Sudoku Players' Forums

FAQ Search Memberlist Usergroups
Profile You have no new messages Log out [denis_berthier]

THE REAL DISTRIBUTION OF MINIMAL PUZZLES

Goto page Previous 1, 2, 3, 4, 5, 6, 7, 8, 9 Next

newtopic postreply **Sudoku Players' Forums Forum Index -> General/puzzle**

View previous topic :: View next topic

Author	Message
Red Ed	Posted: Wed Jul 15, 2009 5:17 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 590

I'm back from my trip and resolved to be polite this time.

> **denis_berthier wrote:**
>
> > **Red Ed wrote:**
> >
> > I said I'd given up trying to be helpful, so I'll limit myself to dropping just a small hint: *think what happens at level 77*.
> >
> > Have fun pondering that. I'm off to catch a plane. 😀
>
> At level 77, nothing happens, because there's no minimal puzzle above.

> **denis_berthier wrote:**
>
> This strategy led me to prove that all the valid puzzles with the same number of clues have the same probability of being reached by a top-down generator. This is more or less obvious, as the relative probabilities of 2 such puzzles depend only on the numbers of paths leading to each of them, which in turn don't depend on what is below any of the minimal puzzles at upper floors.

Are you referring to what you named the "classical top-down algorithm"? Then I think you're wrong for the following reason. Consider a solution grid that contains precisely one U4 and, within it, a 77-clue single-solution puzzle, P, whose holes are incident with 3 of the 4 U4 cells. What's the probability that the generator reaches P?

- In going from 81 to 80 clues, the generator could make any of 81 choices, 4 of which are okay for the puzzle in question. Probability: 4/81.
- In going from 80 to 79 clues, probability = 3/80 for similar reasons.
- In going from 79 to 78 clues, probability = 2/79 for similar reasons.
- In going from 78 to 77 clues, probability = .75*1/78 + .25*1/77. The 1/77 applies in the case that the first three clues deleted by the generator were the 3 U4 holes in P.

It should be obvious that the twist in the tail, in red above, doesn't apply to most other 77-clue puzzles in that solution grid. In short: P-like puzzles are slightly more like to be picked than the other 77-clue puzzles.

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 730
Location: Paris, France

Posted: Wed Jul 15, 2009 5:58 pm   Post subject: quote edit

eleven wrote:

> Denis, this should work now, i tried to keep dukuso's style ☺

Thanks, eleven. dukuso's style is Chinese for me (and probably most styles in C would be Chinese for me), but I trust you.

eleven wrote:

> Note that there are big speed differences between PC's. On an "Intel(R) Core(TM)2 Duo CPU E8400 @ 3.00GHz" running 2 instances of sudo_gen (with different seed) i can generate about 180 puzzles each per second.
> For the new version it took about 10-15 minutes to get a puzzle.

I was surprised by your 180 puzzles per second on a 3 GHz PC. On my Mac (Intel Core 2 Duo 2.66 GHz), it should have been ~ the same, but I got only 88. I tried to add a 64 bits option and I now get ~ 170.
Just to make sure I have the right compilation option, here is what I use: gcc suexg.c -m64

eleven wrote:

> i got 2 26's, a 27 and 2 28's

small sample, but interesting to see that the mean is much higher than for the original version.
I've got my first 6 with seed 0 and I get similar results:

Code:

```
#clues    SER
27          7.2
25          2.6
27          5.6
25          1.5
26          7.2
25          5.6
```

My estimation is about 10 puzzles per hour, 240 per day, 7000 per month.
You'll have to be patient before I have sufficiently many puzzles to apply the Pn+1/Pn formula with reasonable confidence.

Back to top profile pm www

coloin

Joined: 06 May 2005
Posts: 1044
Location: Devon UK

Posted: Wed Jul 15, 2009 6:24 pm Post subject: quote

denis_berthier wrote:

> [...nsmall sample, but interesting to see that the mean is much higher than for the original version....

Well im not so sure you can easily use this method to generate clue freq. distribution.

29 clues - more likely to have 1 sol.
though counteracted by
28 clues more likely to be minimal

maybe not insurmountable, maybe this is what **Ocean** did here

nice work **eleven** even so.

C

Back to top profile pm

Red Ed

Joined: 06 Jun 2005
Posts: 590

Posted: Wed Jul 15, 2009 9:37 pm Post subject: quote

A quick comment on this:

denis_berthier wrote:

> My estimation is about 10 puzzles per hour, 240 per day, 7000 per month. You'll have to be patient before I have sufficiently many puzzles to apply the Pn+1/Pn formula with reasonable confidence.

If one of your aims here is to estimate the number-of-clues distribution then you will need to expend a lot of effort on 28, 29 and maybe 30-clue minimals, since the 28s and 29s together contribute something like 15-20% of the total. I have the same sort of problem with my method (see link) and would welcome ideas for producing a smaller variance estimator for those cases.

Unfortunately, I think that your method will be less successful than mine. Yours is effectively equal to my method run with s=c, for each c in parallel. I described s as a "tuning parameter which takes some effort to get right" (for a given c). For my particular implementation, s=c was a bad choice. We probably need to do a side-by-side test of the methods to be sure one way or the other.

