Coedit: a tool for minimal cograph edge modification

Christophe Crespelle

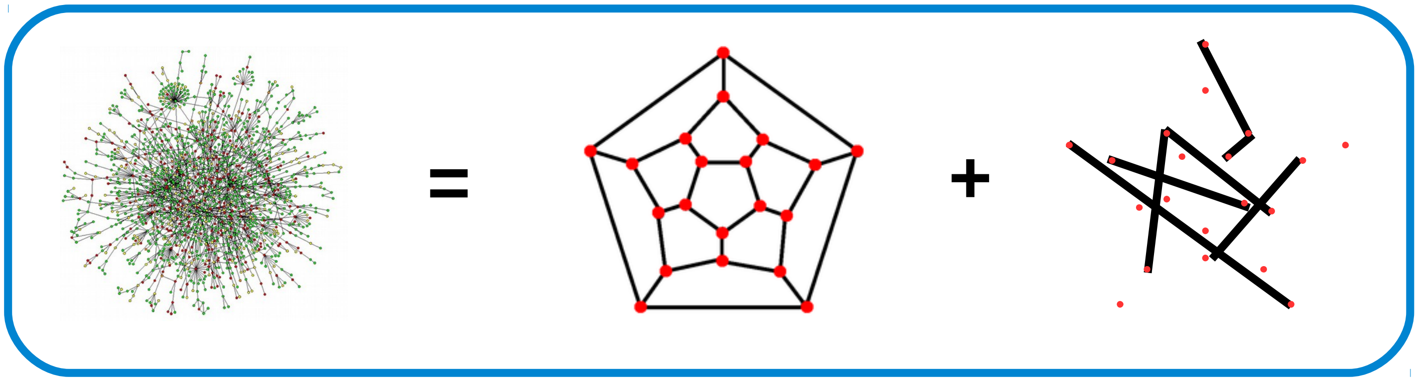
University of Bergen

with Daniel Lokshtanov, Thi Ha Duong Phan and Eric Thierry

Goal of PROXNET project

Representing real-world complex networks
as ***almost*** structured graphs

Complex network = structured graph + noise

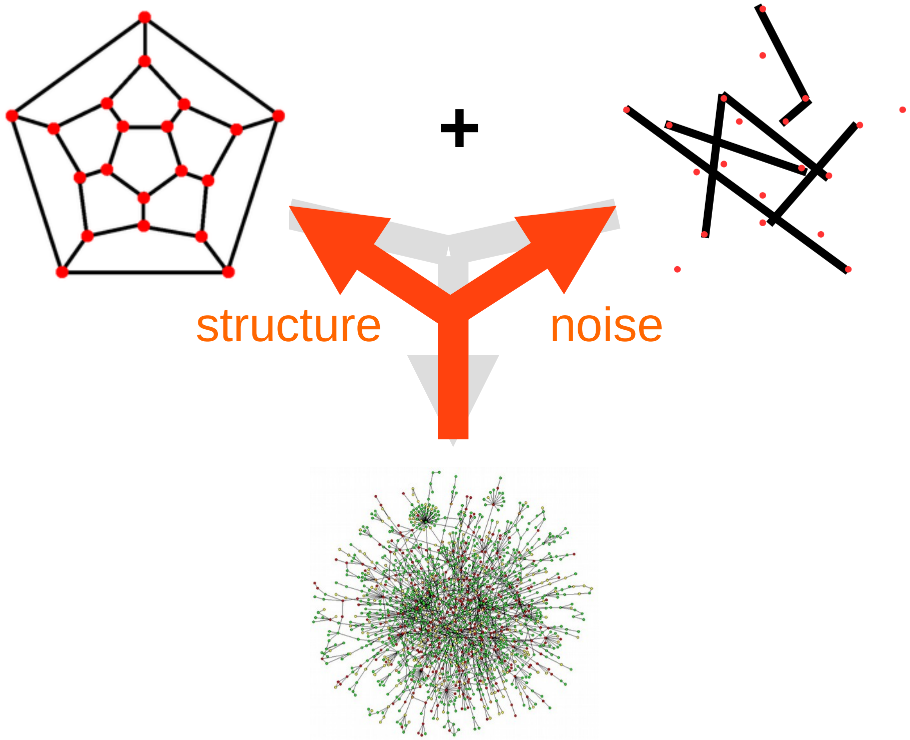


- Modelling
- Efficient encoding : space + query time
- Understand their structure (global organisation, specific roles)
- Algorithmic theory of ***almost*** structured graphs

→ Take advantage of the proximity with a strongly structured graph

Goal of PROXNET project

Representing real-world complex networks
as ***almost*** structured graphs



- Edge modification problems (editing, completion, deletion)
- Polynomial-time algorithms: set of modifications minimal for inclusion

Coedit

INPUT: an arbitrary graph

Computes either:

- a minimal cograph completion
- a minimal cograph deletion
- a minimal cograph editing

OUTPUT: the cotree of the cograph obtained

	Input format:
# of vertices	n
degrees	u d°(u) v d°(v) ⋮
edges	u1 v1 u2 v2 ⋮

	Output format:
# of nodes	n
Label of the root	l (=0 or 1)
# of children	u #child(u) v #child(v) ⋮
Edges of the tree	parent(u) u parent(v) v ⋮

- Written in C
- Sources available at https://www.ii.uib.no/~christophec/coedit/
- Under GNU GPL licence (can do whatever you want with it)

Algorithms

<u>For completion</u>

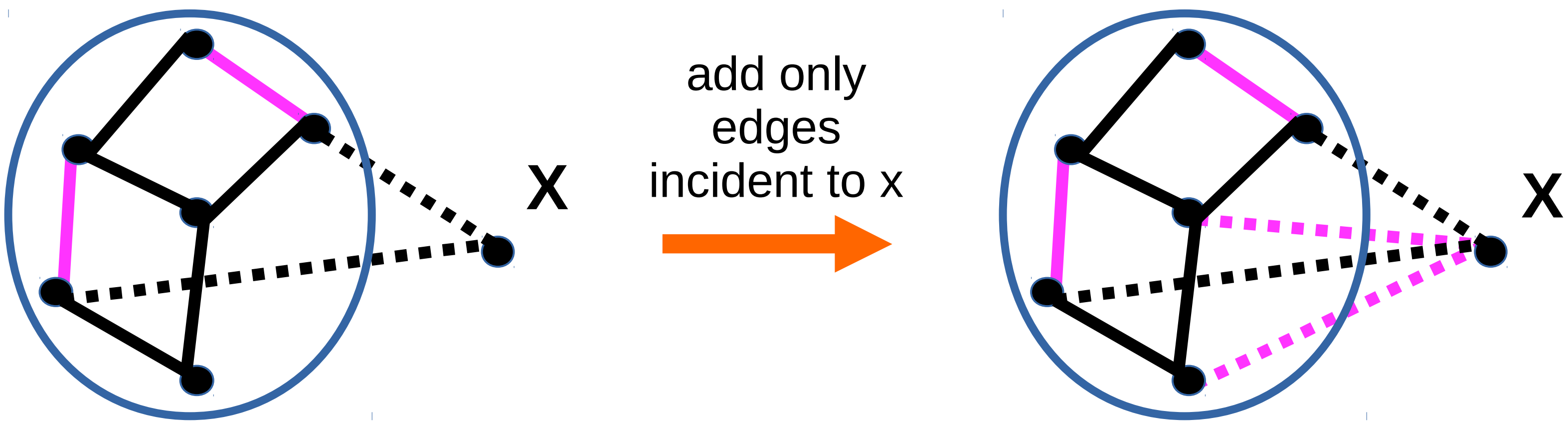
- An O(n+m') algorithm with ***minimum*** at each incremental step
  ➡ improve heuristics

- An O(n+m $\log^2$n) algorithm
  ➡ almost linear in the size of the ***input***

<u>For editing</u>

- An O(n+m) algorithm with ***minimum*** at each incremental step

The vertex incremental approach : vertices are processed one by one

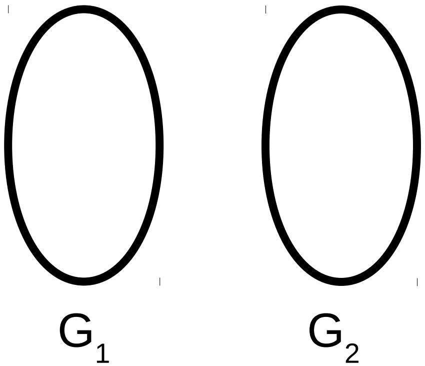


Cographs and incremental app.

