pink pony

Developed by
Pavlina Mitsou
Junior Software
Engineer

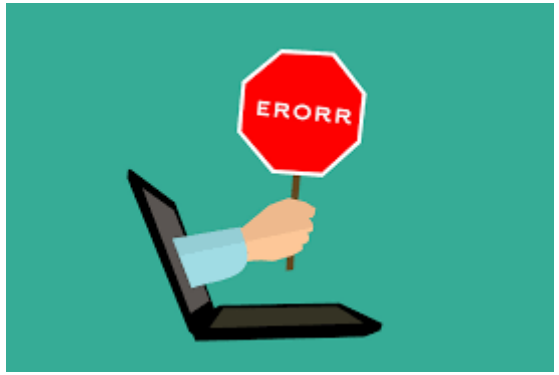
Thesis project: MSR

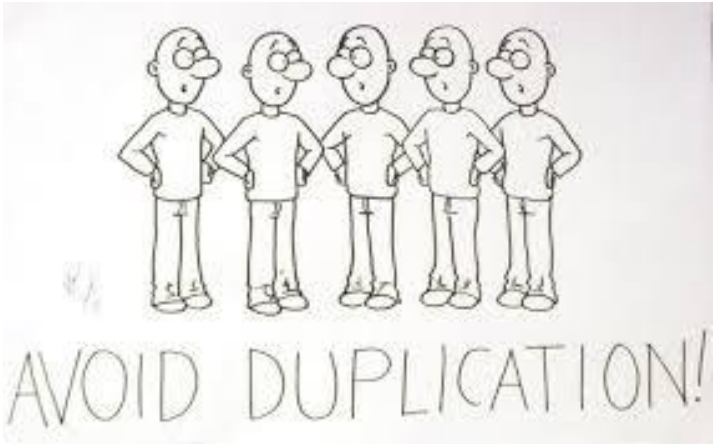
Mining software repositories

The mining software repositories field analyzes the rich data available in software repositories, such as version control repositories, mailing list archives, bug tracking systems, issue tracking systems, etc. to uncover interesting and actionable information about software systems, projects and software engineering. Wikipedia

Source of MSR

Fields of MSR





Git

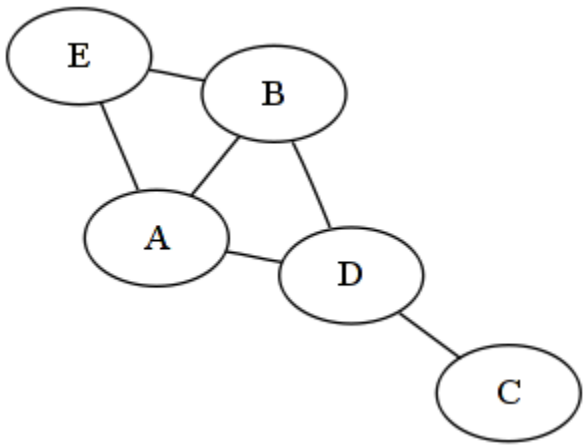
Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

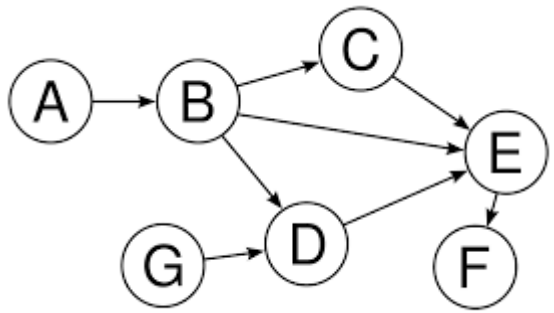
Why Git is important source or MSR?

