DAG program (v0.12)

Causal diagrams:
A computer program to identify minimal sufficient adjustment sets

Sven Knüppel

March 31, 2010

Developed by

Sven Knüppel
Deutsches Institut für Ernährungsforschung Potsdam-Rehbrücke
Abteilung Epidemiologie
Arthur-Scheunert-Allee 114-116
14558 Nuthetal, Germany

e-mail: sven.knueppel@dife.de

Supervised by

Prof. Dr. med. Andreas Stang, MPH
Medizinische Fakultät Martin-Luther-Universität Halle-Wittenberg
Institut für Klinische Epidemiologie
Magdeburger Str. 8
06097 Halle (Saale), Germany

e-mail: andreas.stang@medizin.uni-halle.de

Short description of DAG program

Program version: v0.12 (March 31, 2010)
Algorithm: Backtracking
Language: C++

General notes

Confounding is an important source of bias in epidemiologic studies. With the introduction of causal diagrams (directed acyclic graphs, DAG) a new approach to conceptualize confounding and new rules to identify the minimal sufficient adjustment set have been established (Greenland, Pearl & Robins, Epidemiology, 1999, 10(1):37-48).

The DAG program is a MS DOS command-line analysis tool designed to select minimal sufficient adjustment sets within directed acyclic graphs. The main idea to find all back-door paths is the application of the backtracking algorithm.

Please note that we continuously improve the DAG program. If you have any problems, would like to report a bug or comment on the DAG program, please send an e-mail to sven.knueppel@dife.de.

Citing DAG program

If you use DAG program in any published work, please cite both the software (as an electronic resource/URL) and the research letter describing the methods.

```
Package:    DAG program (including version number)
Author:     Sven Knüppel
URL:        http://epi.dife.de/dag

Knüppel S, Stang A (2010)
DAG program: identifying minimal sufficient adjustment sets.
Epidemiology,21(1):159.
```

Running DAG

DAG program runs as a command-line tool and **cannot be open by clicking on the icon**. Before running the program you need to create an information file and a matrix file or a file of edge statements. If you wish to determine if a given set of covariates is a minimally sufficient adjustment set then create a set file. To run the program, open a MS DOS command prompt and start your analysis by typing a command like this:

```
DAG   -i example1.info    -m example1.txt    -all
```

Please make sure that the input files and the DAG program are located in the same folder.

Command Line Options

-i <filename>	Specify an information input file.
-m <filename>	Specify a matrix file.
-m2 <filename>	Specify a DAG with edge statements.
-set <filename>	Specify sets of covariates to find out if they are minimally sufficient adjustment sets.
-setall	All possible combinations of covariates, which are not affected by the exposure, are specified to find out if they are minimally sufficient, sufficient or insufficient. Option -all is used.
-all	All possible combinations of covariates are tested to be minimally sufficient (see remark 2, p. 11).

Input file formats

Information file

The information file is a white-space (space or tab) delimited file consisting of two columns:

Column 1: Name of variable
Alphanumeric value (without blanks or other white-spaces)
Column 2: Measurement status
Measurement status is set to 1 if the variable is measured otherwise it is set to 0.

The first line must contain the exposure of interest. The second line needs to contain the event of interest. Thereafter, you can add covariates in any order.

Matrix file (option –m)

The matrix file describes the graph by a matrix containing all arrow directions. Create a matrix consisting of all variables in the same order as in the information file and code their connection by 1 (arrow) and 0 (no arrow). For better understanding see the examples section.

File of edge statements (option –m2)

The second possibility to define a DAG is a list of white-space delimited edge statements. All statements are edge statements that name a tail node (parent) and a list of head nodes (childs). Edge starts from the parent and means

edge f**rom** *parent* **to** *child_1*, **to** *child_2* etc.

Only one edge statement is allowed per line. Every parent needs to be described in this file. For further information see section examples.