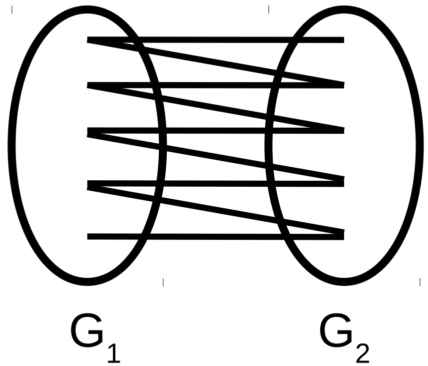
Obtained from single vertices by using 2 operations:

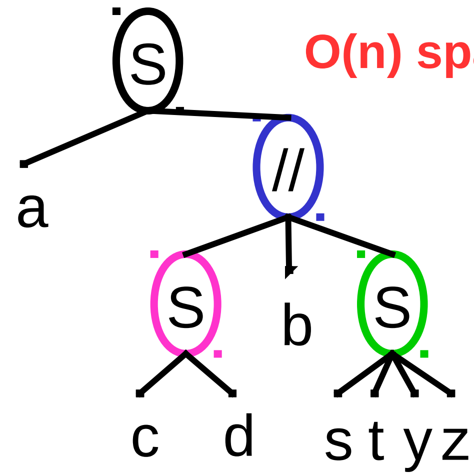
cotree

disjoint union
(//)



complete union
(S)

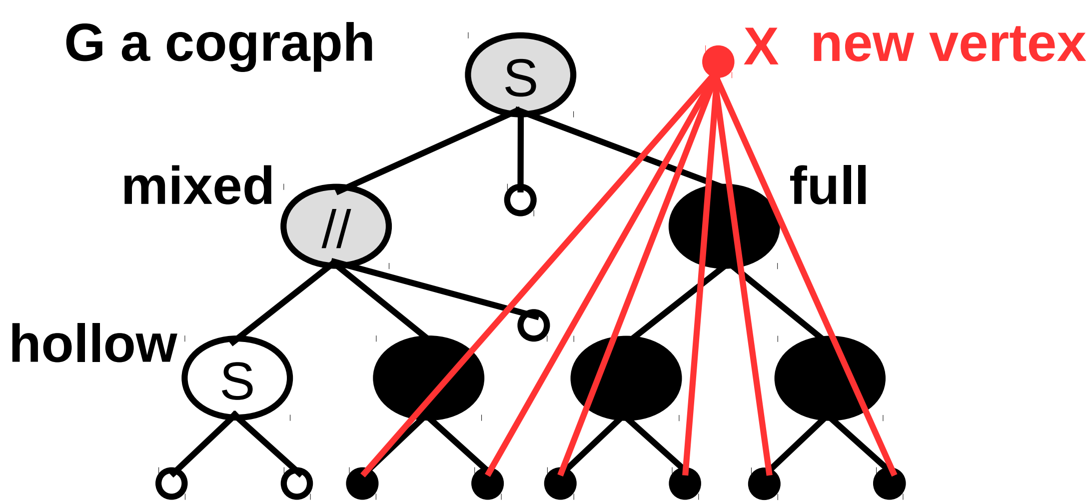




O(n) space

Incremental approach: a **cograph G** and **x** a new incoming vertex

G+x is not a cograph and we want to add (and/or delete) edges incident to x so that G+x become a cograph



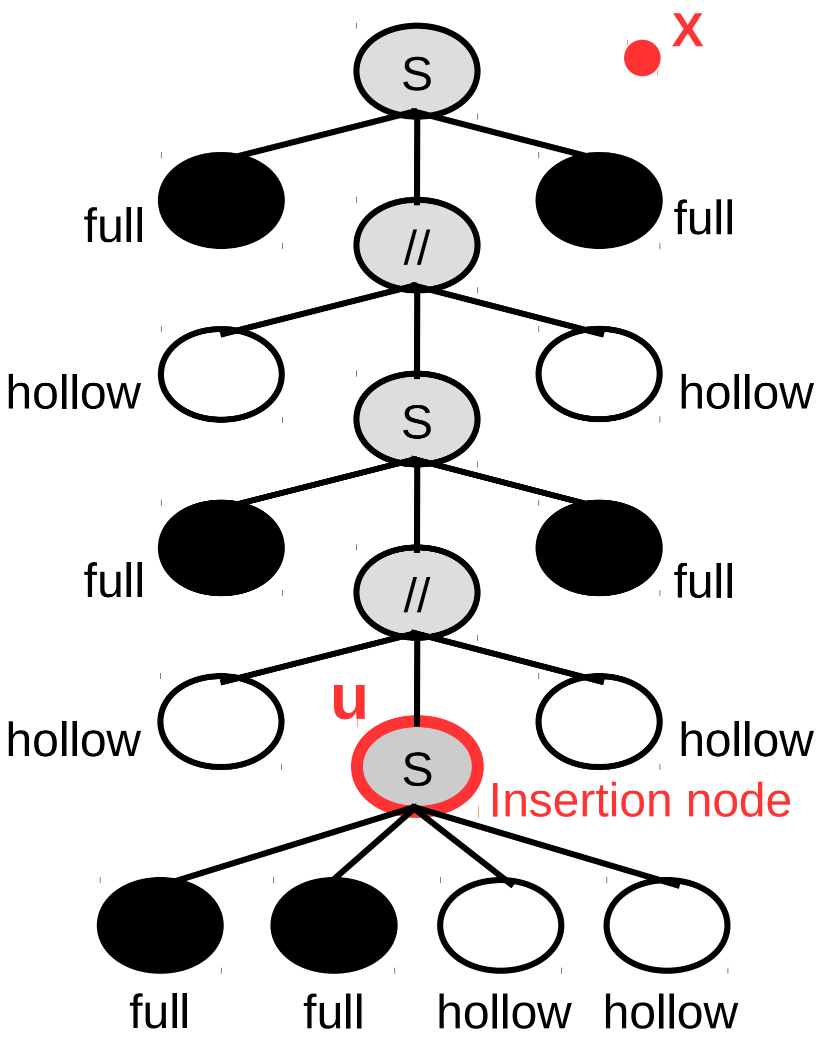
Completion algorithms

First algorithm: $O(n+m')$

A characterisation of cographs

[Corneil, Perl, Stewart 1981]

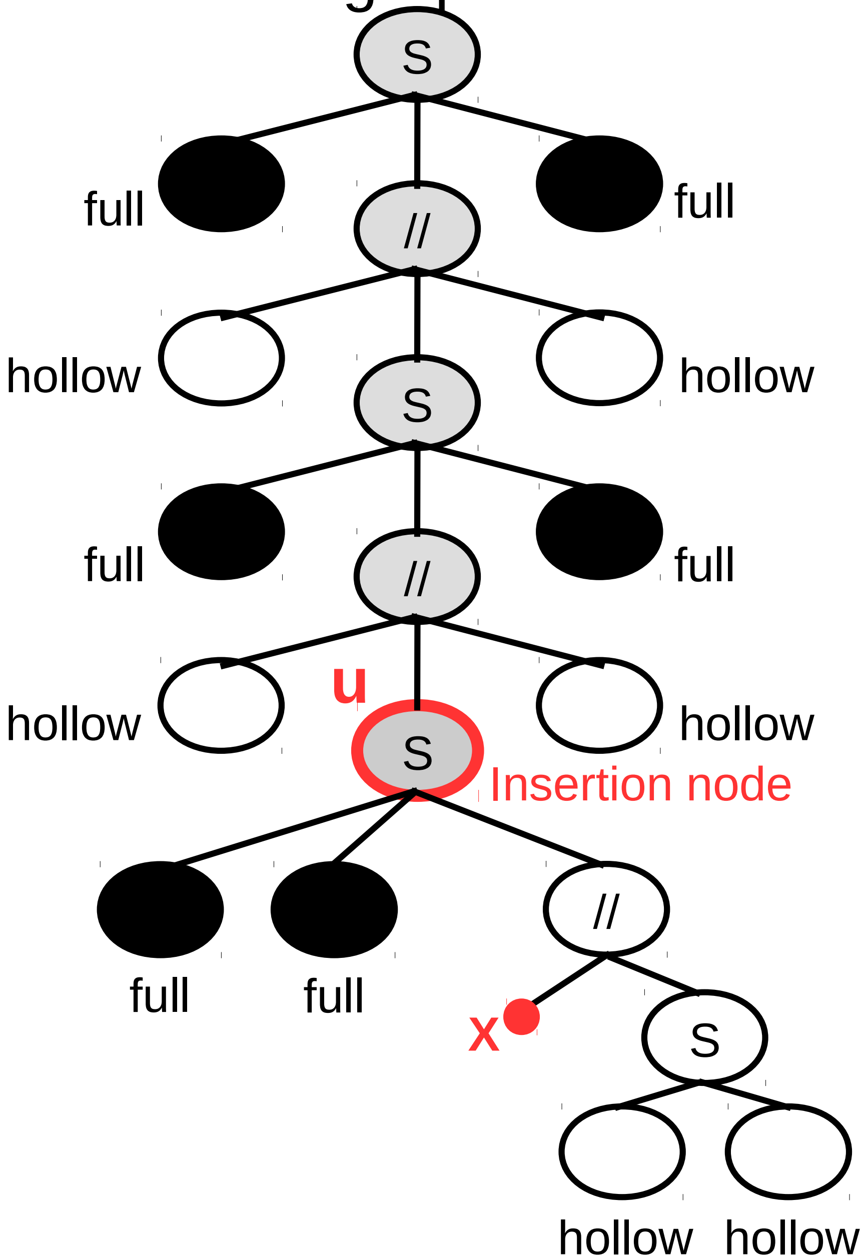
G+x is a cograph iff there exists a node u st.:



