# THE REAL DISTRIBUTION OF MINIMAL PUZZLES

**Goto page** Previous 1, 2, 3 ... , 16, **17**, 18 Next

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

**View previous topic :: View next topic**

| Author | Message |
|---|---|
| **eleven** | Posted: Sun Sep 20, 2009 3:18 pm Post subject: quote |
| Joined: 10 Feb 2008<br>Posts: 474 | **denis_berthier wrote:**<br>> Do we conclude that no optimisation work is underway?<br>You can, as far as i am concerned.<br><br>To say it with your words: "What exactly do we expect of it?". You have found 16 random 30's in months using heavy CPU power. Thus now you can build a collection of i guess 526 random puzzles.<br><br>With a lot of effort this can be made 5 times faster. Then - provided the CPU'S - you can calculate about 10000 random puzzles in a year. Thats not worth for me to invest any time. |
| **Back to top** | profile pm |
| **denis_berthier** | Posted: Sun Sep 20, 2009 4:53 pm Post subject: quote edit |
| Joined: 19 Jun 2007<br>Posts: 814<br>Location: Paris, France | **eleven wrote:**<br>> To say it with your words: "What exactly do we expect of it?".<br>I can't speak for the others, but what I'm expecting is clear:<br>- the real (i.e. unbiased) distribution of clues of minimal puzzles;<br>- something I didn't expect when I started this thread: the real distribution of (SER or NRCZT) complexities for minimal puzzles with n clues (n fixed): I know now that it depends on how these puzzles were generated;<br>- the real distribution of (SER or NRCZT) complexities for minimal puzzles, by combining the previous two results.<br><br>Your estimations of the number of puzzles generated as of now are not correct. See my web pages for the current situation.<br><br>**eleven wrote:**<br>> With a lot of effort this can be made 5 times faster.<br>I think this is a very optimistic estimation.<br><br>My conclusion is that I let suexg-cb run a little longer. |

Back to top profile pm www

**Red Ed**

Posted: Sun Sep 20, 2009 8:59 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 715

> **denis_berthier wrote:**
>
> > **eleven wrote:**
> >
> > > With a lot of effort this can be made 5 times faster.
> >
> > I think this is a very optimistic estimation.

Interesting ... why

Back to top profile pm

**David P Bird**

Posted: Sun Sep 20, 2009 9:23 pm Post subject: quote

Joined: 17 Sep 2008
Posts: 139
Location: Middle England

**Denis**, It appears RE isn't responding to you because you called him by the wrong name and to me because I don't count.

In light of this perhaps you would be so kind as to put me right on this question I put to him:

> **David P Bird wrote:**
>
> Secondly in Scheme II it is possible to reach the stage where all cells are tagged 'Required' or 'Not Required' and we actually know the target minimum set we hope to find by chance. However, as we know the numbers of each, we can also accurately calculate what the chances are for actually hitting that target.
>
> Would there be a way to use that information to add a result to the accumulated list without invalidating the clue distribution that's produced? If that is possible then the time savings would be even greater!

Back to top profile pm

**Red Ed**

Posted: Sun Sep 20, 2009 9:33 pm Post subject: quote

DPB, lighten up. I've been busy.

Joined: 06 Jun 2005
Posts: 715

Back to top profile pm

**David P Bird**

Posted: Sun Sep 20, 2009 9:43 pm Post subject: quote

Joined: 17 Sep 2008
Posts: 139
Location: Middle England

Back to top profile pm

**JPF**

Posted: Sun Sep 20, 2009 11:48 pm Post subject: quote

Joined: 07 Dec 2005
Posts: 2855

> **denis_berthier wrote:**
>
> I can't speak for the others, but what I'm expecting is clear:
> - the real (i.e. unbiased) distribution of clues of minimal puzzles;

Location: Paris, France

> - something I didn't expect when I started this thread: the real distribution of (SER or NRCZT) complexities for minimal puzzles with n clues (n fixed): I know now that it depends on how these puzzles were generated;
> - the real distribution of (SER or NRCZT) complexities for minimal puzzles, by combining the previous two results.

