Tagging SNPs Selection

2 Algorithms

A1 LD-Select

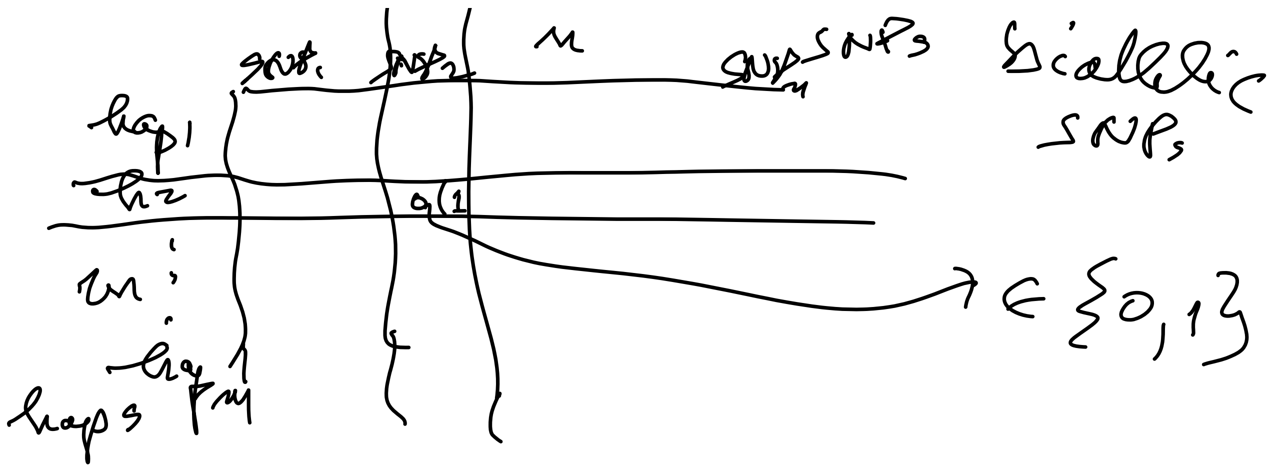
A2 Informativeness

A1 contivuation

„most informative SNPs"

Input is a matrix

haplotypes × SNPs

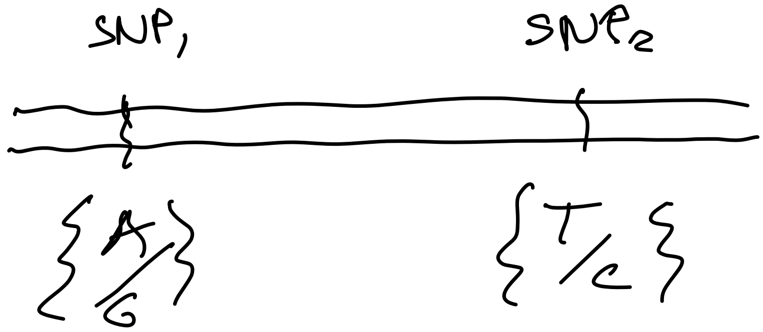


We have individuals genotyped at every SNP of a certain set of SNPs part of an array or assay of SNPs.

The result: for each SNP_i and individual j

output: $\begin{Bmatrix} A \\ T \end{Bmatrix}$ the two alleles of the individ. j at SNP_i.

individual j



sets

$\{A, G\}$, $\{T, C\}$, . . . $\{A, A\}$

We need haplotypes

$\{A, G\}$, $\{T, C\}$

Possible haplotypes:

$$\left.\begin{array}{cc} A & T \\ A & C \\ G & T \\ G & C \end{array}\right\} \text{4 but only 2 occur in individual } j$$

With one individual only
we cannot infer which two
of the 4 haplotypes are
the true haplotypes of
individ j.