A characterisation of cographs

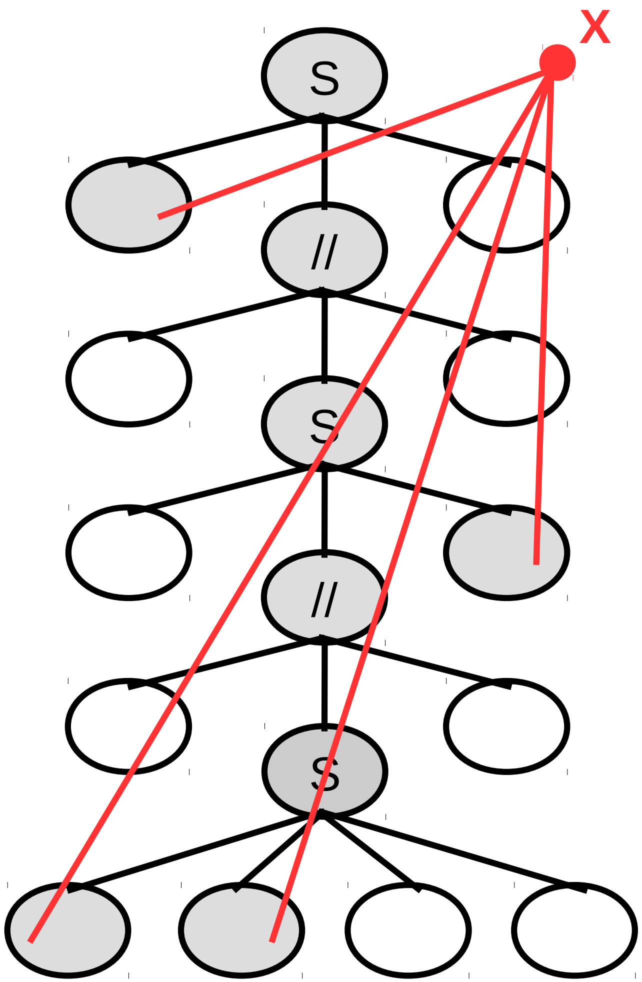
[Corneil, Perl, Stewart 1981]

G+x is a cograph iff there exists a node u st.:



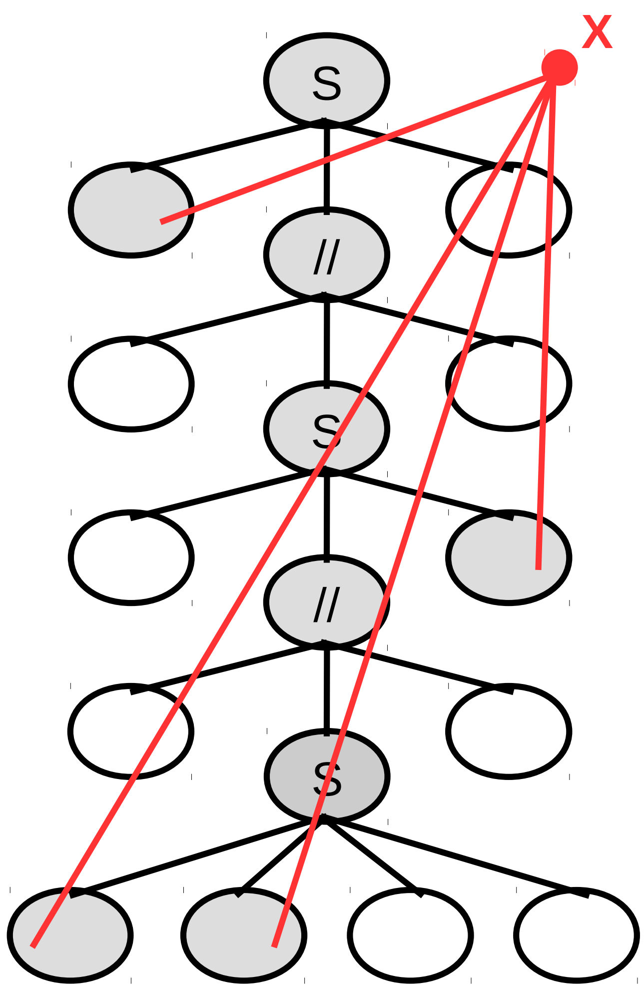
A characterisation of cographs

In our algorithm : G+x is not a cograph



A characterisation of cographs

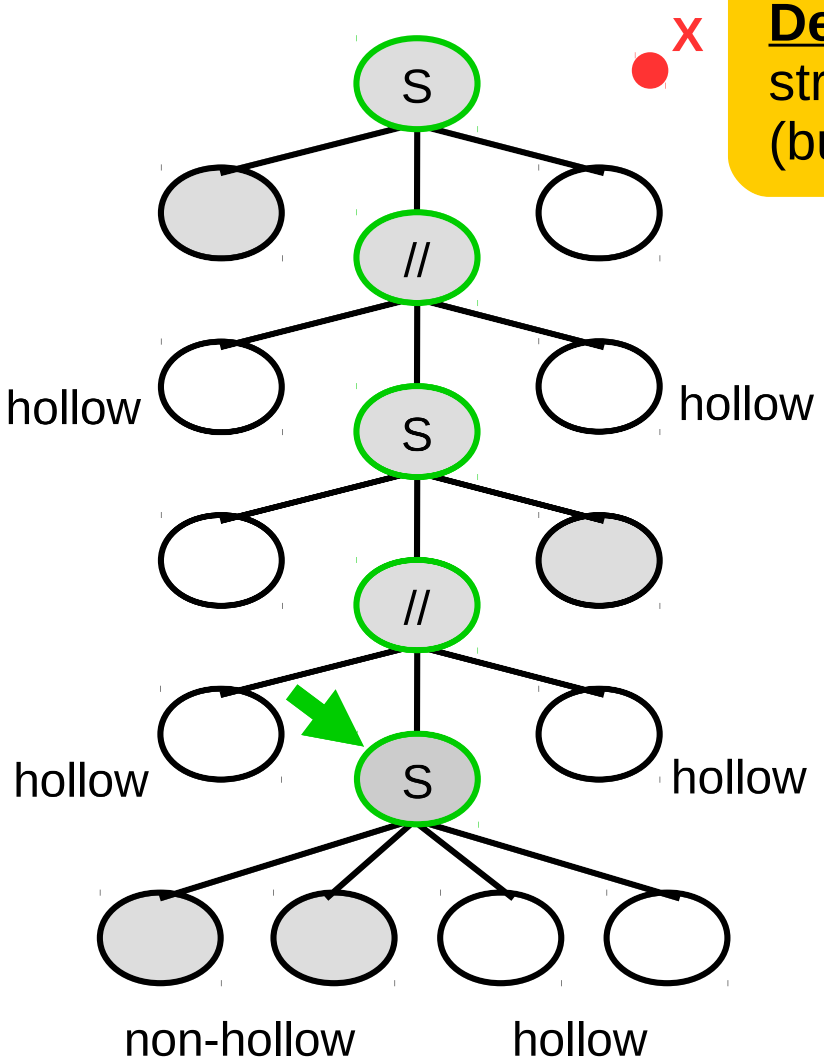
In our algorithm : G+x is not a cograph



Choose one node u for which you make the situation of the [CPS 81]'s theorem happen

