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 ... , 23, 24, 25 Next

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

View previous topic :: View next topic

Author	Message
gsf Joined: 22 Sep 2005 Posts: 3831 Location: NJ USA	Posted: Mon Sep 28, 2009 1:59 pm Post subject: quote

denis_berthier wrote:

> **Eleven**,
> If gsf's list was available, would there be a simple way of using it as a grid source for suexg-cb?
> You speak of obtaining 20000 puzzles without bias. Do you mean with suexg-cb, considering that each minimal puzzle consumes ~ 225000 complete grids?
>
> If we can do something like this, it may be worth contacting gsf. I've never been able to run his program (only for Windows, I think) on my Mac. I'm sure we can find a less mediaeval way of exchanging data than a memory stick.

I'm looking into free hosting for the data
but you will need my solver to read it
(it compresses ~ 1 grid / byte)
the solver runs on all different architectures
you just need to ask for yours and I'll post it
turns out almost all current interested users want the win32 executable
1 requested a linux executable

once storage is arranged I'll post the solver executables with the data

the data includes the #autopmorphisms
maybe that can be factored into the grid selection

reded is right about setting up a 5Gib lookup table
but you could do one pass over the data for one experiment
randomly select grids during the scan
then use those grids for the experiment
it takes ~4hr to scan all of the grids

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 850
Location: Paris, France

Posted: Mon Sep 28, 2009 4:25 pm Post subject: quote edit

gsf wrote:

> I'm looking into free hosting for the data

I can store 5.7 Gb but unfortunately not on a disk connected to the Web.

gsf wrote:

> but you will need my solver to read it
> (it compresses ~ 1 grid / byte)

Wow! < 1 bit per row! How is this possible?

gsf wrote:

> the solver runs on all different architectures
> you just need to ask for yours and I'll post it

I'm using a MacPro (Intel), OSX 10.6, preferably compiled in 64 bits mode.

gsf wrote:

> but you could do one pass over the data for one experiment
> randomly select grids during the scan
> then use those grids for the experiment
> it takes ~4hr to scan all of the grids

You mean your program can output a stream of puzzles and outputting the whole stream of 5.x billion would take only 4hr, right?

Considering that the controlled-bias generator consumes 225000 complete grids for a minimal puzzle, 5.x billion grids would produce ~ 22000 minimal puzzles - a sample large enough for doing interesting comparisons. This should take less than two weeks.

Now, besides the hosting problem, there remains the question of how to use your stream of complete grids, via Unix piping, as an input to suexg-cb. But that's probably obvious for **eleven**.

Back to top profile pm www

gsf

Joined: 22 Sep 2005
Posts: 3831
Location: NJ USA

Posted: Mon Sep 28, 2009 5:22 pm Post subject: quote

denis_berthier wrote:

> **gsf wrote:**
>
> > I'm looking into free hosting for the data
>
> I can store 5.7 Gb but unfortunately not on a disk connected to the Web.

you would have to download the data from that site

Quote:

> **gsf wrote:**
>
> > but you will need my solver to read it
> > (it compresses ~ 1 grid / byte)
>
> Wow! < 1 bit per row! How is this possible?

1 *byte* per row -- still that was a nice surprise
(btw we've written a bunch of compressors that compress all data to 1 bit, but there are bugs in the decompressors)
the grids are stored in minlex order,
so from grid to grid the changes happen in the rightmost cells
the data format encodes those changes
and that encoding is further compressed by bzip2 (the Burrows-Wheeler algorithm)
we have a proprietary compression that does a bit better than bzip2
but I didn't use that for public data

Quote:

> **gsf wrote:**
>
> > it takes ~4hr to scan all of the grids
>
> You mean your program can output a stream of puzzles and outputting the whole stream of 5.x billion would take only 4hr, right?

yes

Quote:

> Considering that the controlled-bias generator consumes 225000 complete grids for a minimal puzzle, 5.x billion grids would produce ~ 22000 minimal puzzles - a sample large enough for doing interesting comparisons. This should take less than two weeks.
>
> Now, besides the hosting problem, there remains the question of how to use your stream of complete grids, via Unix piping, as an input to suexg-cb. But that's probably obvious for **eleven**.

the stream is in minlex order, so that would have to be taken into account

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 850
Location: Paris, France

Posted: Mon Sep 28, 2009 5:55 pm Post subject: quote edit

gsf wrote:

> (it compresses ~ 1 grid / byte)

Wow! < 1 bit per row! How is this possible? [/quote]1 *byte* per row [/quote]
1 byte per grid, 9 rows per grid => < 1 bit per row (I meant row of a puzzle).
But this is of course an outsider's view.

gsf wrote:

> the stream is in minlex order, so that would have to be taken into account

I don't think it is a problem: it doesn't introduce any bias.
The only bias I can see is wrt to automorphic grids, but their proportion must be so small that it shouldn't count.