If we have a set of individuals
human genomes:
This problem is called the haplotype phasing PB.
$\boxed{\underline{PHASE}}$

$\underline{\text{haps}}$: sequences over the alphabet $\{0,1\}$
the two alleles are
0 and 1
0101110

$\underline{\text{genotypes}}$: sequences over the alphabet
$\{0,1,2\}$

individual $\begin{cases} 0101110 \\ 0001001 \end{cases}$ "2" = ambiguous

genotype (summary): 0201222

1, 0 } determined

ex. hap over just 3 SNPs genotypes — " —

022 → {010, 001} → "explanations"

022 → {011, 000} → "explanations"

The phasing problem has a search space exponential in the number of 2's in the genotype sequence

---

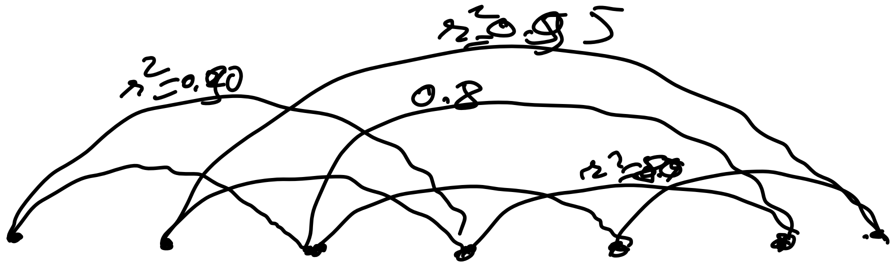
INPUT: for the tagging SNPs selection pb. is NP

haps $\begin{cases} 010110101111 \\ 000101010111 \\ 111010110101 \end{cases}$

i * j *

$r^2(i,j)$

LD-select r^2



SNPs

Complete undirected graph over the n SNPs/vertices and each edge is labeled by the r^2 value between the two vertices joined by the edge.

LD-select is a greedy algorithm. Two constraints as input to the algorithm:

(1) only <u>common</u> SNPs are part of the input

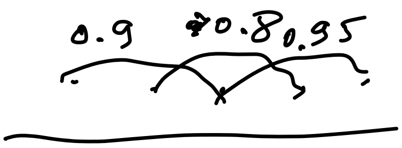
$\geq 10\%$ of people in the population have those SNPs

(2) there is an r^2 threshold

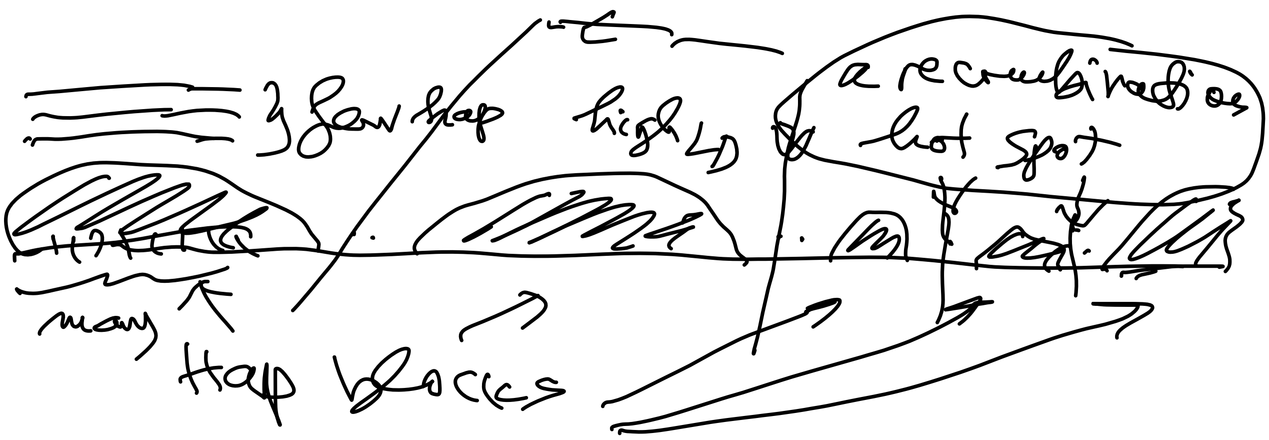
given: τ

Only edges with $r^2 \geq \tau$ are considered

e.g. $r^2 \geq 0.8$



Haplotypes Block



Max informative : vertex that is connected to most other vertices with edges with $r^2 \geq \tau$.