Eligible nodes

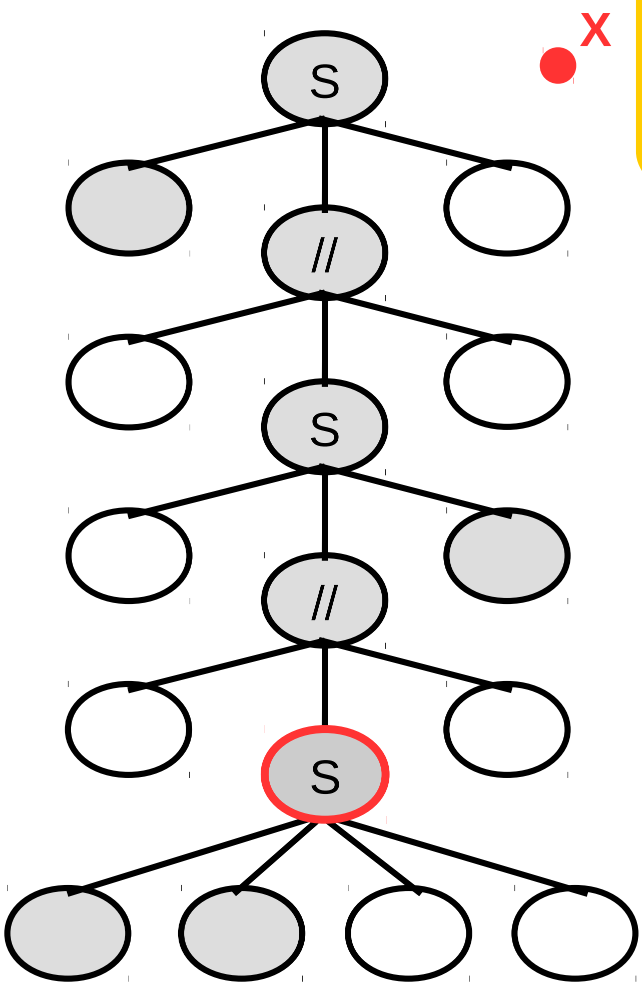
In our algorithm : G+x is not a cograph



Definition: u is an ***eligible node*** Iff all parallel strict ancestors of u are such that all their children (but one) are hollow

Completion anchored at u

In our algorithm : G+x is not a cograph



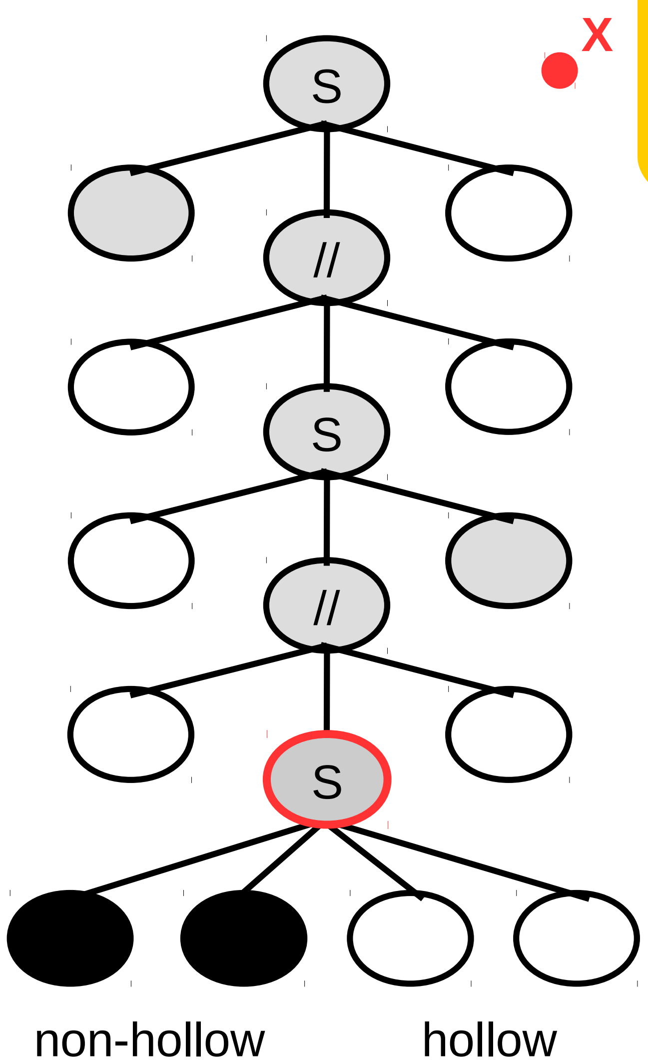
<u>Definition:</u> u is an ***eligible node*** Iff all parallel strict ancestors of u are such that all their children (but one) are hollow

<u>Proceed as follows :</u>

1) choose one eligible node ***u***

Completion anchored at u

In our algorithm : G+x is not a cograph



Definition: u is an ***eligible node*** Iff all parallel strict ancestors of u are such that all their children (but one) are hollow

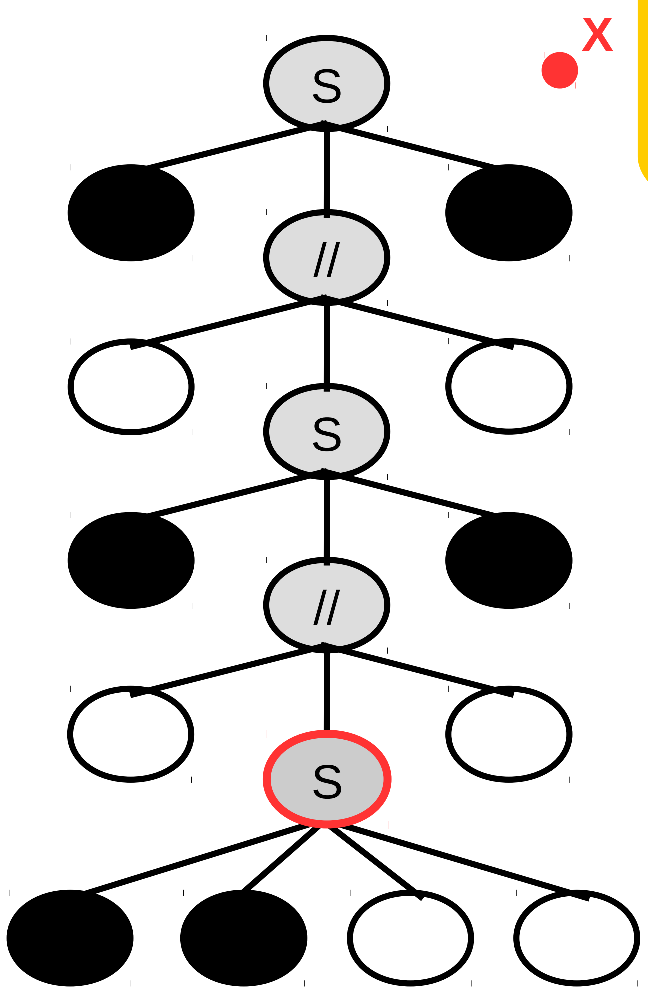
Proceed as follows :

1) choose one eligible node ***u***

2) make the non-hollow children of u become ***full*** (leave the others ***hollow***)

Completion anchored at u

In our algorithm : G+x is not a cograph



<u>Definition:</u> u is an ***eligible node*** Iff all parallel strict ancestors of u are such that all their children (but one) are hollow

<u>Proceed as follows :</u>

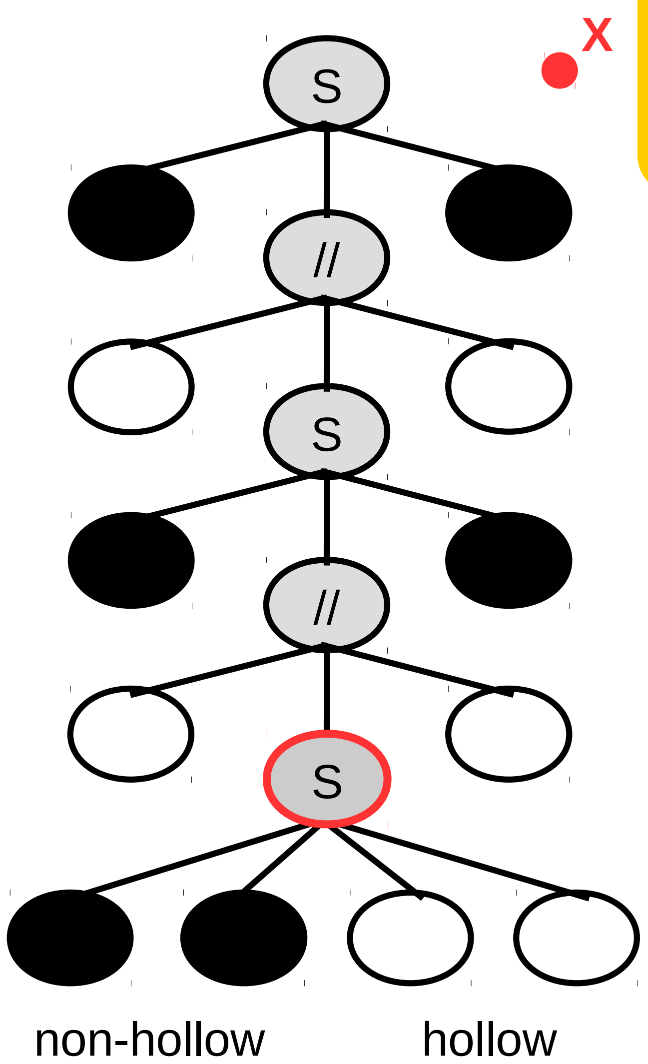
1) choose one eligible node ***u***

2) make the non-hollow children of u become ***full*** (leave the others ***hollow***)

