

CS299 INNOVATION LAB

Abhinav Dutta 1901CS02

Guide: Prof. Raju Halder



ANALYSIS OF PROGRAMS/TEXT

+

•

○

+

○

•

Introduction

- +
 - In this project we will build methods to process the textual content of programs/documents to find some desirable properties. Our initial target is to detect similarity between two documents. Existing software does it by masking all variables with a special character. We improve upon this by allowing the user to define how they want to mask the variables. We , then realized that by defining a relationship between the various masking categories, other security - related issues can also be solved.

Motivation

Current plagiarism detection tools replace all variables by a special character. This is not very flexible since the user might want to treat different variable differently based data type or his own preference.

Current plagiarism detection tools don't allow for cost-accuracy trade-offs. So it becomes infeasible to work with large documents or large collection of documents. Although it is straightforward to allow the users to select the window size while working with large files.

Related Work

Winnowing: Local Algorithms for Document Fingerprinting (Saul Schleimer et.al 2003) Introduces the idea of minimizers for document fingerprinting which is used in MOSS

Weighted minimizer sampling improves long read mapping (Chirag Jain et.al 2018) Introduces the idea of frequency weighted minimizers which is used to build the tool Winnowmap (used for gene sequence analysis).

Contributions

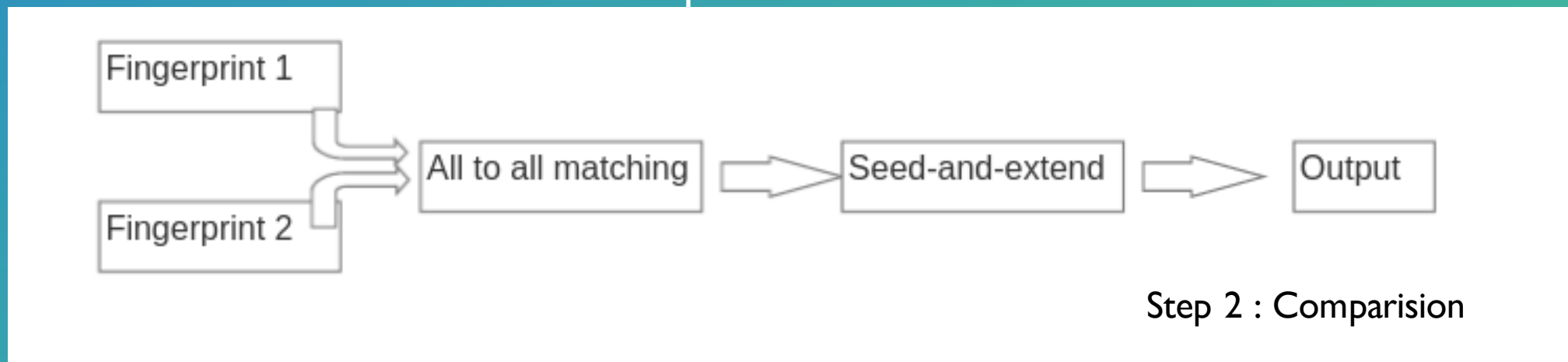
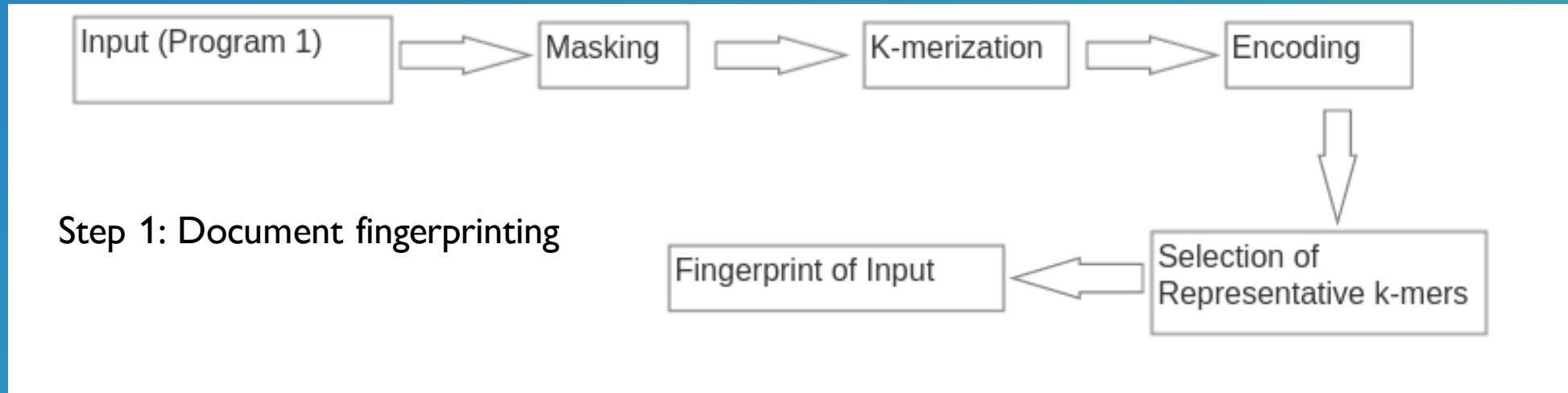
Implemented Colour encoded matchings