---

1) Max informative vertex (MIV) → start a bin

$\underbrace{\{ MIV, \ldots\ldots\ldots \}}$ Bin 1

2) MIV $\underbrace{\{ MIV_2, \ldots \}}$ Bin 2

⋮

$\underbrace{\{ \quad \}}$ Bin 805

The tagging SNPs :

$MIV_1, MIV_2, \ldots, MIV_{805}$

Paper: " Selecting a Maximally informative set of Single Nucleotide polymorphisms for association analysis using linkage disequilibrium"

Am. J Hum. Gen (2004)

Carlson et al

Leonid Kruglyak

Debbie Nickerson

---

2 populations

- African American AA
- European American EA

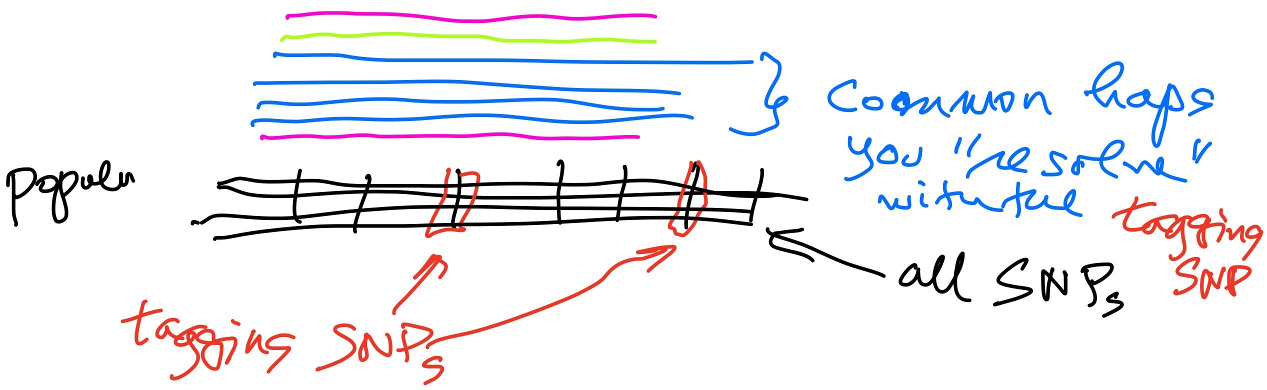
Inferred haplotypes (To Phase or Not to Phase)
PHASE

r^2 was calculated between pairs of SNPs in each population

---

haplotypes → common (many people have those haplotypes)
→ rare

One measure of success for selecting tagging SNPs is



Their LD-blocks captured/resolved $\geq 80\%$ of the common haps

Capture the $\boxed{\text{variation} = \text{haplotypes}}$
$\geq 80\%$ of common haps

Results

100 genes resequenced them
in 24 African Americans
23 European Americans

Gene average: 16.5 kb length
longest 45 kb length

8877 SNPs over all
7793 SNPs in AA pop
4620 SNPs in EA pop

Density 1 SNP in 200 bp

very small # of three-allelic SNPs

when MAF > 10% ⇒ 3178 SNPs in AA
min allele freq
2375 in EA

Common SNPs 1 in 300 bp AA
1 in 700 bp EA

SNPs in coding sequences (genes) cSNPs
8% of seq length (135 kb)

The Second Algorithm

INFORMATIVENESS

LD-Select: the objective function way: Minimum DOMINATING SET

Informativeness: the objective function is: Minimum SET COVER

$\boxed{\text{DOMINATING SET} \longleftrightarrow \text{SET COVER}}$

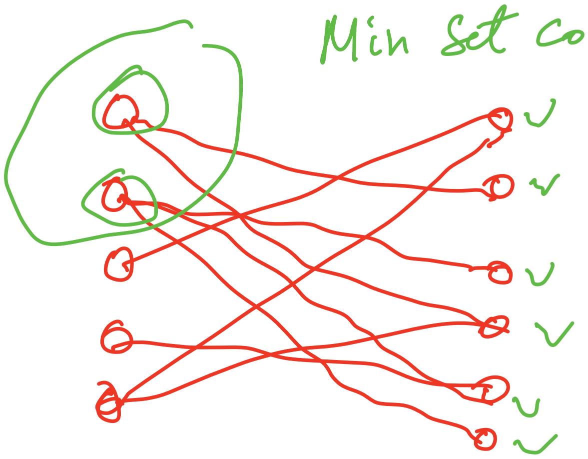
BOTH PBs are NP-Complete

reducible to each other

"unify LD-Select & Informativeness?"