3) for each ***series ancestor*** v of u, make all its children (but one) ***full***

Completion anchored at u

In our algorithm : G+x is not a cograph



<u>Definition:</u> u is an ***eligible node*** Iff all parallel strict ancestors of u are such that all their children (but one) are hollow

<u>Proceed as follows :</u>

1) choose one eligible node ***u***

2) make the non-hollow children of u become ***full*** (leave the others ***hollow***)

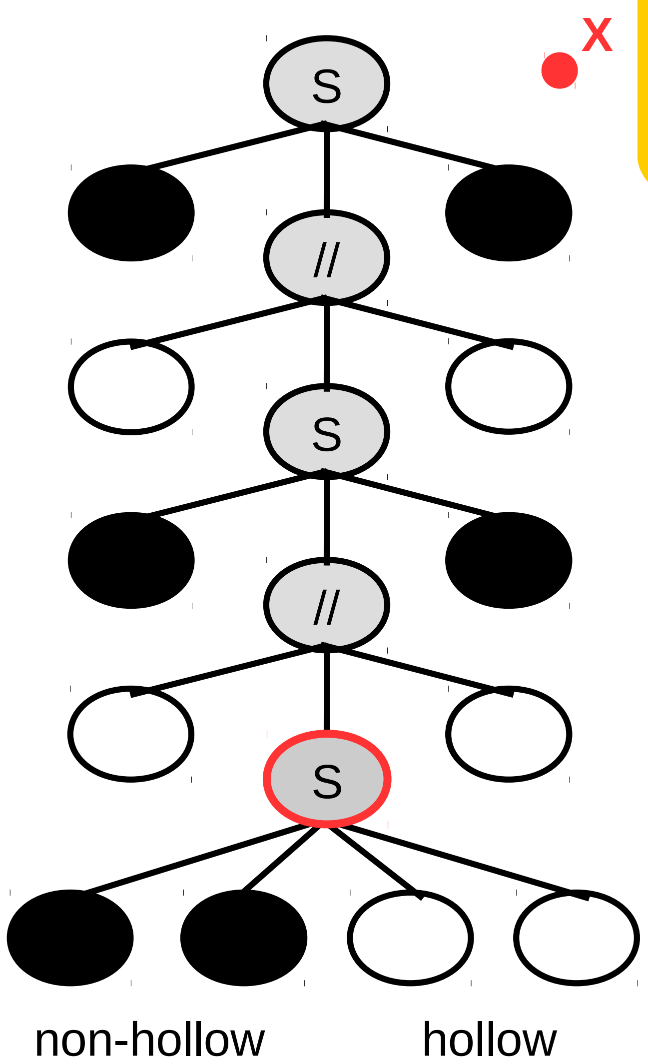
3) for each ***series ancestor*** v of u, make all its children (but one) ***full***

→ **you obtain a cograph completion of G+x**

called the ***completion anchored at u***

Completion anchored at u

In our algorithm : G+x is not a cograph



<u>Definition:</u> u is an ***eligible node*** Iff all parallel strict ancestors of u are such that all their children (but one) are hollow

<u>Proceed as follows :</u>

1) choose one eligible node ***u***

2) make the non-hollow children of u become ***full*** (leave the others ***hollow***)

3) for each ***series ancestor*** v of u, make all its children (but one) ***full***

➡ **you obtain a cograph completion of G+x**

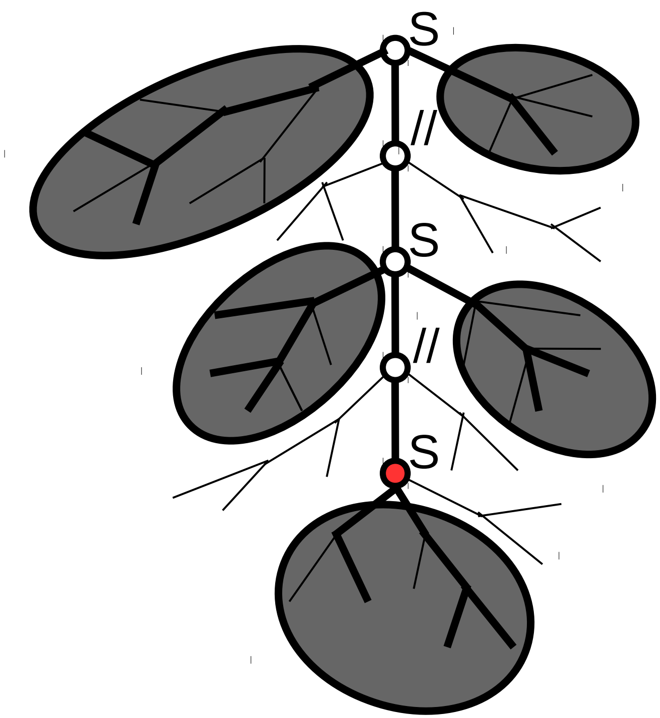
called the ***completion anchored at u***

<u>Question:</u> Is it minimal ?

➡ **We have a characterization for this**

First algorithm : O(n+m')

- Search the tree bottom up from the leaves adjacent to x
- Find the eligible nodes that satisfy the characterization



Note : we search only non-hollow nodes

Complexity : O(d')

[LMP 10]

- Choose one u of ***minimum cost*** and update the data structure by running **[CPS 81]**'s algorithm.

Complexity : O(d') for one incremental step
O(n+m') for the whole algorithm

Completion algorithms

Second algorithm: $O(n + m \log^2 n)$

Why is O(n+m') not necessarily optimal?

- No reason to use adjacency lists to encode the output

→ there is an O(n) space representation of cographs

Why is O(n+m’) not necessarily optimal?

- No reason to use adjacency lists to encode the output

→ there is an O(n) space representation of cographs

- What is the expected number of edges m’ in a cograph completion?
  - If the input G has the vertex-expansion property, then G’ has $O(n^2)$ edges
  - Random graphs with fixed average degree, **O(n) edges**, have the expansion property with high probability

→ **In practice, O(n+m’) ~ $O(n^2)$**

→ **We achieve $O(n+m \log^2 n)$ time**

Why is O(n+m') not necessarily optimal?

- No reason to use adjacency lists to encode the output

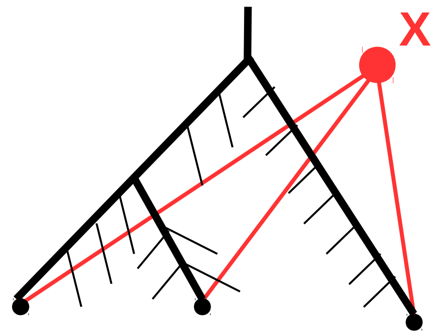
➡ there is an O(n) space representation of cographs

- What is the expected number of edges m' in a cograph completion?
  - If the input G has the vertex-expansion property, then G' has $O(n^2)$ edges
  - Random graphs with fixed average degree, **O(n) edges**, have the expansion property with high probability

➡ **In practice, O(n+m') ~ $O(n^2)$**

➡ **We achieve O(n+m $\log^2$n) time**

- Where is the room for improvement of the complexity?