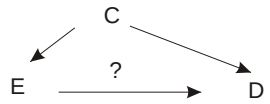
Set file (option –set)

You can use the set file (option -set) to specify one or more sets of covariates to find out which set is minimally sufficient, sufficient, or insufficient. If a set of covariates is sufficient but not minimally sufficient, it will be shown which minimally sufficient set(s) it contains. In case it is insufficient, the paths that aren't accounted for will be displayed. In every line of the set file you can define one set of covariates by typing the names defined in the information file.

Examples

Example 1: Classical confounder triangle

Suppose the following DAG is given.
E is the exposure of interest, D is the outcome, and C is a possible confounder.
All three variables are measured.



Make a matrix of all arrow directions. Set 1 if there is an arrow otherwise set 0.

		to E	to D	to C
	E	0	1	0
From	D	0	0	0
	C	1	1	0

Create two files: information and matrix file.

Information file: example1.info

```
E       1
D       1
C       1
```

Matrix file: example1.txt

Input from table above.

```
0       1       0
0       0       0
1       1       0
```

MS DOS Commando Prompt

```
dag -i example1.info -m example1.txt

*------------------------------------------------------------*
|                                                            |
|    DAG program                                             |
|                                                            |
|    v0.12                                                   |
|    March 31, 2010                                          |
|                                                            |
|    Sven Knueppel                                           |
|    http://epi.dife.de/dag/                                 |
|                                                            |
*------------------------------------------------------------*

Information file:   example1.info
Graph description:  example1.txt

3 measured and 0 unmeasured variables
from information file read.

No closed loop found.

No covariates affected by the exposure.

1 front-door path(s) from E to D found.

    1:      causal             E->D

1 back-door path(s) from E to D found.

    1:  non-causal  unblocked  E<-C->D

Summary:

    1 front-door path(s) found.
    1 back-door path(s) found.

    1 causal path(s) found.
    1 non-causal unblocked back-door  path(s) found.
    0 non-causal   blocked front-door path(s) found.
    0 non-causal   blocked back-door  path(s) found.

0 causal intermediate(s) found.

No colliders found.

1 minimally sufficient adjustment set(s) found:

    1: {C}

Minimally sufficient adjustment set(s)
with the lowest number of covariates:

    1: {C}
```

Example 1: Classical confounder triangle (continue)

Alternatively, you can specify the DAG with edge statements.

File of edge statements: example1_b.txt

```
E D
C E D
```

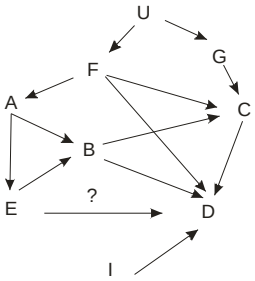
The edge statement E D makes an edge from E to D. The next edge statement C E D makes an edge from C to E and another from C to D. Because the syntax of edge statements is simple, graph definitions can be changed easily.
To run the DAG program use now the option –m2 instead of -m.

MS DOS Commando Prompt

```
dag -i example1.info -m2 example1_b.txt

*-----------------------------------------------------------*
|                                                           |
|    DAG program                                            |
|                                                           |
|    v0.12                                                  |
|    March 31, 2010                                         |
|                                                           |
|    Sven Knueppel                                          |
|    http://epi.dife.de/dag/                                |
|                                                           |
*-----------------------------------------------------------*

Information file:   example1.info
Graph description:  example1_b.txt

3 measured and 0 unmeasured variables
from information file read.

No closed loop found.

No covariates affected by the exposure.

1 front-door path(s) from E to D found.

    1:      causal             E->D

1 back-door path(s) from E to D found.

    1:  non-causal  unblocked  E<-C->D

Summary:

    1 front-door path(s) found.
    1 back-door path(s) found.

    1 causal path(s) found.
    1 non-causal unblocked back-door  path(s) found.
    0 non-causal   blocked front-door path(s) found.
    0 non-causal   blocked back-door  path(s) found.

0 causal intermediate(s) found.

No colliders found.

1 minimally sufficient adjustment set(s) found:

    1: {C}

Minimally sufficient adjustment set(s)
with the lowest number of covariates:

    1: {C}
```

Example 2: Complex DAG

Next DAG is more complex.