Some comments :
- the real distribution of (SER or NRCZT) complexities for minimal puzzles is a real series of numbers , not to be mixed up with their estimations.
Therefore, the real distribution has nothing to do with the way the puzzles are generated.
- the (estimation of the) correlation between the number of clues and the complexity as defined here (SER,...) seems extremely weak. So, I don't see why combining the two first results could give a better estimation of the third one.

Btw, instead of using SER (and its arbitrary values) as a measure of complexity, one could calculate the ratio of "all singles" puzzles over the number of puzzles.

JPF

Back to top profile pm

**denis_berthier**

Posted: Mon Sep 21, 2009 4:13 am Post subject: quote edit

Joined: 19 Jun 2007
Posts: 814
Location: Paris, France

**JPF wrote:**

> **denis_berthier wrote:**
>
> > I can't speak for the others, but what I'm expecting is clear:
> > - the real (i.e. unbiased) distribution of clues of minimal puzzles;
> > - something I didn't expect when I started this thread: the real distribution of (SER or NRCZT) complexities for minimal puzzles with n clues (n fixed): I know now that it depends on how these puzzles were generated;
> > - the real distribution of (SER or NRCZT) complexities for minimal puzzles, by combining the previous two results.
>
> Some comments :
> - the real distribution of (SER or NRCZT) complexities for minimal puzzles is a real series of numbers , not to be mixed up with their estimations.
> Therefore, the real distribution has nothing to do with the way the puzzles are generated.

You're right, my formulation was loose. I was of course speaking of estimations. What I meant is that the real distribution of (SER or NRCZT) complexities for n-clue minimal puzzles, n fixed, which we don't know but can estimate from a random sample, depends in fact on how (by which kind of "random" generator) the sample was generated (sampling errors put aside). This should be obvious but we thought that the bias was mainly due to the number of clues; the examples on my website or the full bottom-up generator in the "rating" thread show that this expectation doesn't hold.

**JPF wrote:**

> the (estimation of the) correlation between the number of clues and the complexity as defined here (SER,...) seems extremely weak. So, I don't see why combining the two first results could give a better

> estimation of the third one.

Combining the distribution of complexities P(c, n) for each n and the distribution NP(n) of the n's is standard probabilistic reasoning. It gives an estimation of the real distribution of complexities CP(c):
CP(c) = sum P(c, n) * NP(n)
It doesn't depend on correlations.

> **JPF wrote:**
>
> instead of using SER (and its arbitrary values) as a measure of complexity, one could calculate the ratio of "all singles" puzzles over the number of puzzles.

My main measure of complexity is not the SER, but the purely logical NRCZT. The SER is only a statistical approximation, faster to compute and statistically well correlated. I use the SER for a first analysis and I can generally extend the results to the NRCZT.
Of course, once we have a controlled-bias sample, we can compute the statistics for any other measure of complexity we want - provided the program for computing this complexity measure is available. Is it the case for the one you're proposing?

Last edited by denis_berthier on Mon Sep 21, 2009 6:14 am; edited 2 times in total

Back to top profile pm www

**denis_berthier**

Joined: 19 Jun 2007
Posts: 814
Location: Paris, France

Posted: Mon Sep 21, 2009 4:17 am Post subject: quote edit

> **David P Bird wrote:**
>
> Secondly in Scheme II it is possible to reach the stage where all cells are tagged 'Required' or 'Not Required' and we actually know the target minimum set we hope to find by chance. However, as we know the numbers of each, we can also accurately calculate what the chances are for actually hitting that target.
> Would there be a way to use that information to add a result to the accumulated list without invalidating the clue distribution that's produced? If that is possible then the time savings would be even greater!

A more detailed analysis may be needed, but I fear that any use we make of information deduced from the current state will introduce some form of bias. The problem with all these potential improvements is that they are very difficult to analyse formally. The only worth of the controlled-bias algorithm is that it is provably controlled-bias. Loose this property and you can discard it.

Another problem I see with all the proposed "improvements" is that they maybe very hard to implement as changes to the existing version of suexg. Though I'm not very qualified in C, I can see that the current implementation isn't straightforward.