Devised a novel masking scheme which can be used to detect vulnerabilities that are not otherwise detected by compilers (for eg. if public variables can store the value of private variable it is not detected but in reality this is a vulnerability). Also such masking schemes improves plagiarism detection.

Implemented frequency weighted minimizers for plagiarism tool which has not been done used in MOSS and other plagiarism tools based on k-mers (preferential selection of rare k-mers)

The winnowing paper and also github implementations of plagiarism tools based on minimizers has complexity $O(n*w)$ (where n is number of characters and w is window size) . However this can be improved to $O(n)$ as suggested in the Winnowmap paper (2018) using monotone queues and this is what I have used in this project.

WORKFLOW



Overview of Similarity detection

- We tokenize the input (text/masked code) into k-mers. And we select some of the k-mers as representative of the sequence. This step requires 2 parameters k and w . k controls the selectivity-specificity trade off whereas w primarily controls the cost-accuracy trade off. (Out of w adjacent k-mers at least 1 k-mer is always selected). There are 2 schemes used to select the representative k-mers :
 - + ● ○
- Minimizers (S.Schleimer, 2003) - Current similarity detectors (MOSS) like use this.
- Syncmers (R. Edgar , 2021) - This is another scheme which selects more evenly spaced k-mers as representatives. Has not yet been used in plagiarism tools.

In my project I have mainly focused on minimizers but can be easily extended to syncmers

- Step 1: k-mers are selected according to the schemes (Minimizer/frequency weighted minimizer).
- Step 2: The selected k-mers from two files are compared and the similarity metrics (Jaccard index) are determined and the similar text is highlighted.

Algorithms

- **Masking algorithm**

The user associates a category with every variable. The user can define a poset on the categories. In this case we have defined the relation based on sensitivity of data.

If $\text{var1} = f(\text{var2}, \text{var3}, \dots, \text{var}_k)$ then $\text{category}(\text{var1}) = \text{lowest upper bound}(\text{category}(\text{var1}, \text{var2}, \dots, \text{var}_k))$.

For. eg if the value stored by private var1 (say password) is more sensitive than the value stored by public var2 (say name) then $\text{category of var1} \geq \text{category of var2}$. Relations maybe be assigned based on data type as well like integers and floats maybe masked separately such that $\text{category}(\text{float}) > \text{category}(\text{integer})$ in the poset (since a float can store a int but not vice versa)

In case the poset does not form a lattice, instead of lowest upper bound, I have considered any upper bound randomly. We formulate the problem of finding $\text{Join}(a, b)$ following as a graph based question : Finding the lowest common ancestor of a and b in the poset graph. (Note : Normal lowest common ancestor algorithm does not work here since the graph is not a tree – each vertex has multiple parents)

Pseudo Code :

Input - the poset as a graph with directed edges stored in adjacency list `adj[]`, source vertex

FindJoin(vertex a, vertex b)

