal{S}_n \mid \mathcal{P}]$ will be asymptotically negligible by comparison with $n^2\ln^2 n$. One candidate for this property would be the number of stable sub-matchings. Define a sub-matching of size k to be a matching of two size k subsets of men and women. A sub-matching of men $i_1,\ldots,i_k$ and women $j_1,\ldots,j_k$ is stable if it is not blocked by any pair (m_{i_t}, w_{j_s}) $(t,s\in[k])$. We could condition on s_k :=the maximum number of sub-matchings, and we might find that

$$\mathrm{VAR}[\,\mathrm{E}[\mathcal{S}_n \mid s_k]\,] \leq f(k)\mathrm{E}[\mathcal{S}_n]^2, \quad f(k)\downarrow 0,\ k\to\infty.$$

At the present, however, it is not clear how we might estimate the variance of $\mathrm{E}[\mathcal{S}_n \mid s_k]$.

Another approach is based on techniques of Achlioptas, Moore and Peres, see [AM] and [AP]. This approach would involve giving different weights to different matchings. With $\mathbf{I}_A$ denoting the indicator of the event A, we have

$$\mathcal{S}_n := \sum_M \mathbf{I}_{\{M \text{ is stable}\}},$$

where the above sum is over all distinct matchings. Perhaps we could assign to each matching a weight $w(M) \leq 1$. Then we could consider

$$\mathcal{W}_n := \sum_M w(M)\mathbf{I}_{\{M \text{ is stable}\}} \leq \mathcal{S}_n.$$

It may be possible to find a function ϵ_n that goes to zero slowly such that

$$\mathrm{E}\left[\mathcal{W}_n\right] \sim \epsilon_n n \ln n, \quad \mathrm{E}\left[\mathcal{W}_n^2\right] \sim \mathrm{E}\left[\mathcal{W}_n\right]^2.$$

Then with high probability $\mathcal{W}_n$ would be on the order of $\epsilon_n n \ln n$, and $\mathcal{S}_n \geq \mathcal{W}_n$ would be of at least that order (with high probability). The challenge is choosing an appropriate weighting scheme, and at the present it is not clear what that scheme might be. Thus a better estimate of the likely value of $\mathcal{S}_n$, possibly using the methods discussed above, remains an open problem.

3. Locality Results for the Prüfer code

Background

A genetic algorithm typically requires: 1) a mathematical structure which constitutes a possible solution to the problem we are attempting to solve (a genotype), 2) a "fitness function" which evaluates how well the genotype solves the problem, 3) a method for selecting an initial population of genotypes, and 4) a method of generating new genotypes from the current population.

A tree is an object frequently represented, with a tree on n labeled vertices having as its genotype a length $n-2$ string on the letters $[n] := \{1, 2, \ldots, n\}$. Such a representation is advantageous because it is then easy to choose members of an initial population uniformly at random from among all trees. One widely known method for encoding a tree is the Prüfer code. Let us describe the process of encoding a tree as a string of the type described above (we will henceforth refer to such strings as "P-strings").

If we are given a tree T, we encode T as a P-string as follows: at step i $(1 \leq i \leq n-2)$ of the encoding process, the lowest number leaf is removed, and it's neighbor is recorded as p_i, the ith element of the P-string

$$P = (p_1, \ldots, p_{n-2}), \quad p_i \in [n], \quad (1 \leq i \leq n-2).$$

Avoiding a discussion of fitness functions (which are problem specific), we turn to methods for generating new populations. A typical genetic algorithm will involve some method of "mating" two highly fit genotypes (called recombination), and a method for making small changes in the genotype (called mutation). Ideally, the mutation operation will help a fit genotype move toward a local optimum, while the recombination operation will ensure that all local optima are explored, resulting in the discovery of a globally optimal solution. We want mutation to be a small change in the genotype which results in a small change in the possible solution [Ro].

A genotype representation is said to have high locality if small changes in the representation correspond to small changes in the represented object. For example, we might hope that changing one entry of a P-string results in a change of one edge of the represented tree (at least most of the time). To be more precise, let us introduce notation for the number of edges in one tree which are not in the other:

$$\Delta = \Delta^{(n)} = \Delta^{(n)}(T, T^*) := n - 1 - |E(T) \cap E(T^*)|,$$

where $E(T)$ is the edge set of tree T.