SET COVER

Min set cover



Boys Girls

Compute: The min # of boys that know all the girls

"knows"

boy i girl j

INFORMATIVENESS ALGORITHM

"information theory"

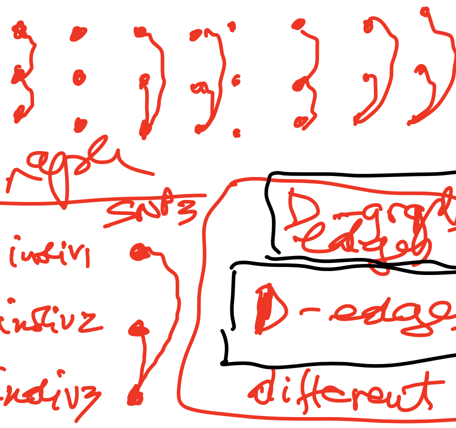
"bit": what a SNP does? what information it gives

SNP_3 $m = 8$

$n = 3$

indiv 1	0	0	0	0	1	0	1	1
2	1	0	1	0	1	1	1	1
3	0	0	0	1	1	0	0	0

column graphs 1 2 3



SNP_3 graph

SNP3

indiv1

indiv2

indiv3

D-graphs edges

D-edges

different alleles

A SNP provides one "bit" of information

indiv 1 and indiv3 have different alleles at the SNP

"

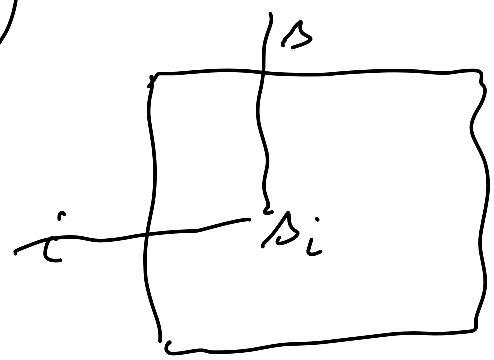
D-edge "distinguishing" edge

"different" edge

INPUT. A matrix n haplotypes × m SNPs

0-1 matrix

M



$s_i = M[i, s]$

allele

Def For SNP s and two haplotypes i and j

let us consider

$$M[i, s] \overset{=}{\underset{\neq}{}} M[j, s]$$

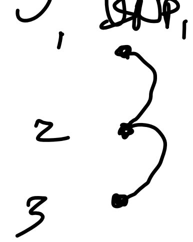
How much information exists in one SNP about another SNP?

Can i predict something about j?

Let $E(s)$ = edgeset of s in M
is the set of D-edges in the
associated <u>Column graph</u>

In the above example:

Column graph of SNP_1



$E(s_1) = \{(1,2),(2,3)\}$

SNP_3

$E(s_3) = \{(1,3),(2,3)\}$

$E(s_1), \ldots, E(s_8)$

<u>The information in SNP s with respect to SNP t</u>

$$I(s,t) = \frac{|E(s) \cap E(t)|}{|E(t)|}$$

$|E(s)|$ = size of the edge set

$E(s_1) = \{(1,2), (2,3)\}$

$E(s_2) = \emptyset$

$E(s_3) = \{(1,3), (2,3)\}$

$E(s_4) = \{(1,3), (2,3)\}$

$E(s_5) = \emptyset$

$E(s_6) = \{(1,2), (2,3)\}$

$E(s_7) = \{(1,3), (2,3)\}$

$E(s_8) = \{(1,3), (2,3)\}$

$$I(s_1, s_8) = \frac{1}{2}$$

$$I(s_1, s_4) = \frac{1}{2}$$

Generalize to sets of SNPs

S, T two sets of SNPs

$$I(S,T) = \frac{|S \cap T|}{|T|} = \frac{\left|\left(\bigcup_{s\in S} E(s)\right) \cap \left(\bigcup_{s\in T} E(s)\right)\right|}{\left|\bigcup_{s\in T} E(s)\right|}$$