A ***constant*** number of neighbours of x
can force to search an ***Ω(n)*** part of the co tree

Second algorithm : $O(n + m \log^2 n)$

- Note: we abandon the minimum incremental → ***only minimal***

- we use a dynamic data-structure for *lowest ancestor queries*
  [Sleator, Tarjan 1983]
  - In O(log n) time: w=lca(u,v) and w_u the child of w that is an ancestor of u
  - Update the structure in O(log n) time under elementary tree modifications

- we use *ordered lists*
  [Dietz, Sleator 1987]
  - In O(1) time: order between two elements in the list
  - Update the structure in O(1) time under deletion and insertion of an element

Second algorithm : $O(n + m \log^2 n)$

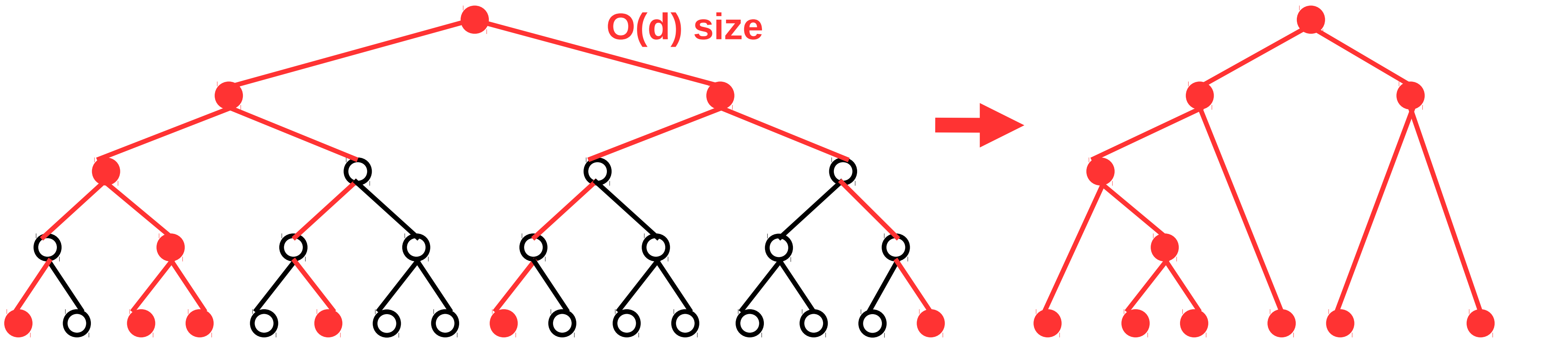
<u>Our goal :</u> determine the ***lowest*** eligible, non-hollow and non-forced nodes

➡ **minimal completion**

- Lowest eligible nodes

  ➡ highest parallel nodes with ≥2 non-hollow children

  - build T' : the subtree of lowest common ancestors of neighbours of x
  - Keep the highest parallel nodes in T'



1) sort neighbours of x from left to right : $O(d \log^2 n)$ time

2) insert neighbours one by one
Total : $O(d \log n)$ time

<u>Complexity :</u> $O(d \log^2 n)$ for one incremental step
$O(n+m \log^2 n)$ for the whole algorithm

Editing algorithm

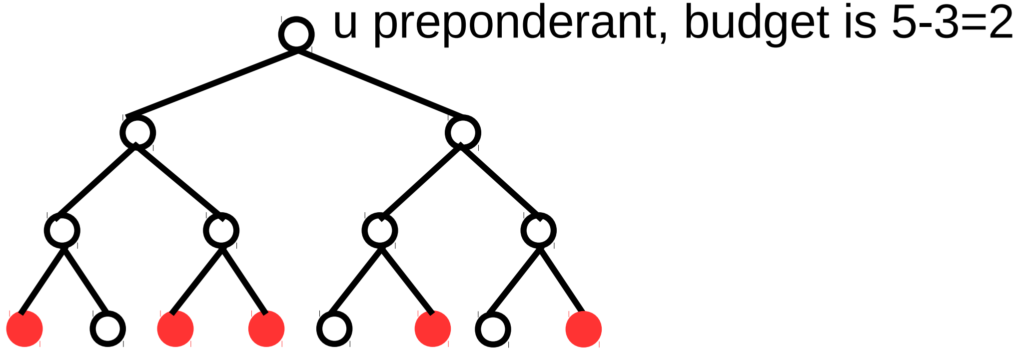
$O(n + m)$ time

Cograph editing

- Use ***both addition and deletion*** of edges
- Find a ***minimum cardinality modification*** at each incremental step
- Complexity **O(n+m)** time, O(d) time at each incremental step

Obs.: a minimum editing is not worse than deleting all edges incident to x

1) compute all maximal *preponderant* nodes and their budget



2) for each parent u of some preponderant node, climb in the tree and try to fill what must be by using only the budgets of the children of u
reach the root : success, otherwise : failure

→ ensures an O(d) time complexity

Coedit : use case

Cograph edition of real-world graphs

35 real-world graphs

+

8 random graphs

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

Cograph edition of real-world graphs

RESULTS

35 real-world graphs

+

8 random graphs

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

- Some networks are very close from cographs

Cograph edition of real-world graphs

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

35 real-world graphs

+

8 random graphs

RESULTS

- Some networks are very close from cographs
- Random graphs are never

Cograph edition of real-world graphs

35 real-world graphs

+

8 random graphs

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

RESULTS

- Some networks are very close from cographs
- Random graphs are never
- A wide range of proximity : 12% to 93%

Cograph edition of real-world graphs

35 real-world graphs

+

8 random graphs

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

RESULTS

- Some networks are very close from cographs
- Random graphs are never
- A wide range of proximity : 12% to 93%
- The proximity with cographs highly depends on the real-world context

Cograph edition of real-world graphs

Close to cographs

- www
- software

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

- The proximity with cographs highly depends on the real-world context

Cograph edition of real-world graphs

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

Not close not far

- internet
- road

- The proximity with cographs highly depends on the real-world context

Cograph edition of real-world graphs

Context	Network	n	m	d°	%mod
WWW	in-2004	1 148 875	12 281 937	21.4	12 %
WWW	cnr-2000	227 058	2 187 201	19.3	19 %
PROTEIN	reactome	5 973	145 778	48.8	22 %
SOFTWARE	jdk	6 434	53 658	16.7	29 %
SOFTWARE	jung-j	6 120	50 290	16.4	29 %
WWW	eu-2005	835 044	15 718 784	37.7	29 %
CO-AUTHOR	ca-GrQc	4 158	13 422	6.5	34 %
CO-AUTHOR	ca-HepPh	11 204	117 619	21.0	34 %
SPECIES	foodweb	183	2 434	26.6	43 %
CO-AUTHOR	dblp	317 080	1 049 866	6.6	45 %
WORD-REL.	wordnet	145 145	656 230	9.0	48 %
COMMUNIC.	wiki-Talk	2 388 953	4 656 682	3.9	49 %
CO-SOLD	amazon	334 863	925 872	5.5	49 %
CO-AUTHOR	ca-CondMat	21 363	91 286	8.6	52 %
RANDOM	ER-Gnm_1M-2	796 208	958 827	2.4	52 %
CO-AUTHOR	ca-HepTh	8 638	24 806	5.7	54 %
INTERNET	as2000	6 474	12 572	3.9	54 %
ROAD	roadNet-TX	1 351 137	1 879 201	2.8	54 %
INTERNET	as-caida2007	26 475	53 381	4.0	55 %
CO-AUTHOR	ca-AstroPh	17 903	196 972	22.0	59 %
INTERNET	topology	34 761	107 720	6.2	61 %
RANDOM	ER-Gnm_1M-3	940 987	1 494 643	3.2	63 %
INTERNET	as-skitter	1 694 616	11 094 209	13.1	64 %
CO-OCCUR	bible-names	1 707	9 059	10.6	67 %
PROTEIN	figeys	2 217	6 418	5.8	67 %
CITATION-SCI.	cora	23 166	89 157	7.7	68 %
SOCIAL	youtube	1 134 890	2 987 624	5.3	69 %
CO-ACTOR	actor-col.	374 511	15 014 839	80.2	71 %
P2P-CONNECT.	p2p-Gnutella	62 561	147 878	4.7	71 %
RANDOM	ER-Gnm_1M-4	980 191	1 999 203	4.1	71 %
CITATION-SCI.	citeseer	365 154	1 721 981	9.4	75 %
CITATION-PAT.	cit-Patents	3 764 117	16 511 740	8.8	76 %
SOFTWARE	linux	30 817	213 208	13.8	77 %
SOCIAL	LiveJournal	3 997 962	34 681 189	17.4	78 %
CITATION-SCI.	cit-HepTh	27 400	352 021	25.7	79 %
RANDOM	ER-Gnm_1M-6	997 479	2 999 988	6.0	79 %
CITATION-SCI.	cit-HepPh	34 401	420 784	24.5	81 %
RANDOM	ER-Gnm_1M-8	999 684	3 999 999	8.0	84 %
RANDOM	ER-Gnm_1M-10	999 952	5 000 000	10.0	87 %
RANDOM	ER-Gnm_1M-15	1 000 000	7 500 000	15.0	91 %
SOCIAL	orkut	3 072 441	117 185 083	76.3	91 %
RANDOM	ER-Gnm_1M-20	1 000 000	10 000 000	20.0	93 %
WORD-REL.	Thesaurus	23 132	297 094	25.7	93 %