Information file: example2.info

E	1
D	1
A	1
B	1
C	1
F	1
G	1
I	1
U	0

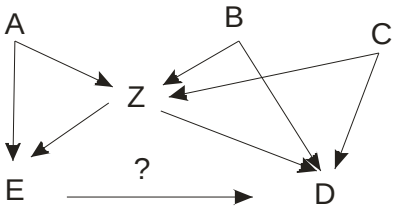
Matrix file: example2.txt

from \ to	E	D	A	B	C	F	G	I	U
E	0	1	0	1	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0
A	1	0	0	1	0	0	0	0	0
B	0	1	0	0	1	0	0	0	0
C	0	1	0	0	0	0	0	0	0
F	0	1	1	0	1	0	0	0	0
G	0	0	0	0	1	0	0	0	0
I	0	1	0	0	0	0	0	0	0
U	0	0	0	0	0	1	1	0	0

MS DOS Commando Prompt

```
dag -i example2.info -m example2.txt

*-----------------------------------------------------------*
|                                                           |
|    DAG program                                            |
|                                                           |
|    v0.12                                                  |
|    March 31, 2010                                         |
|                                                           |
|    Sven Knueppel                                          |
|    http://epi.dife.de/dag/                                |
|                                                           |
*-----------------------------------------------------------*

Information file:   example2.info
Graph description:  example2.txt

8 measured and 1 unmeasured variables
from information file read.

No closed loop found.

Covariates affected by the exposure:

    1: B
    2: C

8 front-door path(s) from E to D found.

    1:      causal             E->D
    2:      causal             E->B->D
    3:  non-causal    blocked  E->B<-A<-F->D          (blocked at B)
    4:  non-causal    blocked  E->B<-A<-F->C->D       (blocked at B)
    5:  non-causal    blocked  E->B<-A<-F<-U->G->C->D (blocked at B)
    6:      causal             E->B->C->D
    7:  non-causal    blocked  E->B->C<-F->D          (blocked at C)
    8:  non-causal    blocked  E->B->C<-G<-U->F->D    (blocked at C)

9 back-door path(s) from E to D found.

    1:  non-causal  unblocked  E<-A->B->D
    2:  non-causal  unblocked  E<-A->B->C->D
    3:  non-causal    blocked  E<-A->B->C<-F->D       (blocked at C)
    4:  non-causal    blocked  E<-A->B->C<-G<-U->F->D (blocked at C)
    5:  non-causal  unblocked  E<-A<-F->D
    6:  non-causal  unblocked  E<-A<-F->C->D
    7:  non-causal    blocked  E<-A<-F->C<-B->D       (blocked at C)
    8:  non-causal  unblocked  E<-A<-F<-U->G->C->D
    9:  non-causal    blocked  E<-A<-F<-U->G->C<-B->D (blocked at C)

Summary:

    8 front-door path(s) found.
    9 back-door path(s) found.

    3 causal path(s) found.
    5 non-causal unblocked back-door  path(s) found.
    5 non-causal   blocked front-door path(s) found.
    4 non-causal   blocked back-door  path(s) found.

2 causal intermediate(s) found: B, C

Collider(s) found:

    1: B   >> descendant(s): C
    2: C   >> no descendants

1 minimally sufficient adjustment set(s) found:

    1: {A}

Minimally sufficient adjustment set(s)
with the lowest number of covariates:

    1: {A}
```

Example 3: DAG including collider adjustment



Information file: example3.info

E	1
D	1
A	1
B	1
C	1
Z	1

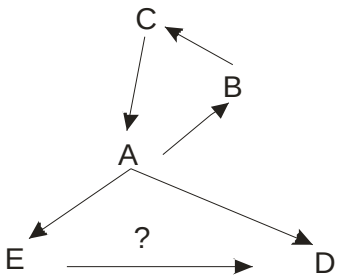
Matrix file: example3.txt

	to					
from	E	D	A	B	C	Z
E	0	1	0	0	0	0
D	0	0	0	0	0	0
A	1	0	0	0	0	1
B	0	1	0	0	0	1
C	0	1	0	0	0	1
Z	1	1	0	0	0	0