$S = \{s_1, s_3\}$

$T = \{s_2, s_4, s_5, s_6, s_7, s_8\}$

$$I(S,T) = \frac{3}{3} = 1$$

Look at $S = \{s_1, s_3\}$

<u>Def</u> the <u>minimum informative subset of SNPs</u> (more precisely the <u>min set</u> of SNPs of <u>maximum information</u>) (MIS)

$\boxed{S = \{s_1, s_3\} \text{ is the } \underline{\text{MIS}} \text{ for our example}}$

Another <u>MIS</u> is $S' = \{\Delta_4, \Delta_6\}$

~~Extendable to many SNPs not only pairwise.~~

A conservative extension of the pairwise case.

This was not a property of r^2: Cannot be generalized conservatively to many SNPs.

"break the curse of the pairwise"

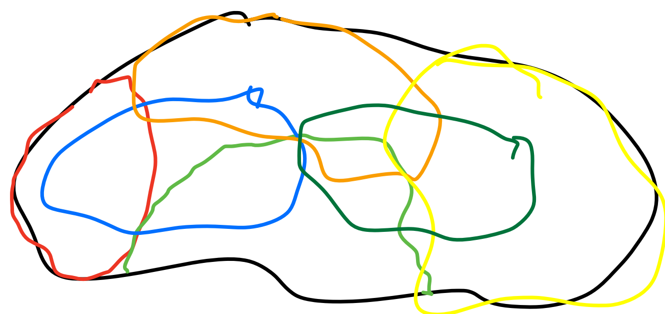
<u>Information-Theory based: 1 bit</u>

<u>The Set Cover Problem</u>

Given: U = a universe (a set of elements)

$R_i \subseteq U$, $\{R_i \mid i \in I\}$

Compute: A set cover is a collection of the R_i subsets whose union is the entire universe U.



An example:

$U = \{a, b, c, d, e\}$

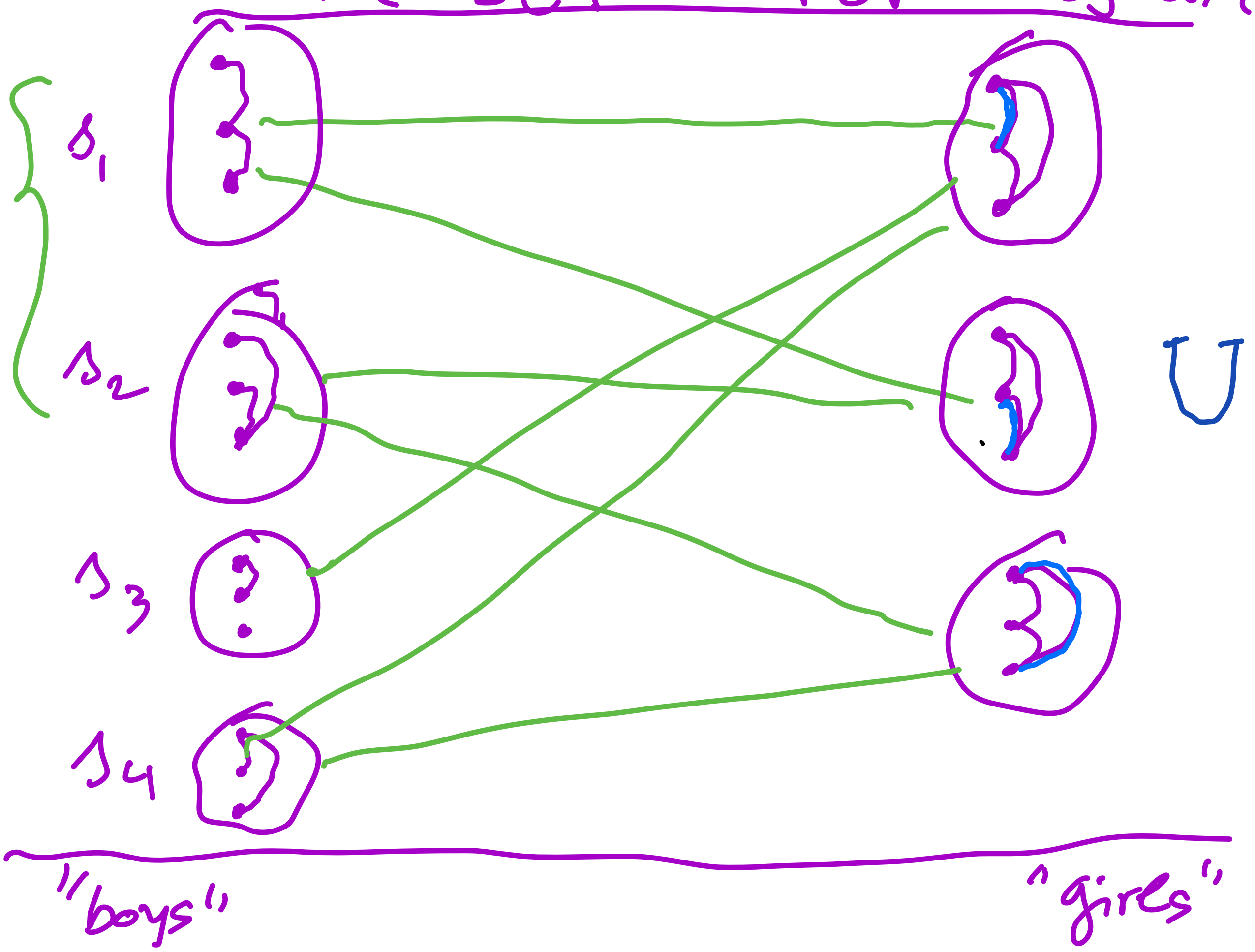
$R_1 = \{a, b, e\}$, $R_2 = \{a, b\}$

$R_3 = \{b, c, d\}$, $R_4 = \{c, d, e\}$

$\boxed{R_1, R_4}$ = a set cover

R_2, R_4 also a set cover

MIS = is a set cover PB
i.e. the objective function
is the SET COVER objective



~~add~~ a threshold Σ
MIS $\Sigma \geq 80\%$
find the min ~~set of~~ boys that know at least 80% of the girls

HW2 PB3

LD-select with $r^2 \geq \tau$ e.g. 0.8

Informative with $\xi \geq \rho$ e.g. 0.8

Find a correlation between the two:

In fact DOMINATING SET, and SET COVER are NP-complete, but

Reducible to each other:

- If you can solve one efficiently you can solve the other with same time complexity alg.
- The same for approximation algorithms.

Def <u>dominating set</u>

$G = (V, E)$ an undirected graph

V = vertex set

E = edge set

A subset $D \subseteq V$ is called a <u>dominating set</u> if every vertex outside D is connected by at least one edge to a vertex in D.

A reduction between DOMINATING SET ~~and~~ and SET COVER

$\boxed{\text{① From DOMINATING SET to SET COVER}}$

Take a general graph $G = (V, E)$ with $V = \{1, 2, \dots, n\}$

We construct a SET COVER for G as follows.

$U = V$ universe of elements

consider a family of subsets of U defined as follows:

$$S = \{ S_1, S_2, \dots, S_n \}$$

one for each vertex of G.

$S_v = \{$ the set of vertices adjacent to v together with v $\}$

Consider D any dominating set of G.

let $C = \{ S_v \mid v \in D \}$.

want to show that C is a set cover. Indeed!

Moreover: $|C| = |D|$

Conversely: $W \subseteq V$

if $C = \{S_v : v \in W\}$ is a set cover

then want to show that W is a dominating set for G.

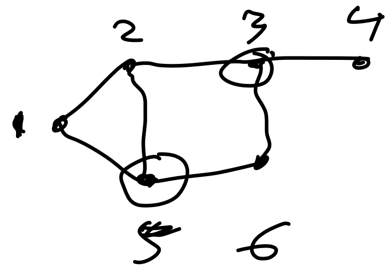
indeed!

One-to-one correspondence between dominating sets in G and set covers in (U, S).

min — min

A very simple algorithm to construct a set cover solution from a dominating set in G.