Far from cographs

citation

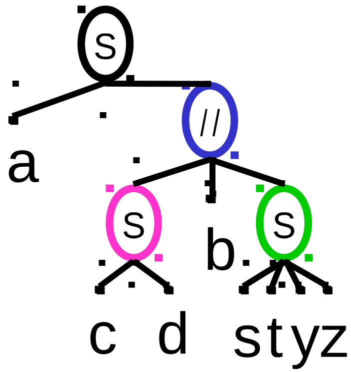
social

- The proximity with cographs highly depends on the real-world context

Testing the modelling approach

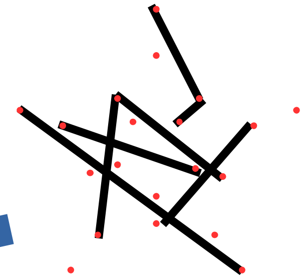
1 strongly structured

cographs obtained from edition

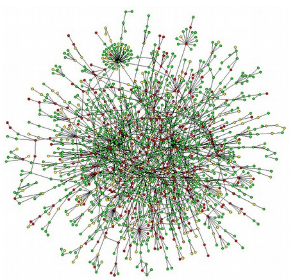


+

2 random modifications



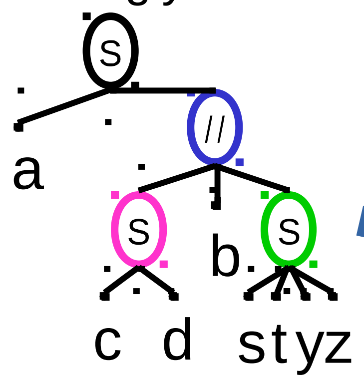
same number as in edition problem



compare with the original real-world network

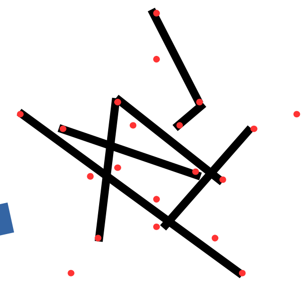
Conclusion

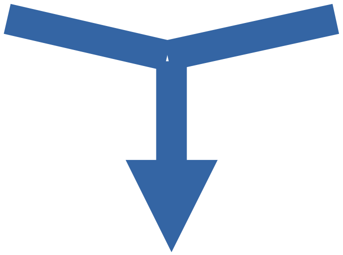
1 strongly structured

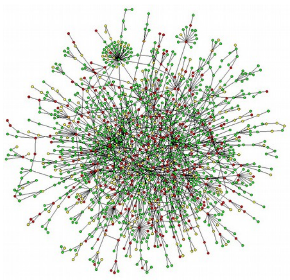


\+

2 random modifications



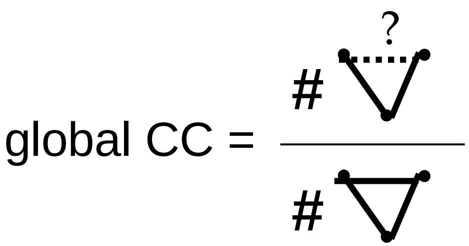


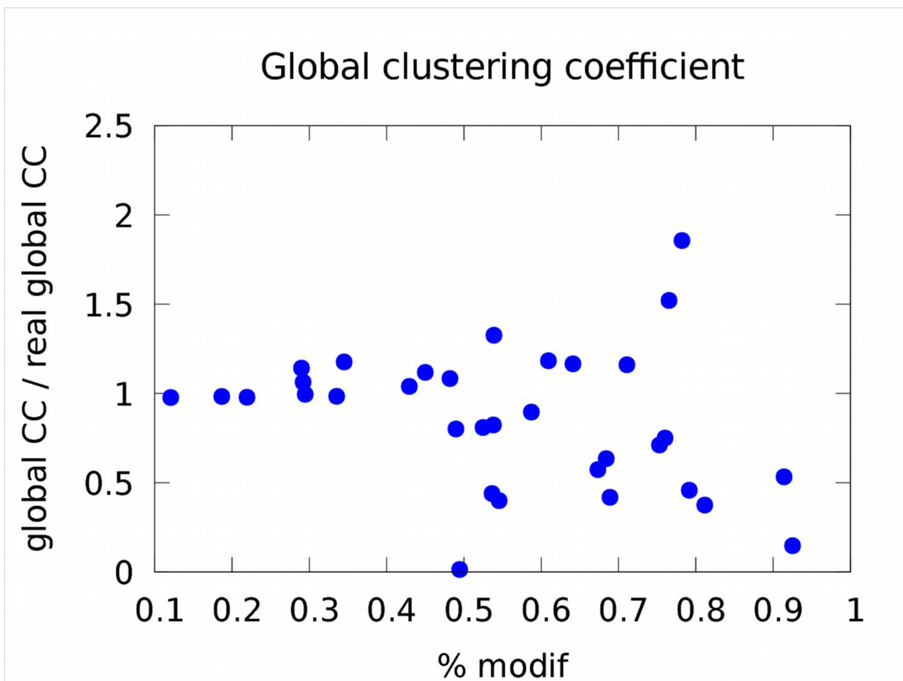


- ✔ global density
- ✔ distances
- **?** degree distribution
- **?** local density