 Mark all the ancestors of a with blue

 Mark all the ancestors of b with black

 DFS to find the height of all vertices

 V = Choose the vertex of greatest height that is both blue and black

 Return V

Algorithm 1: Standard procedure for computing minimizers

Input: $w, arr[]$ //array of k -mers in the order as they occur in seq.
Output: $M[]$ //array of minimizers

```

1 Array  $M \leftarrow []$ ; Deque  $Q \leftarrow []$ ;
2 for  $i \leftarrow 1$  to  $arr.size()$ 
3   //each  $k$ -mer is represented by a pair of its position and order
4    $\langle pos, ord \rangle \leftarrow (i, getOrder(arr[i]))$ ;
5   while  $\neg Q.empty()$  and  $Q.back().second > ord$ 
6      $Q.pop\_back()$ ;
7    $Q.push\_back(\langle pos, ord \rangle)$ ;
8   if  $Q.front().first \leq i - w$  then
9     while  $Q.front().first \leq i - w$ 
10        $Q.pop\_front()$ ; // discard out-of-range  $k$ -mer
11     furtherPop( $Q$ );
12   sample( $Q.front(), M$ );
13   furtherSample( $Q, M$ ); // process ties
14 Function sample( $min, M$ )
15   if  $M.back() \neq min$  then
16      $M.push\_back(min)$ ; // add to index
17 Function getOrder( $mer$ )
18   return  $h(mer)$ ; // hash fn
19 Function furtherPop( $Q$ )
20   return; // do nothing
21 Function furtherSample( $Q, M$ )
22   while  $Q.front().second = Q.next\_to\_front().second$ 
23      $Q.pop\_front()$ ;
24   sample( $Q.front(), M$ );

```

Seed-and-Extend Pseudo Code:**Input** - location of common k -mers in file 1 and 2**Output** - pair<int,int> match1, match2 -which contains the regions which have matched**Seed-and-Extend(pos1,pos2)**

```

i=pos1, j=pos2
while(file[i]==file[j])
  i--; j--;
left1=i; left2=j;

```

```

i=pos1, j=pos2;
while(file[i]==file[j])
  i++; j++;
right1=i; right2=j;

```

Return <left1,right1> , <left2,right2>

Credit : Chirag Jain
 Winnowmap paper

Experiments (similarity detection)

She feels much warmer once they are out in the car park,
The girl is so small her feet dangle off the folding plastic bench

The girl is so small her feet dangle off the folding plastic bench. She stares at her grandparents' names and oval photos in which they don't smile. She feels much warmer once they are out in the car park, flat and not like a frying pan, sizzling the metal boxes in the afternoon sun. Mama says you can be so morbee, the girl remembers out loud, fishing a lollipop out of the glove compartment, and her father laughs. If trees are forever, the girl reasons, then surely, they ought to plant one so that it remembers them, so that it's grateful for the gift of existence, so that it whispers of them when they're gone.

Jaccard value= 0.444444
Cosine similarity= 1.54222

Tearing the page away and crumpling it into a ball, he threw it over his shoulder to join its growing mass of brethren on the floor. It was dark in the office.
He licked his dry lips, ignoring his thirst and the beginnings of a headache. When last had he drunk anything? No, that wasn't important. This was.
He turned, almost expecting her to be standing behind him, but there were only the shadows and his books. Those books that had taken so many hours.
He got up, reaching for the first book on the shelf, his first novel. The snarling face of an undead childrella stared up at him.
He tore the page out and then the next and the next...and the next. One by one, they fell to the floor like snow until his movements. Book after book, was torn apart, leaving him vent his rage in that dark room. He finally collapsed on the floor, exhausted to his bones.
The world was bright. Sunlight dabbled down through the leaves of the trees in the forest as he roared out on his white horse. It had been a long and treacherous journey. The princess had been stolen away from them suddenly by a wicked fairy who cast a terrible curse on her.
"Paul? Paul!" the voice was coming to him from far away, slowly dragging him out of the dream, "Honey, wake up!"

He turned, almost expecting her to be standing behind him, but there were only the shadows and his books. Those books that had taken so many hours.
A gilded, glass coffin lay in the center of a clearing. As he approached, he could see her sweet face through the opaque glass, gently drawing his eyes sting.
Her angelic face, framed by the golden halo of her hair remained impassive. He bent down, pressing his lips to her cool forehead and feeling.
The world was bright. Sunlight dabbled down through the leaves of the trees in the forest as he roared out on his white horse. It had been a long and treacherous journey. The princess had been stolen away from them suddenly by a wicked fairy who cast a terrible curse on her.
Tearing the page away and crumpling it into a ball, he threw it over his shoulder to join its growing mass of brethren on the floor. It was dark in the office.
She tried to pull him up, but he gripped her by the arm, gaze searching. She looked so much like the little princess.
He leaped forward, wrapping her arms around him and bringing them close together. He shuddered in her arms, letting himself get lulled in by her warmth. Hot tears found their way down his face as he clutched close to her.

Small k-values (k=4) has many small random matches

Required Density ?1

She sits for hours, watching the stem, trying to catch it in the act of growing. It doesn't look like anything just yet. It could be an iris, it could be a rose bush, it could be an elm. The girl wonders whether the tree knows what it will become.
She stares, but nothing ever happens. It grows constantly, just very slowly, her mother tries to explain, taking the chance to braid her distracted daughter's hair.