Back to top profile pm www

udosuk

Joined: 17 Jul 2005
Posts: 2827
Location: Sydney, Australia

Posted: Mon Sep 28, 2009 6:55 pm Post subject: quote

denis_berthier wrote:

> **gsf wrote:**
>
> > **denis_berthier wrote:**
> >
> > > **gsf wrote:**
> > >
> > > > (it compresses ~ 1 grid / byte)
> > >
> > > Wow! < 1 bit per row! How is this possible?
> >
> > 1 *byte* per row
>
> 1 byte per grid, 9 rows per grid => < 1 bit per row (I meant row of a puzzle). But this is of course an outsider's view.

Yep it's 1 *bit* per row **on average**. For a single grid or even a collection of 10 grids this is impossible, but if you stack millions of grids together with each adjacent grids just a few cells different then a good compression software can just store the minor changes between adjacent grids, which makes 1 *byte* per grid possible. The trade-off is it will take a *long* time to decompress, and you can't access/search the database in its compressed format.

Back to top profile pm

gsf

Joined: 22 Sep 2005
Posts: 3831
Location: NJ USA

Posted: Mon Sep 28, 2009 7:02 pm Post subject: quote

denis_berthier wrote:

> 1 byte per grid, 9 rows per grid => < 1 bit per row (I meant row of a puzzle). But this is of course an outsider's view.

sorry, I had my database lexicon out (row==record)

Back to top profile pm www

denis_berthier

Posted: Tue Sep 29, 2009 5:11 am Post subject: quote edit

Joined: 19 Jun 2007
Posts: 850
Location: Paris, France

gsf, udosuk, thanks for your explanations of the compression method. The result is very impressive. And udosuk's remarks about access to the data don't apply to sequential access, which is the case of interest for me.
A quick computation:
sudogen0_1M-solutions (1 million complete grids) size in plain text format = 82 Mb.
=> 5472730538 essentially different grids in plain text format ~ 449 GB
Compression rate ~ 78 (wow! again).
The 4 hrs you announce for reading the full file, the equivalent of 449 GB of grids in plain text format, sequentially with your program is a very attractive option. Compared to the time necessary to find minimal puzzles with suexg-cb, this is almost 0.

gsf, one more question. In this thread http://www.sudoku.com/boards/viewtopic.php?t=6679, you say:

> **gsf wrote:**
>
> there are 416 bands, and earlier bands compress better than later bands

How are the bands defined? Are they related to minlex order?

On second thoughts, the question of practical interest for my purposes is: if I use only the first part of the data (e.g. if, for some reason, the program stops before the estimated 2 weeks necessary to generate minimal puzzles with suexg-cb using the whole file), do I introduce a bias?

While you are trying to find hosting for the whole file, is there any possibility of sending me a small fraction of it (e.g. via my personal email, limited to 4 or 5 Mb) so that I can test the whole idea and see if anything is missing?

Back to top

profile pm www

gsf

Joined: 22 Sep 2005
Posts: 3831
Location: NJ USA

Posted: Tue Sep 29, 2009 5:20 am   Post subject:

quote

> **denis_berthier wrote:**
>
> **gsf**, one more question. In this thread http://www.sudoku.com/boards/viewtopic.php?t=6679, you say:
>
> > **gsf wrote:**
> >
> > there are 416 bands, and earlier bands compress better than later bands
>
> How are the bands defined? Are they related to minlex order?

a *band* in this context is the first three rows of a minlex grid
so the bands are in minlex order by definition

> **Quote:**
>
> On second thoughts, the question of practical interest for my purposes is: if I use only the first part of the data (e.g. if, for some reason, the program stops before the estimated 2 weeks necessary to generate

> minimal puzzles with suexg-cb using the whole file), do I introduce a bias?

there will be a bias because some bands will not be represented in
the minlex grid pool you draw from
but I don't know the consequences of such bias

Back to top profile pm www

gsf

Joined: 22 Sep 2005
Posts: 3831
Location: NJ USA

Posted: Tue Sep 29, 2009 5:44 am Post subject: quote

denis_berthier wrote:

> While you are trying to find hosting for the whole file, is there any possibility of sending me a small fraction of it (e.g. via my personal email, limited to 4 or 5 Mb) so that I can test the whole idea and see if anything is missing?