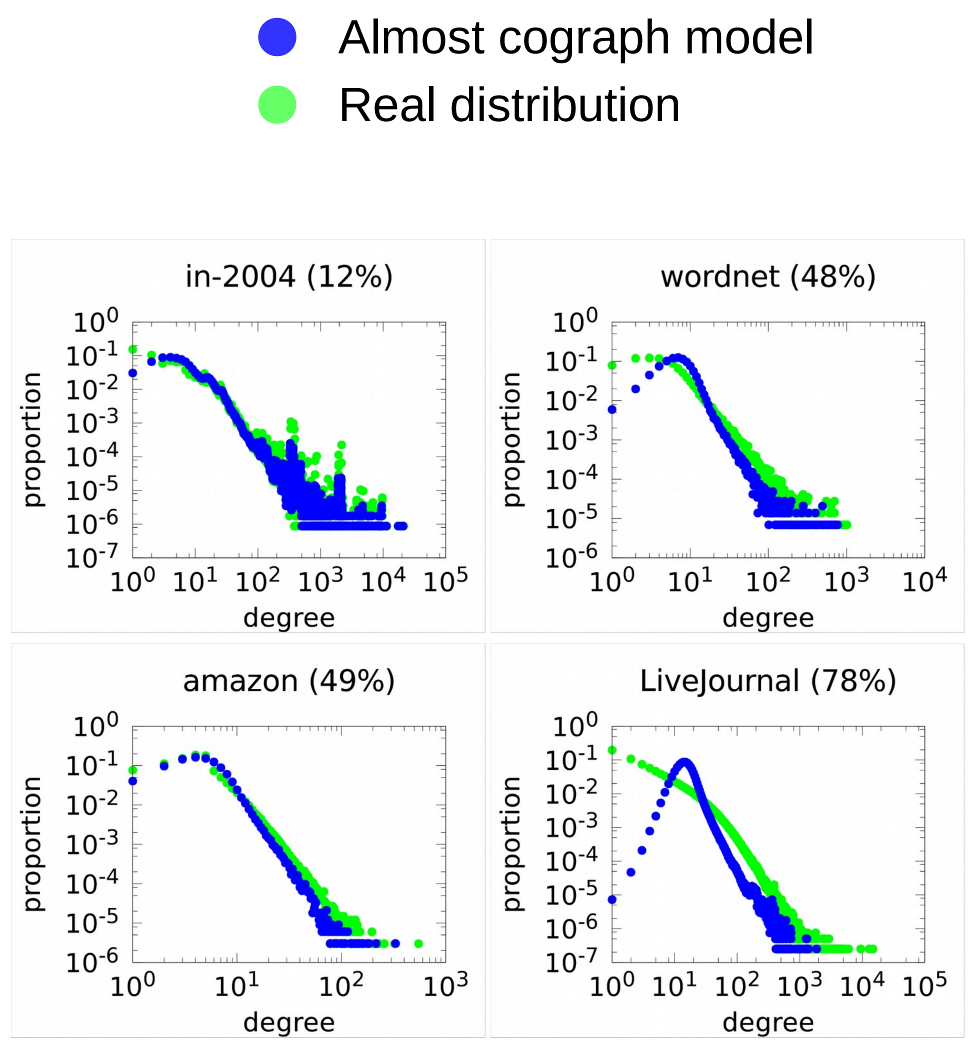
Results of generation

Local density



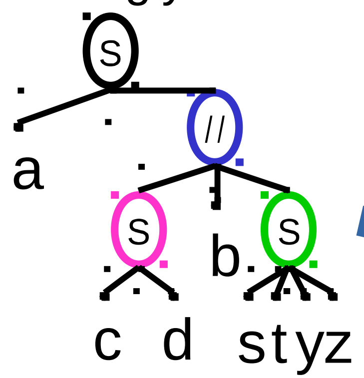


Degree distribution



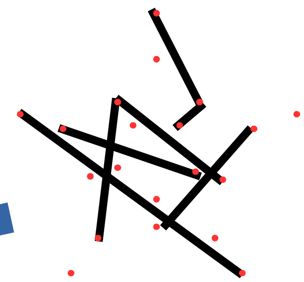
Conclusion

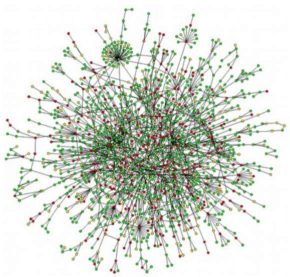
1 strongly structured



+

2 random modifications

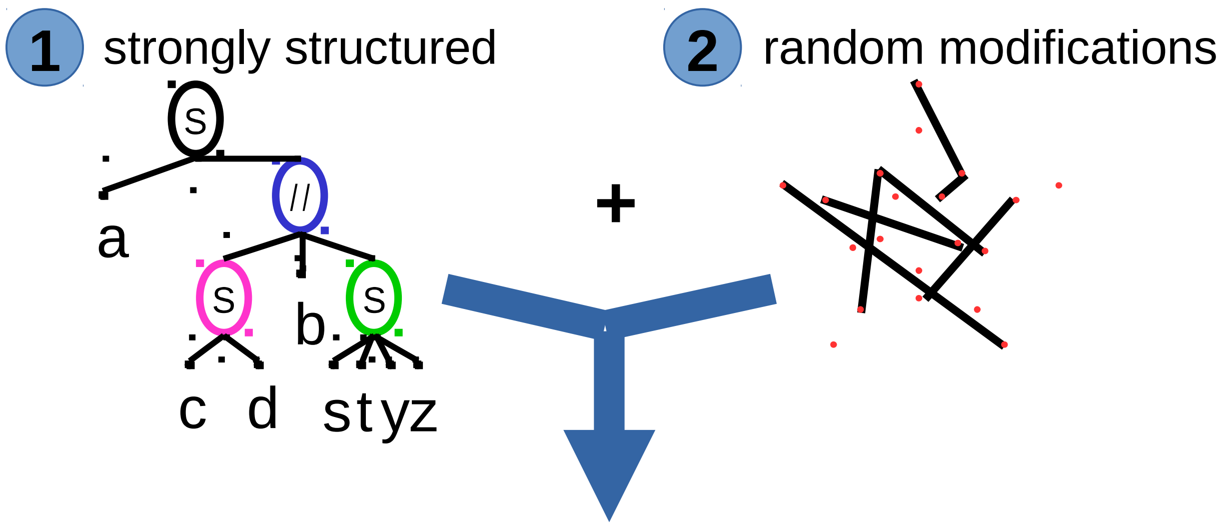




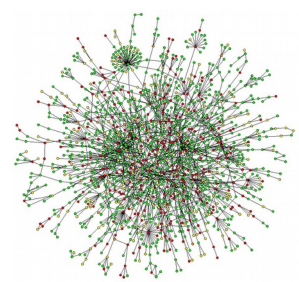
- global density
- distances
- degree distribution
- local density

The cograph structure successfully captures these properties

Conclusion



The cograph structure successfully captures these properties



- global density
- distances
- degree distribution
- local density

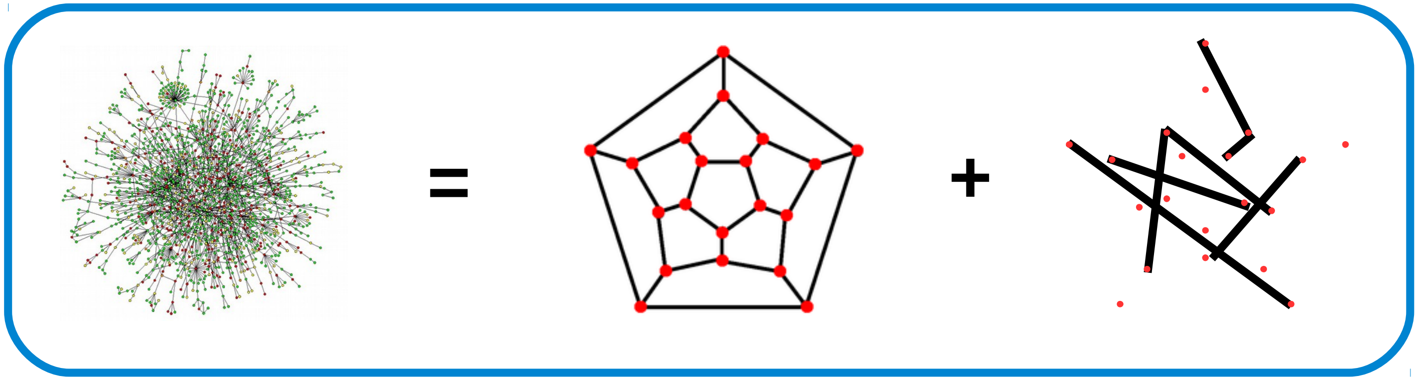
To complete the model

- Edit a real-world graph into a cograph
- Generate a similar cotree
- Apply random modifications to the cograph

Perspectives

- Assess the quality of the set of modifications obtained from the inclusion-minimal approach

- Consider other graph classes suitable for other kinds of networks
  - Chordal graphs → social networks, citations
  - Related to planar graphs → internet, road networks



- Other possibilities of this representation
  - Efficient encoding
  - Algorithmics of almost structured graphs

PROXNET - Modelling Complex Networks Through Graph Editing Problems

Marie Sklodowska-Curie Actions of the European Union

About PROXNET

PROXNET is a project funded by the MSCA program of the European Union. It is hosted at the University of Bergen, with principal researcher Christophe Crespelle and supervised by Pinar Heggernes.

The goal of the PROXNET project is to open a new way for analysing, modelling and managing complex networks, through graph editing problems. The reason why these networks are said to be complex is that they are loosely structured, due to the part of uncertainty and randomness they contain. On the other hand, the real-world context where they come from strongly constrains their organisation and gives them some specific structure. The difficulty in retrieving this structure is that it is altered by the noise resulting from the uncertainty and randomness that these networks contain. In the PROXNET project, we retrieve the hidden structures of complex networks thanks to graph editing problems, which consist in changing some adjacencies of the graph in order to obtain a desired property. We develop the algorithms necessary to solve graph editing problems on huge instances of graphs, we apply them to real-world datasets and use the results obtained in order to design new models of complex networks.

Contact information

e-mail christophe dot crespelle at uib dot no (dot = ".", at = "@")

Postal address Postboks 7803
5020 Bergen, Norway

Location Høyteknologisenteret
Thormøhlens Gate 55, Bergen

News

01/23/2020 Workshop on Graph Modification: Algorithms, Experiments and New Problems in Bergen, Norway.
06/03/2019 Workshop on Kernelization in Bergen, Norway.
03/04/2019 Conference on Algorithms, Optimization and Learning in Dynamics Environments in Hanoi, Vietnam.
11/15/2018 Graph Theory and Applications Workshop in Hanoi, Vietnam.
09/17/2018 Operation Research + Parameterized Complexity Workshop in Solstrand, Norway.
08/09/2018 China-Norway FPT workshop in Bergen, Norway.
03/21/2018 16th Annual Winter School in Algorithms, Graph Theory and Combinatorics in Geilo, Norway.

Software

Coedit Minimal completion, deletion and editing of an arbitrary graph into a cograph.
Released January 2020.
sources

Coedit: a tool for minimal cograph edge modification

Christophe Crespelle

University of Bergen

with Daniel Lokshtanov, Thi Ha Duong Phan and Eric Thierry