She stares, but nothing ever happens. It grows constantly, just very slowly, her mother tries to explain, taking the chance to braid her distracted daughter's hair.
The tree strives for sun, trying to catch light onto its leaves. It needs all the sugar it can get, her father explains, 'but it'll grow crooked if we don't give it a spin.'
She sits for hours, watching the stem, trying to catch it in the act of growing. It doesn't look like anything just yet. It could be an iris, it could be a rose bush, it could be an elm. The girl wonders whether the tree knows what it will become.

Jaccard value= 0.411111

(program exited with code: 0)
Press return to continue

She stares, but nothing ever happens. It grows constantly, just very slowly, her mother tries to explain, taking the chance to braid her distracted daughter's hair.
'That's very sneaky, the girl nods, her entire being focused on the tree so much that her mother's poor combing technique doesn't make her flinch and run away.
She sits for hours, watching the stem, trying to catch it in the act of growing. It doesn't look like anything just yet. It could be an iris, it could be a rose bush, it could be an elm.

She sits for hours, watching the stem, trying to catch it in the act of growing. It doesn't look like anything just yet. It could be an iris, it could be a rose bush, it could be an elm.
She stares, but nothing ever happens. It grows constantly, just very slowly, her mother tries to explain, taking the chance to braid her distracted daughter's hair.

Larger k values (k=8) gives better matches

The Cost-Accuracy Tradeoff

Required Density ? 1

They⁰⁰⁰⁰re visiting the quiet grandparents when her father says, ⁰⁰⁰⁰One day, you⁰⁰⁰⁰ll find me here, too. Nothing is forever.⁰⁰⁰⁰ His voice sounds like it⁰⁰⁰⁰s coming from under the ground, like he⁰⁰⁰⁰s already training her for what⁰⁰⁰⁰s to come.

The girl is so small her feet dangle off the folding plastic bench. She stares at her grandparents' names and oval photos in which they don't smile, and then at the oak casting a shadow on their granite home. 'What about that?' she points at the tree.

She feels much warmer once they⁰⁰⁰⁰re out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

They⁰⁰⁰⁰re visiting the quiet grandparents when her father says, ⁰⁰⁰⁰One day, you⁰⁰⁰⁰ll find me here, too. Nothing is forever.⁰⁰⁰⁰ His voice sounds like it⁰⁰⁰⁰s coming from under the ground, like he⁰⁰⁰⁰s already training her for what⁰⁰⁰⁰s to come.

If trees are forever, the girl reasons, then surely, they ought to plant one so that it remembers them, so that it⁰⁰⁰⁰s grateful for the gift of existence, so that it whispers of them when they⁰⁰⁰⁰re gone.

She feels much warmer once they⁰⁰⁰⁰re out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

Jaccard value= 0.54895

Required Density ? 0.1

They⁰⁰⁰⁰re visiting the quiet grandparents when her father says, ⁰⁰⁰⁰One day, you⁰⁰⁰⁰ll find me here, too. Nothing is forever.⁰⁰⁰⁰ His voice sounds like it⁰⁰⁰⁰s coming from under the ground, like he⁰⁰⁰⁰s already training her for what⁰⁰⁰⁰s to come.

The girl is so small her feet dangle off the folding plastic bench. She stares at her grandparents' names and oval photos in which they don't smile, and then at the oak casting a shadow on their granite home. 'What about that?' she points at the tree.

She feels much warmer once they⁰⁰⁰⁰re out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

They⁰⁰⁰⁰re visiting the quiet grandparents when her father says, ⁰⁰⁰⁰One day, you⁰⁰⁰⁰ll find me here, too. Nothing is forever.⁰⁰⁰⁰ His voice sounds like it⁰⁰⁰⁰s coming from under the ground, like he⁰⁰⁰⁰s already training her for what⁰⁰⁰⁰s to come.

If trees are forever, the girl reasons, then surely, they ought to plant one so that it remembers them, so that it⁰⁰⁰⁰s grateful for the gift of existence, so that it whispers of them when they're gone.

She feels much warmer once they⁰⁰⁰⁰re out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

Jaccard value= 0.506173

Required Density ? 0.01

They⁰⁰⁰⁰re visiting the quiet grandparents when her father says, ⁰⁰⁰⁰One day, you⁰⁰⁰⁰ll find me here, too. Nothing is forever.⁰⁰⁰⁰ His voice sounds like it⁰⁰⁰⁰s coming from under the ground, like he⁰⁰⁰⁰s already training her for what⁰⁰⁰⁰s to come.