MS DOS Commando Prompt

```
dag -i example3.info -m example3.txt

*-----------------------------------------------------------*
|                                                           |
|    DAG program                                            |
|                                                           |
|    v0.12                                                  |
|    March 31, 2010                                         |
|                                                           |
|    Sven Knueppel                                          |
|    http://epi.dife.de/dag/                                |
|                                                           |
*-----------------------------------------------------------*

Information file:   example3.info
Graph description:  example3.txt

6 measured and 0 unmeasured variables
from information file read.

No closed loop found.

No covariates affected by the exposure.

1 front-door path(s) from E to D found.

    1:      causal             E->D

6 back-door path(s) from E to D found.

    1:  non-causal  unblocked  E<-A->Z->D
    2:  non-causal    blocked  E<-A->Z<-B->D (blocked at Z)
    3:  non-causal    blocked  E<-A->Z<-C->D (blocked at Z)
    4:  non-causal  unblocked  E<-Z->D
    5:  non-causal  unblocked  E<-Z<-B->D
    6:  non-causal  unblocked  E<-Z<-C->D

Summary:

    1 front-door path(s) found.
    6 back-door path(s) found.

    1 causal path(s) found.
    4 non-causal unblocked back-door  path(s) found.
    0 non-causal   blocked front-door path(s) found.
    2 non-causal   blocked back-door  path(s) found.

0 causal intermediate(s) found.

Collider(s) found:

    1: Z   >> no descendants

2 minimally sufficient adjustment set(s) found:

    1: {A,Z}
    2: {B,C,Z}

Minimally sufficient adjustment set(s)
with the lowest number of covariates:

    1: {A,Z}
```

Example 4: Causal graph with closed loop



MS DOS Commando Prompt

```
dag -i example4.info -m example4.txt

*-------------------------------------------------------------*
|                                                             |
|    DAG program                                              |
|                                                             |
|    v0.12                                                    |
|    March 31, 2010                                           |
|                                                             |
|    Sven Knueppel                                            |
|    http://epi.dife.de/dag/                                  |
|                                                             |
*-------------------------------------------------------------*

Information file:   example4.info
Graph description:  example4.txt

5 measured and 0 unmeasured variables
from information file read.


Closed loop found:
>> A->B->C->A
```

Information file: example4.info

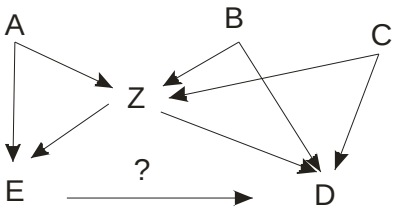
E	1
D	1
A	1
B	1
C	1

Matrix file: example4.txt

	to				
from	E	D	A	B	C
E	0	1	0	0	0
D	0	0	0	0	0
A	1	1	0	1	0
B	0	0	0	0	1
C	0	0	1	0	0

Example 5: Specification of a set of covariates

Suppose the DAG from example 3 is given.



We would like to test three different sets of covariates if they are sufficient.

Set 1 = {A}
Set 2 = {A,Z}
Set 3 = {A,B,C,Z}

Additional to the information file (example3.info) and matrix file (example3.txt) we need a file which specified the test sets.