An example: From Dominating Set to Set Cover

G: 

example of dominating set

$D = \{3, 5\}$ min DS

Construct the Set Cover Instance

$S_1 = \{1, 2, 5\}$, $S_2 = \{1, 2, 3, 5\}$, $S_3 = \{2, 3, 4, 6\}$

$S_5 = \{1, 2, 5, 6\}$, $S_6 = \{3, 5, 6\}$

In G $D = \{3, 5\}$ is a dominating set

This corresp. to $C = \{S_3, S_5\}$ a set cover

② From SET COVER to DOMINATING SET

A general Set Cover instance

(U, S) U = universe

S = collection of subsets of U

$$S = \{ S_i : i \in I \}$$

Assume U and I are disjoint

Construct a graph $G = (V, E)$ as follows:

- the set of vertices $V = U \cup I$
- the set of edges:

type I $\{i, j\} \in E$

for every $i, j \in I$

type II $\{i, u\} \in E$
for every $i \in I$
$u \in S_i$

Note. I is a clique (completely connected graph)

Suppose that C is a set cover solution to (U, S) pb.

$C = \{S_i : i \in Z\}$
want to show that Z is a dominating set in G.

$Z \subseteq I$
then Z is a dominating set for G.

. First for each $u \in U$ there

is $i \in Z$ such that $u \in S_i$
and by construction
u and i are adjacent
in G.
hence u is dominated by i

- Second, since Z must be nonempty
each $i \in I$ is adjacent to
a vertex in Z.

So Z is a dominating set.

<u>Conversely</u>: let D be a dominating
set for G
Then it is possible to construct
another dominating set X
such that $|X| \leq |D|$

and $X \subseteq I$:

simply replace each $u \in D \cap V$ by a neighbor $i \in I$ of u.

Then $C = \{S_i : i \in X\}$ is a set cover solution for (U, S) and $|C| = |X| \leq |D|$.

Example: From Set Cover $\Rightarrow$ Dominating Set

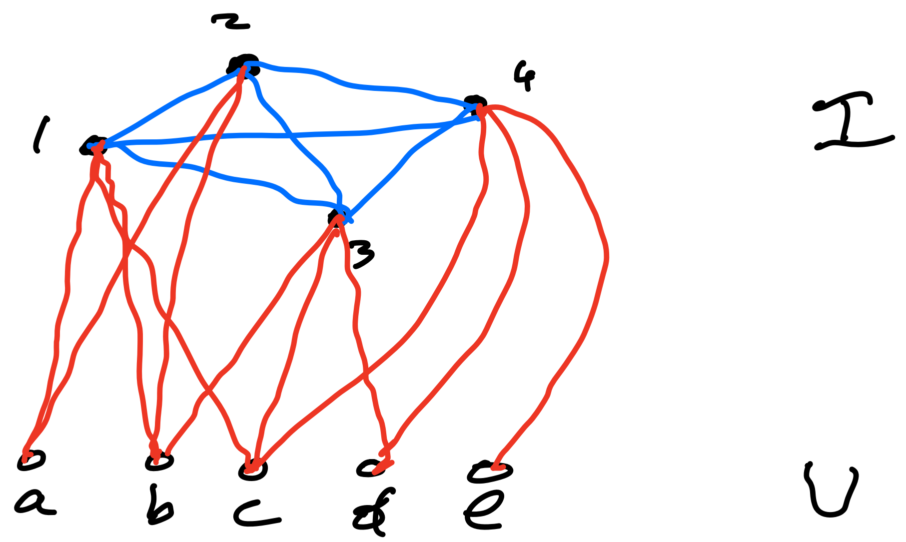
$U = \{a, b, c, d, e\}$

$S_1 = \{a, b, c\}$, $S_2 = \{a, b\}$,

$S_3 = \{b, c, d\}$, $S_4 = \{c, d, e\}$

Construct G: $V = U \cup I$

$I = \{1, 2, 3, 4\}$



In this example:

$C = \{S_1, S_4\}$ is a set cover

correspondin $D = \{1, 4\}$ it is a dominating set

· $D = \{a, 3, 4\}$ is also a dominating set

X constructed as ann

$X = \{1, 3, 4\}$ a dominating set

$\subseteq I$.