I do at least agree that your method is unbiased (for fixed c) this time, though 😀

Back to top profile pm

coloin

Joined: 06 May 2005

Posted: Wed Jul 15, 2009 11:06 pm Post subject: denis quote

just back from the pub..............

Posts: 1044
Location: Devon UK

1 pint plus 4 ? grouse.......cant wait to do a compare and contrast with the "Ardbeg"..........I think Ive actually been to Ardbeg...many years ago.

I dont think denis will take back any of his comments...........

The statistics of this is intriguing, perhaps the work done by you and Ocean needs expanding for the secondary school maths student !

The more I think about the unavoidable sets - and just how many there are ......I remember a long tme ago when I suddenly realised the extent of this and just how many there were - there has got to be a "bulge" at the 27 + clue level whiuch explains the distribution. Except it doesnt come out when you examine 40-clue subgrids........

The fact that a grid can be defined by a puzzle with less than 21 or so clues should be considered remarkable.

But its difficult to tell people [i know Im one].....

C

Back to top profile pm

ronk

Joined: 02 Nov 2005
Posts: 2396
Location: Southeastern USA

Posted: Thu Jul 16, 2009 1:49 am Post subject: quote

> **denis_berthier wrote:**
>
> > **ronk wrote:**
> >
> > > **denis_berthier wrote:**
> > >
> > > Coloin, beware: the result on the same probabilities is valid only when you start from complete grids. If you start from partial grids, some paths won't appear.
> >
> > If, for n >= 40, all n-clue subgrids are non-minimal, how can starting with a 40-clue subgrid lead to misleading results ?
>
> ... if you consider only the descendants of a 40-clue valid subgrid, for some of them paths that would have come from other parts of the complete grid by deleting clues in a different order will be missing. These puzzles will miss part of their heritage.

Descendants, paths, heritage? Gee, let me rephrase in simpler terms. I take your answer to mean ...

Some of the minimal puzzles that can be generated from a complete grid may be lost when starting from a sub-grid, even a large subgrid.

Is that correct?

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 730
Location: Paris, France

Posted: Thu Jul 16, 2009 6:43 am Post subject: quote edit

> **ronk wrote:**
>
> I take your answer to mean ...
> Some of the minimal puzzles that can be generated from a complete grid may be lost when starting from a sub-grid, even a large subgrid. Is that correct?

Not at all.
I mean that the number of ways of reaching valid puzzles P with n clues generated from a fixed subgrid (with >n clues) will not be the same for all these P.

I won't have much time today for more details.

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 730
Location: Paris, France

Posted: Thu Jul 16, 2009 6:47 am Post subject: quote edit

> **Red Ed wrote:**
>
> If one of your aims here is to estimate the number-of-clues distribution then you will need to expend a lot of effort on 28, 29 and maybe 30-clue minimals, since the 28s and 29s together contribute something like 15-20% of the total. I have the same sort of problem with my method (see link) and would welcome ideas for producing a smaller variance estimator for those cases.
> Unfortunately, I think that your method will be less successful than mine. Yours is effectively equal to my method run with s=c, for each c in parallel. I described s as a "tuning parameter which takes some effort to get right" (for a given c). For my particular implementation, s=c was a bad choice. We probably need to do a side-by-side test of the methods to be sure one way or the other.

I won't have time to study this today. One thing I suggest you do is a synthetic post on your method and your current results. It would make it easier to understand.

> **Red Ed wrote:**
>
> I do at least agree that your method is unbiased (for fixed c) this time, though 😀

Great.

Back to top profile pm www

eleven

Joined: 10 Feb 2008
Posts: 355

Posted: Thu Jul 16, 2009 9:34 am Post subject: quote

Thats the distribution of the first 100 puzzles with seed 0 and 1, multiplied with Denis' correction factor

Edit: i had an error when copying the factors (1 level wrong), but it did not change much to the result:

Code:

```
     found    corr.     %
23:      4       81     0.14
24:     16      786     1.40
25:     44     4927     8.79
26:     79    19053    34.00
27:     50    24564    43.83
28:      7     6632    11.83
29:      0        0     0
```

200 puzzles, av. 25.88, corrected av. 26.5547

Probably it will get higher, when also 29's and/or 30's are found.

Last edited by eleven on Thu Jul 16, 2009 11:45 am; edited 1 time in total

Back to top profile pm

coloin

Joined: 06 May 2005
Posts: 1044
Location: Devon UK

Posted: Thu Jul 16, 2009 9:54 am   Post subject: quote

Nice work again.........

Duh...... its just dawned on me !

top down mean ~ 24.3

mean for ALL the minimal puzzles in a specific 40 clue subgrid ~ 25.2

number of 40 clue subgrids in a specific grid ~ 81/40!*41! of which ~ 35% have 1 sol.

Now............

If you were to get *all* puzzles from *all* the the valid 40-subgrids [a rather big if - even for just one complete grid !], there would be more "double counting" of the smaller size puzzles. So there will be a higher mean for each individual grid - and hence all grids.