Set file: example3.set

```
A
A Z
A B C Z
```

MS DOS Commando Prompt

```
dag -i example3.info -m example3.txt –set example3.set

*-----------------------------------------------------------*
|                                                           |
|    DAG program                                            |
|                                                           |
|    v0.12                                                  |
|    March 31, 2010                                         |
|                                                           |
|    Sven Knueppel                                          |
|    http://epi.dife.de/dag/                                |
|                                                           |
*-----------------------------------------------------------*

Information file:   example3.info
Graph description:  example3.txt
Set file:           example3.set

6 measured and 0 unmeasured variables
from information file read.

No closed loop found.

No covariates affected by the exposure.

1 front-door path(s) from E to D found.

    1:      causal             E->D

6 back-door path(s) from E to D found.

    1:  non-causal  unblocked  E<-A->Z->D
    2:  non-causal    blocked  E<-A->Z<-B->D (blocked at Z)
    3:  non-causal    blocked  E<-A->Z<-C->D (blocked at Z)
    4:  non-causal  unblocked  E<-Z->D
    5:  non-causal  unblocked  E<-Z<-B->D
    6:  non-causal  unblocked  E<-Z<-C->D

Summary:

    1 front-door path(s) found.
    6 back-door path(s) found.

    1 causal path(s) found.
    4 non-causal unblocked back-door  path(s) found.
    0 non-causal   blocked front-door path(s) found.
    2 non-causal   blocked back-door  path(s) found.

0 causal intermediate(s) found.

Collider(s) found:

    1: Z   >> no descendants

2 minimally sufficient adjustment set(s) found:

    1: {A,Z}
    2: {B,C,Z}

Minimally sufficient adjustment set(s)
with the lowest number of covariates:

    1: {A,Z}

------------------------------------------------------------
Option: -set

3 specified set(s) of covariates found.

    1: {A} is insufficient.


       Front-door paths:

             1 front-door (causal) paths remain open:
               1
             0 front-door paths are blocked by conditioning on specified set.

             0 front-door paths become open by conditioning on specified set.

             0 front-door paths remain closed.


-continue on next page-
```

MS DOS Commando Prompt: Example 5 -continued-

```
Back-door paths:

      3 back-door paths remain open:
        4 5 6
      1 back-door paths are blocked by conditioning on specified set:
        1
      0 back-door paths become open by conditioning on specified set.

      2 back-door paths remain closed.

1 minimally sufficient set(s) conditional on specified set:
     {A, Z}


2: {A, Z} is minimally sufficient adjustment set.


3: {A, B, C, Z} is sufficient and
   contains minimally sufficient adjustment set(s):
   {A,Z}
   {B,C,Z}
```

Remark

Remark 1: How many paths can be covered by the DAG program?

The backtracking algorithm is a fast algorithm to identify paths and can handle a high number of paths. The operating time depends on the number of existing paths. So far the upper limit for the number of paths, which can be handled, is not tested.

Remark 2: How many sets have to be tested to find the minimally sufficient adjustment sets?

With n variables you get 2^n possible combinations of the variables.

$n = 3 \quad \Rightarrow \quad 2^3 = 8$
$n = 10 \quad \Rightarrow \quad 2^{10} = 1{,}024$
$n = 20 \quad \Rightarrow \quad 2^{20} = 1{,}048{,}576$
$n = 100 \quad \Rightarrow \quad 2^{100} = 1.3 \cdot 10^{30}$

If the number of variables is high the DAG program cannot compute all combinations in a limited amount of time.

The DAG program applies the following steps.

1) Empty set is checked. The existence of one or more confounding paths makes that set insufficient.
2) One-covariate sets are checked that do not contain colliders.
3) Two-covariate sets are checked and sufficient sets which contains proper subsets are removed.
4) This process is repeated for candidate covariate sets containing 3, 4, 5 or more covariates and stopped, if in one step no minimally sufficient sets are found.

In this way the DAG program identifies at the minimum the minimally sufficient adjustment sets with the lowest number of elements!

If all combinations should be tested, you can specify option -all in the command line.

Changelog

This section contains a version history recording changes and additions to DAG program.

v0.12 (31-Mar-2010)

- Added –all option
- Added –setall option
- Added summary of count of paths
- Extension of the output of option -set
- Empty set {} could be identified as minimally sufficient adjustment set
- Consideration of descendents of colliders
- Some changes of the output

V0.11 (21-Oct-2009)

- Added –m2 option
- Added –set option
- All minimally sufficient adjustment sets are suggested instead of sufficient sets
- Fixed problem with colliders
- Fixed problem with covariates affected by the exposure

v0.10 (15-Jun-2009)

- First released version