As a mutation in a P-string is the change of exactly one element of the P-string, we denote the set of all ordered pairs of P-strings differing in exactly one coordinate (the mutation space) by $\mathcal{M}$, and by $\mathcal{M}_\mu$ we mean the subset of the mutation space in which the P-strings differ in the μ th coordinate:

$$\mathcal{M} = \bigcup_{1=\mu}^{n-2} \mathcal{M}_\mu, \quad \mathcal{M}_\mu := \left\{(P, P^*) \,:\, p_i = p_i^* \text{ for } i \neq \mu, \text{ and } p_\mu \neq p_\mu^*\right\},$$

where

$$P = (p_1, \ldots, p_{n-2}), \quad P^* = (p_1^*, \ldots, p_{n-2}^*).$$

We choose a pair $(P, P^*) \in \mathcal{M}$ uniformly at random, and the random variable Δ measures the distance between the trees corresponding to (P, P^*). In order to say that the Prüfer code has high locality, we might require that

$$\lim_{n\to\infty} \mathrm{P}\left(\Delta^{(n)} = 1\right) = 1,$$

so that for large enough trees, it is highly likely that a mutation in the P-string results in a change of exactly one edge of the represented tree. Unfortunately, the experience of programmers suggested that the Prüfer code did not have high locality [GJRR].

Using $\mathrm{P}(\{\text{event}\}|\circ)$ to denote conditional probability, we will represent the event $(P, P^*) \in \mathcal{M}_\mu$ by μ, as in

$$\mathrm{P}(\{\text{event}\} \,|\, \mu) := \mathrm{P}(\{\text{event}\} \,|\, (P, P^*) \in \mathcal{M}_\mu).$$

Computer assisted experiments conducted by Thompson (see [Th1] page 195-196) for trees with a vertex size as large as $n = 100$ led him to conjecture that:

$$\lim_{n\to\infty} \mathrm{P}\left(\Delta^{(n)} = 1\right) = \frac{1}{3}, \tag{4}$$

and that if $\mu/n \to \alpha$, then

$$\lim_{n\to\infty} \mathrm{P}\left(\Delta^{(n)} = 1 \,\middle|\, \mu\right) = (1-\alpha)^2. \tag{5}$$

In a recent paper [PaS1], Paulden and Smith used combinatorial and numerical methods to develop conjectures about the exact value of $\mathrm{P}(\Delta = \ell \,|\, \mu)$ for $\ell = 1, 2$, and about the generic form that $\mathrm{P}(\Delta = \ell \,|\, \mu)$ would take for $\ell > 2$. These conjectures imply (4)-(5), but so far they have not been proved.

Results

In [Le], the author shows via a probabilistic method that (4)-(5) are indeed correct – in particular that

$$\mathrm{P}\left(\Delta^{(n)} = 1 \,\middle|\, \mu\right) = (1-\mu/n)^2 + O\left(n^{-1/3}\ln^2 n\right), \tag{6}$$

and that

$$\mathrm{P}\left(\Delta^{(n)} = \ell \,\middle|\, \mu\right) = O\left(n^{-1/3}\ln^2 n\right), \qquad (\ell > 1). \tag{7}$$

Of course (6) implies (4), because $\int_0^1 (1-\alpha)^2\, d\alpha = 1/3$.

The proof makes use of a decoding algorithm that reads the P-string from back to front, constructing a nested sequence of subtrees of the tree represented by the P-string. Applying the principle of deferred decisions, we find an event $\mathcal{E}$ occurring at step μ which implies the event $\Delta = 1$. We then show the the event $\Delta = 1$ is the union of $\mathcal{E}$ and several unlikely events, and estimate the probability of $\mathcal{E}$ to obtain (6)-(7).

Future research

The results above suggest that the Prüfer code is indeed a poor tree representation for genetic algorithms. There are, however, several other representations whose locality is currently being studied, and which seem more promising [PaS2], [PaS3]. Looking beyond tree representations, genetic and evolutionary algorithms seem like a fruitful area for probabilistic research. In the case of string representations, we found it easy to choose an initial population uniformly at random. But for other genotypes, a simple algorithm may not exist for uniformly random selection. Further, a uniformly random selection may not always be the best choice, for such a choice may lead the algorithm toward a solution that is locally, but not globally, optimal. Of course this depends on the problem in question. Additionally, questions of locality will arise with most genotypes, and probability theory will likely play a role in answering them. Thus for the present, the field of genetic and evolutionary algorithms presents an excellent area for those interested in combinatorial probability.

References

[AM] D. Achlioptas, C. Moore, *The Asymptotic Order of the Random k-SAT Threshold*, 2002; available at http://www.santafe.edu/∼moore/pubs.html.

[AP] D. Achlioptas, Y. Peres, *The Threshold for Random k-SAT is* $2^k log2 - O(k)$, Journal of the American Mathematical Society **4** (2004) pp. 947–973.

[CKSS] M. Cho, Kim, S. Seo, and H Shin, "Colored Prüfer codes for k-edge colored trees," *The Electronic Journal of Combinatorics*, vol. 10, 2004.