So the mean puzzle size will be considerably higher than 25.2, making **Red Ed**'s and more recently **eleven**'s estimation of ~ 26.5 - 27 perhaps completely explainable.

Denis - Is there a missing factor [in the top-down generation] which you have not added ? Perhaps related to sample size and differential rate of duplicates with each n[c].

eleven - Are you really sure you can use those corrective factors for this data ? Are they really correcting for

Code:

```
29 clues - more likely to have 1 sol.
though counteracted by
28 clues more likely to be minimal
```

......maybe it is - though im not sure !

C

Back to top profile pm

eleven

Joined: 10 Feb 2008
Posts: 355

Posted: Thu Jul 16, 2009 12:55 pm Post subject: quote

coloin wrote:

> **eleven** - Are you really sure you can use those corrective factors for this data ?

Not at all, dont take that serious. Denis never said, how he will calculate the correction factors for these data. I just did that to see, what we would get. Compared to Red Ed's current results we have about the same mean value, but a different distribution:

Code:

```
       corr.    Red Ed
22:    0        0.0034
23:    0.14     0.15
24:    1.40     2.35
25:    8.79    13.64
26:   34.00    31.88
27:   43.83    32.77
28:   11.83    15.65
29:    0         3.56
```

Edit:
Ah, he did say it before the idea with the modified generator:

denis_berthier wrote:

> Do you have a more precise estimation of the % of minimal puzzles (wrt to the valid ones) for each number of clues.
> If we had this, we could have a more precise idea of the top-down generators bias.

It will be hard to find the percentage for high clue puzzles.

Back to top profile pm

Red Ed

Joined: 06 Jun 2005
Posts: 590

Posted: Thu Jul 16, 2009 1:27 pm Post subject: quote

Let p(c) be the probability that a random c-clue puzzle is minimal. Your method is being used at present to estimate p(c), for all c, directly. That's okay, but it's tricky to analyse the variance of the estimator when done that way ... in other words, hard to know how much to trust the results.

But you can do better by using the *same* method to estimate n(c), the *number* of minimal c-clue puzzles. p(c) is then an obvious byproduct of all the n(c) values. Let a "trial" be: { pick a random solution grid; remove clues in turn until minimal or multi-solution puzzle }. Let A(t,c) be the number of c-clue minimals found in t trials. Then the random variable N(t,c) = A(t,c) * choose(81,c) / t is an unbiased estimator for n(c). Mathematically, then, this equals the s=c case of my method.

You're already recording A(t,c) for each c. I'm just asking you to record t as well, and to report N(t,c) alongside those percentages. It will be interesting to see if your N(t,c) are significantly different from mine: if so, then we might guess that your biased solution grid generator is not suitable for this experiment (it may or may not be suitable - I honestly don't know).

As for measuring the variance ... I'll come back to that, later, on the minimal puzzles thread where my method originated.

Back to top profile pm

eleven

Joined: 10 Feb 2008
Posts: 355

Posted: Thu Jul 16, 2009 3:58 pm Post subject: quote

Red Ed wrote:

> Your method ...

Well its not my method, i just wanted to know, how the results of the modified generator can be used.

Quote:

> Let a "trial" be: { pick a random solution grid; remove clues in turn until minimal or multi-solution puzzle }. Let A(t,c) be the number of c-clue minimals found in t trials. Then the random variable N(t,c) = A(t,c) * choose(81,c) / t is an unbiased estimator for n(c). ...I'm just asking you to record t as well, and to report N(t,c) alongside those percentages.

I see. I only reported the number of trials (new grids), when i tested it for the first numbers. From that i guess, that about 3 in a million tries were found. Inserting this number gives

Code:

```
23: 5.7e12
24: 5.6e13
25: 3.5e14
26: 1.3e15
27: 1.7e15
28: 4.7e14
```

i.e. only half of your 24's and 25's
But before i start a new run, i will change it to count the trials again. So thanks.

Back to top profile pm

Red Ed

Joined: 06 Jun 2005
Posts: 590

Posted: Thu Jul 16, 2009 4:34 pm Post subject: quote

Excellent, just the ticket. And yes, of course, I meant Denis' method, not yours.

Just one other thing: please print the number of trials in addition to the per-clue estimates as in your table above. Then we can generate confidence intervals.

Back to top profile pm

denis_berthier

Posted: Thu Jul 16, 2009 9:37 pm Post subject: quote edit

[Deleted]

Joined: 19 Jun 2007
Posts: 730
Location: Paris, France

Last edited by denis_berthier on Thu Jul 16, 2009 10:05 pm; edited 1 time in total

Back to top

profile pm www

Display posts from previous: All Posts Oldest First Go

newtopic postreply

Sudoku Players' Forums Forum Index -> General/puzzle

All times are GMT

Goto page Previous 1, 2, 3, 4, 5, 6, 7, 8, 9 Next

Page 8 of 9

Stop watching this topic

Jump to: General/puzzle Go

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** vote in polls in this forum