Last edited by denis_berthier on Mon Sep 21, 2009 6:20 am; edited 1 time in total

Back to top profile pm www

**m_b_metcalf**

Posted: Mon Sep 21, 2009 6:17 am Post subject: quote

**Red Ed wrote:**

> Let P80, P79, P78, ... be the subgrids of some solution grid on a path from 80 to 0 clues. Suppose that P40 is the first one with multiple solutions. Then, excusing 1-off errors in my description, suexg does this:
>
> count solutions for P80 : answer=1
> count solutions for P79 : answer=1
> count solutions for P78 : answer=1
> ...
> count solutions for P41 : answer=1
> count solutions for P40 : answer>1
>
> ... which is 41 expensive calls to the solver.
>
> A better strategy (one of many possible better strategies) ...

Joined: 15 May 2006
Posts: 2344
Location: Berlin

It is with great trepidation that I enter this discussion, but I really don't understand why the calls to the solver *have* to be expensive. Given a sub-grid of a solution grid, one implementation of my own code is:

- does any clue value occur 6 or more times? => not minimal
- are there fewer than eight clue values? => not minimal
- locate, in turn, all 4- and 6-unavoidable sets in the solution grid and check that each is covered by the clues in question, otherwise => not minimal
- solve for singles and some other simple logic (remember that about 80% of all 17-clue puzzles yield to this method). This will *either* find the unique solution *or* prune the search tree.
- **only if we get this far**: solve by backtracking, with the search order starting in the dense regions of the sub-grid and working through to the sparse regions, stopping at the second found solution, if more than one.

What I am trying to show here is that all calls to an *optimized* solver are not equal, so statements about speed improvements require some tests. But maybe *suexg* is not optimized.

Regards,

Mike Metcalf

Back to top profile pm

**denis_berthier**

Posted: Mon Sep 21, 2009 6:28 am Post subject: quote edit

**Mike**, Glad to see you join this discussion.

Joined: 19 Jun 2007
Posts: 814
Location: Paris, France

Needless to say I plainly agree with your comments.

Considering you have a different top-down generator, would it be easy to modify it into a controlled-bias one - along the same lines as suexgx.x was?

It would be interesting to compare the computation times, the mean number of

complete grids consumed for a minimal puzzle, and, last but not least, the distribution of clues. Unfortunately, all this supposes a long run time.

Back to top profile pm www

**Red Ed**

Posted: Mon Sep 21, 2009 6:30 am Post subject: quote

Joined: 06 Jun 2005
Posts: 715

Yep, *suexg*'s solver can be improved using some (not all) of the techniques you suggest. But my interest in this revolved around treating the solver as a black box -- for someone else to improve if they like -- and optimising the probing strategy instead.

Anyway, I tried it and as expected the path-probing part now runs 4.9 times quicker. Owing to the overhead of generating the solution grid in the first place remaining unchanged, that translates to a 3.5 times speed-up per puzzle generated. The only surprise, because I didn't think to check it before implementation, is that accounting for "more clues => quicker to solve" in the strategy makes no discernible difference to the run time.

Now I might see if there's a Huffman coding type description of the probing strategy: that would be a nice finale.

Back to top profile pm

**m_b_metcalf**

Posted: Mon Sep 21, 2009 6:40 am Post subject: quote

Joined: 15 May 2006
Posts: 2344
Location: Berlin

> **denis_berthier wrote:**
>
> Considering you have a different top-down generator, would it be easy to modify it into a controlled-bias one - along the same lines as suexgx.x was?

I'm now lost in the thicket of this discussion. Please remind of, or point me to, the algorithm for a 'controlled-bias' generator.

Regards,

Mike Metcalf

Back to top profile pm

**Red Ed**

Posted: Mon Sep 21, 2009 6:47 am Post subject: quote

Joined: 06 Jun 2005
Posts: 715

Denis will point you to something different, but for the purposes of understanding my posts on path probing (to speed up the controlled bias generator, a.k.a. "modified top-down generator") the best place to start is <here>.

Back to top profile pm

**denis_berthier**

Posted: Mon Sep 21, 2009 6:50 am Post subject: quote edit

Joined: 19 Jun 2007
Posts: 814
Location: Paris, France

> **m_b_metcalf wrote:**
>
> Please remind of, or point me to, the algorithm for a 'controlled-bias' generator.

Either my web page (the most up to date place)

http://carva.org/denis.berthier/HLS/Classification,
or here : http://www.sudoku.com/boards/viewtopic.php?t=14615&start=134

**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 ... , 16, **17**, 18 Next

**Page 17 of 18**

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