[GS] D. Gale, L. S. Shapley, *College Admissions and the Stability of Marriage*, American Mathematical Monthly **69** (1962) pp. 9–15.

[GJRR] J. Gottlieb, B. Julstrom, G. Raidl, F. Rothlauf. "Prüfer numbers and genetic algorithms: A lesson how the low locality of an encoding can harm the performance of GAs." *Lecture Notes in Computer Science* vol. 1917, Proc. PPSN VI Paris, France, pp. 395-404, September 2000.

[GI] D. Gusfield, R. W. Irving, *The Stable Marriage Problem: Structure and Algorithms*, MIT Press, Cambridge, 1989.

[IL] R. Irving, P. Leather, *The Complexity of Counting Stable Marriages*, SIAM Journal of Computing **15** (1986) pp. 655–667.

[K1] D. E. Knuth, *Stable Marriage and its Relation to Other Combinatorial Problems: An Introduction to the Mathematical Analysis of Algorithms*, American Mathematical Society, Providence, RI, 1997.

[KMP] D. E. Knuth, R. Motwani, and B. Pittel, *Stable Husbands*, Random Structures and Algorithms **1** (1990) pp. 1–14.

[LPi] C. Lennon, B. Pittel, *On the Likely Number of Stable Marriages*, dissertation at The Ohio State University (2007).

[MW] D. G. McVitie, L. B. Wilson, *The Stable Marriage Problem*, Comm. ACM **14** (1971) pp. 486–490.

[Le] C. Lennon, *On the Locality of the Prüfer Code*, submitted to *The Electronic Journal of Combinatorics* Spring 2008, available at http://aps.arxiv.org/abs/0802.3514.

[PaS1] T. Paulden and D. K. Smith, "Developing new locality results for the Prüfer Code using a remarkable linear-time decoding algorithm," *The Electronic Journal of Combinatorics*, vol. 14, August 2007.

[PaS2] T. Paulden and D. K. Smith, "Some Novel Locality Results for the Blob Code Spanning Tree Representation," *Genetic and Evolutionary Computation Conference: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pp. 1320-1327, 2007.

[PaS3] T. Paulden and D. K. Smith, "From the Dandelion Code to the Rainbow Code: A class of bijective spanning tree representations with linear complexity and bounded locality," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 2, pp. 108-122, April 2006.

[Pi1] B. Pittel, *Personal communication*, March 2007.

[Pi2] B. Pittel, *The Average Number of Stable Matchings*, SIAM Journal of Discrete Mathematics **2** (1989) pp. 520–549.

[Pi3] B. Pittel, *The "Stable Roommates" Problem with Random Preferences*, Annals of Probability **21** (1993) pp.1441–1477.

[Pi4] B. Pittel, *On a Random Instance of the "Stable Roommates" Problem: Likely Behavior of the Proposal Algorithm*, Combinatorics, Probability and Computing **2** (1991) pp. 53–92.

[Pi5] B. Pittel, *On Likely Solutions of a Stable Marriage Problem*, Annals of Applied Probability **2** (1992) pp. 358–401.

[IP] B. Pittel, R. W. Irving, *An Upper Bound for the Solvability Probability of a Random Stable Roommates Instance*, Random Structures and Algorithms **5** (1994) pp. 465–486.

[RW1] R. W. Robinson, N. C. Wormald, *Almost All Cubic Graphs are Hamiltonian*, Random Structures and Algorithms **3** (1992) pp.117–125.

[RW2] R. W. Robinson, N. C. Wormald, *Almost All Regular Graphs are Hamiltonian*, Random Structures and Algorithms **5** (1994) pp.363–374.

[RW3] R. W. Robinson, N. C. Wormald, *Existence of Long Cycles in Random Cubic Graphs*, Enumeration and Design, Ed. Jackson, D. and Vanstone, S., (1984) pp. 363–374.

[Ro] F. Rothlauf, *Representations for Genetic and Evolutionary Algorithms*, Second edition. Heidelberg, Germany: Physica-Verlag, 2006.

[Th1] E. B. Thompson, "The application of evolutionary algorithms to spanning tree problems," Ph.D. dissertation, University of Exeter, U.K., 2003.

[TPaS] E. Thompson, T. Paulden, and D. K. Smith, "The Dandelion Code: A new coding of spanning trees for genetic algorithms," *IEEE Transactions on Evolutionary Computing*, vol. 1,no. 1, 1 pp. 91-100, February 2007.

[Wi] L. B. Wilson, *An Analysis of the Stable Marriage Assignment Problem*, BIT **12** (1972) pp. 569–575.