there will be 300 files in all
001.sudz ... 299.sudz for the first 299 bands
and 300-416.sudz for the remaining bands
I posted 300-416.sudz
it contains all 2,097,068 grids from bands 300..416
(~2.8 bytes per grid because there are a lot of bands for a small group of grids)
also posted in the same place is sudoku-darwin.i386, a mac osx i386 binary of my solver
this command should print 13 grids with their automorphism counts

Code:

```
sudoku -e '(%#An)>9' -f'%v # %#An' 300-416.sudz
```

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 850
Location: Paris, France

Posted: Tue Sep 29, 2009 5:58 am Post subject: quote edit

gsf wrote:

> I posted 300-416.sudz

Thanks, I could download it as a .sudz file

gsf wrote:

> also posted in the same place is sudoku-darwin.i386, a mac osx i386 binary of my solver

but I can't find this.

Back to top profile pm www

gsf

Joined: 22 Sep 2005
Posts: 3831

Posted: Tue Sep 29, 2009 6:03 am Post subject: quote

denis_berthier wrote:

> **gsf wrote:**

Location: NJ USA

> > also posted in the same place is sudoku-darwin.i386, a mac osx i386 binary of my solver
>
> but I can't find this.

here's the complete url sudoku-darwin.i386

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 850
Location: Paris, France

Posted: Tue Sep 29, 2009 6:43 am Post subject: quote edit

gsf wrote:

> this command should print 13 grids with their automorphism counts
>
> **Code:**
>
> ```
> sudoku -e '(%#An)>9' -f'%v # %#An' 300-416.sudz
> ```

I've now downloaded the 2 files and succeeded outputting the 13 grids.
But where did you specify in this command line that you wanted 13?
More importantly for me, how do you specify that you want all the grids?

Back to top profile pm www

gsf

Joined: 22 Sep 2005
Posts: 3831
Location: NJ USA

Posted: Tue Sep 29, 2009 6:51 am Post subject: quote

denis_berthier wrote:

> **gsf wrote:**
>
> > this command should print 13 grids with their automorphism counts
> >
> > **Code:**
> >
> > ```
> > sudoku -e '(%#An)>9' -f'%v # %#An' 300-416.sudz
> > ```
>
> I've now downloaded the 2 files and succeeded outputting the 13 grids.
> But where did you specify in this command line that you wanted 13?
> More importantly for me, how do you specify that you want all the grids?

this option lists a (terse) man page on stderr

Code:

```
--man
```

this

Code:

```
-e '(%#An)>9'
```

is a filter expression that lists all grids with #automorphisms > 9

to just list the grids use these options

Code:

```
-q- -f%v
```

where:
-q- : don't solve
-f%v : list grid as 81 chars, "." for empty cell

Back to top profile pm www

denis_berthier

Joined: 19 Jun 2007
Posts: 850
Location: Paris, France

Posted: Tue Sep 29, 2009 7:34 am Post subject: quote edit

In this sample of 2097068 grids, there are 5203 with more than 1 automorphism (the trivial one - I suppose you count it). This is ~ 0.25 %. Even if one of these grids has 108 automorphisms, most of them have less than 4.
If these percentages are not larger in the other bands (which remains to be checked), we can safely neglect automorphisms and take the output as a source of unbiased grids.

Back to top profile pm www

gsf

Joined: 22 Sep 2005
Posts: 3831
Location: NJ USA

Posted: Tue Sep 29, 2009 7:56 am Post subject: quote

denis_berthier wrote:

> In this sample of 2097068 grids, there are 5203 with more than 1 automorphism (the trivial one - I suppose you count it). This is ~ 0.25 %. Even if one of these grids has 108 automorphisms, most of them have less than 4.
> If these percentages are not larger in the other bands (which remains to be checked), we can safely neglect automorphisms and take the output as a source of unbiased grids.

only 560,151 grids have non-trivial automorphisms
here are the frequencies by #automorphisms

Code:

```
548449 2
7336 3
2826 4
1257 6
  29 8
  42 9
  92 12
  85 18
   2 27
  15 36
  11 54
   2 72
   3 108
   1 162
   1 648
```

Back to top profile pm www

Display posts from previous: All Posts ⬍ Oldest First ⬍ Go

Sudoku Players' All times are GMT

newtopic postreply **Forums Forum Index -> General/puzzle** **Goto page Previous 1, 2, 3 ... , 23, 24, 25 Next**

Page 24 of 25

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