The girl is so small her feet dangle off the folding plastic bench. She stares at her grandparents' names and oval photos in which they don't smile, and then at the oak casting a shadow on their granite home. 'What about that?' she points at the tree.

She feels much warmer once they⁰⁰⁰⁰re out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

They⁰⁰⁰⁰re visiting the quiet grandparents when her father says, ⁰⁰⁰⁰One day, you⁰⁰⁰⁰ll find me here, too. Nothing is forever.⁰⁰⁰⁰ His voice sounds like it⁰⁰⁰⁰s coming from under the ground, like he⁰⁰⁰⁰s already training her for what⁰⁰⁰⁰s to come.

If trees are forever, the girl reasons, then surely, they ought to plant one so that it remembers them, so that it's grateful for the gift of existence, so that it whispers of them when they're gone.

She feels much warmer once they⁰⁰⁰⁰re out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

Jaccard value= 0.5

Required Density ? 0.005

They're visiting the quiet grandparents when her father says, 'One day, you'll find me here, too. Nothing is forever.' His voice sounds like it's coming from under the ground, like he's already training her for what's to come.

The girl is so small her feet dangle off the folding plastic bench. She stares at her grandparents' names and oval photos in which they don't smile, and then at the oak casting a shadow on their granite home. 'What about that?' she points at the tree.

She feels much warmer once they're out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

They're visiting the quiet grandparents when her father says, 'One day, you'll find me here, too. Nothing is forever.' His voice sounds like it's coming from under the ground, like he's already training her for what's to come.

If trees are forever, the girl reasons, then surely, they ought to plant one so that it remembers them, so that it's grateful for the gift of existence, so that it whispers of them when they're gone.

She feels much warmer once they're out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

Jaccard value= 0.333333

Required Density ? 0.0001

They're visiting the quiet grandparents when her father says, 'One day, you'll find me here, too. Nothing is forever.' His voice sounds like it's coming from under the ground, like he's already training her for what's to come.

The girl is so small her feet dangle off the folding plastic bench. She stares at her grandparents' names and oval photos in which they don't smile, and then at the oak casting a shadow on their granite home. 'What about that?' she points at the tree.

She feels much warmer once they're out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

They're visiting the quiet grandparents when her father says, 'One day, you'll find me here, too. Nothing is forever.' His voice sounds like it's coming from under the ground, like he's already training her for what's to come.

If trees are forever, the girl reasons, then surely, they ought to plant one so that it remembers them, so that it's grateful for the gift of existence, so that it whispers of them when they're gone.

She feels much warmer once they're out in the car park, flat and hot like a frying pan, sizzling the metal boxes in the afternoon sun. A wave of heat so potent it distorts vision hits her when she opens the passenger door.

Jaccard value= -nan

This demonstrates that as density decreases the smaller matches are undetected and finally all matches are lost Density values are (1, 0.1, 0.01, 0.005, 0.001 from top -bottom left-right). User has to choose the correct density as per his needs

Experiments (Masking)

```
int main()
{
alpha = beta + gamma + delta; /* add some numbers */
if(alpha>=beta)                /* do some comparison */
gamma = beta;
if(delta== gamma)
alpha = delta * 50;
}
```

```
int main()
{
R1 = R2 + R3 + R4;
if(R1>=R2)
R3 = R2;
if(R4== R3)R1 = R4 * INTEGER;
}
```

```
5
4
1 2
1 3
2 4
2 5
1
4
alpha 1
beta 2
gamma 3
delta 4
```

```
Required Density ?0.1
int main()
{
R1 = R2 + R3 + R4;
if(R1>=R2)
R3 = R2;
if(R4== R3)R1 = R4 * INTEGER;
}

int main()
{
R1 = R2 + R3 + R4;
if(R1>=R2)
R3 = R2;
if(R4== R3)R1 = R4 * INTEGER;
}

Jaccard value= 0.875
```

Improvements/future work

- Building it as a web service (underway)
- Using syncmers in place of minimizers as syncmers are shown to have better spacing distribution compared to minimizers.
- Improve the colour highlighting scheme to deal with overlaps in a better way.
- Currently the hasse diagram is directly taken from user however if a program uses many variables (say 100) it is difficult for the user to input the poset graph with 100 vertices