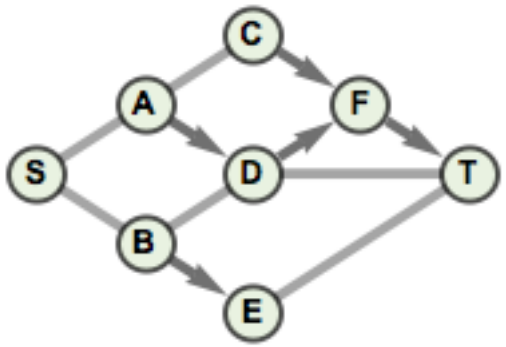
Git can provide to you in questions such as:

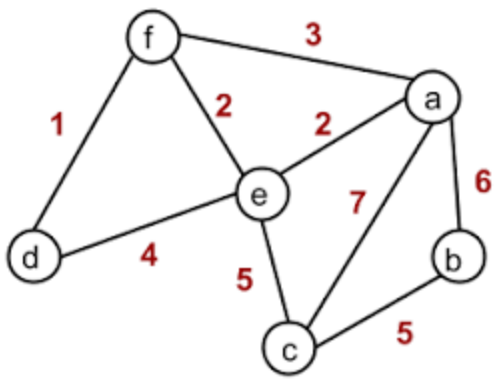
- Which files do change together?
- Who did this change?
- When did this change happen?
- How many contributors are in this project?
- What are the commit messages are?
- How many commits are there in this branch?

Graphs

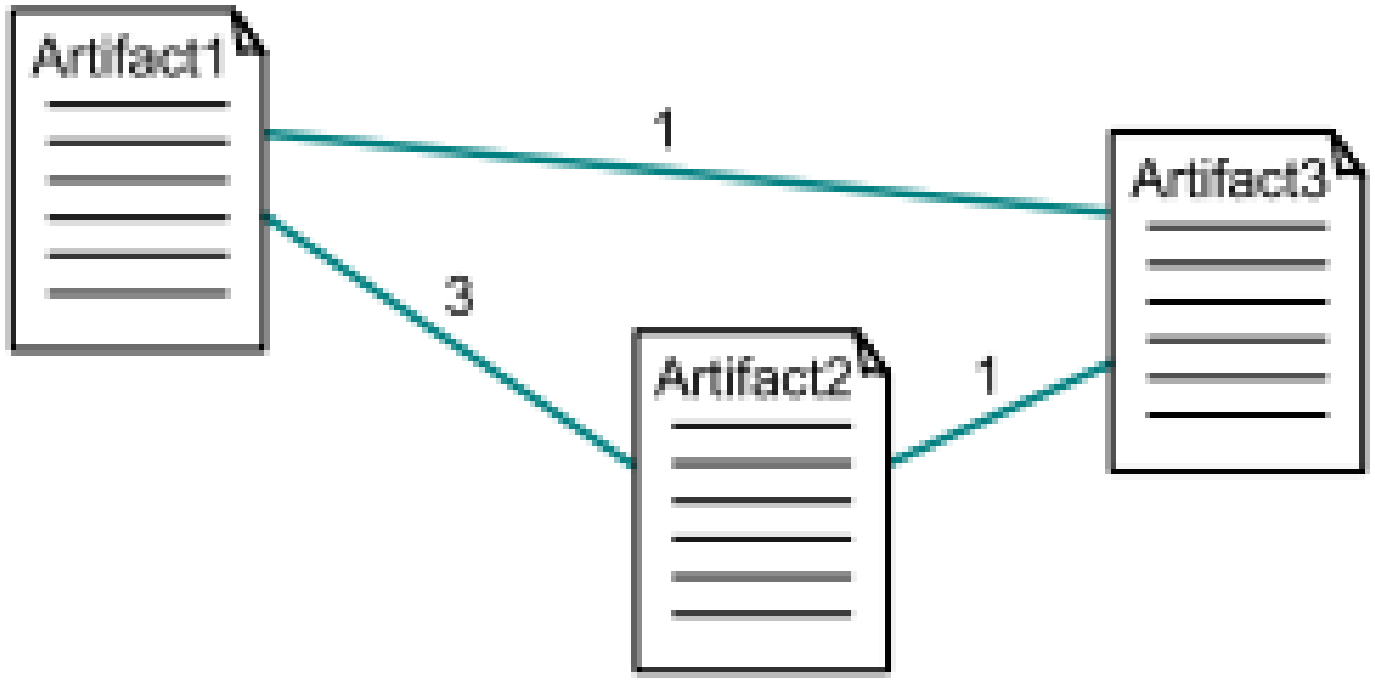




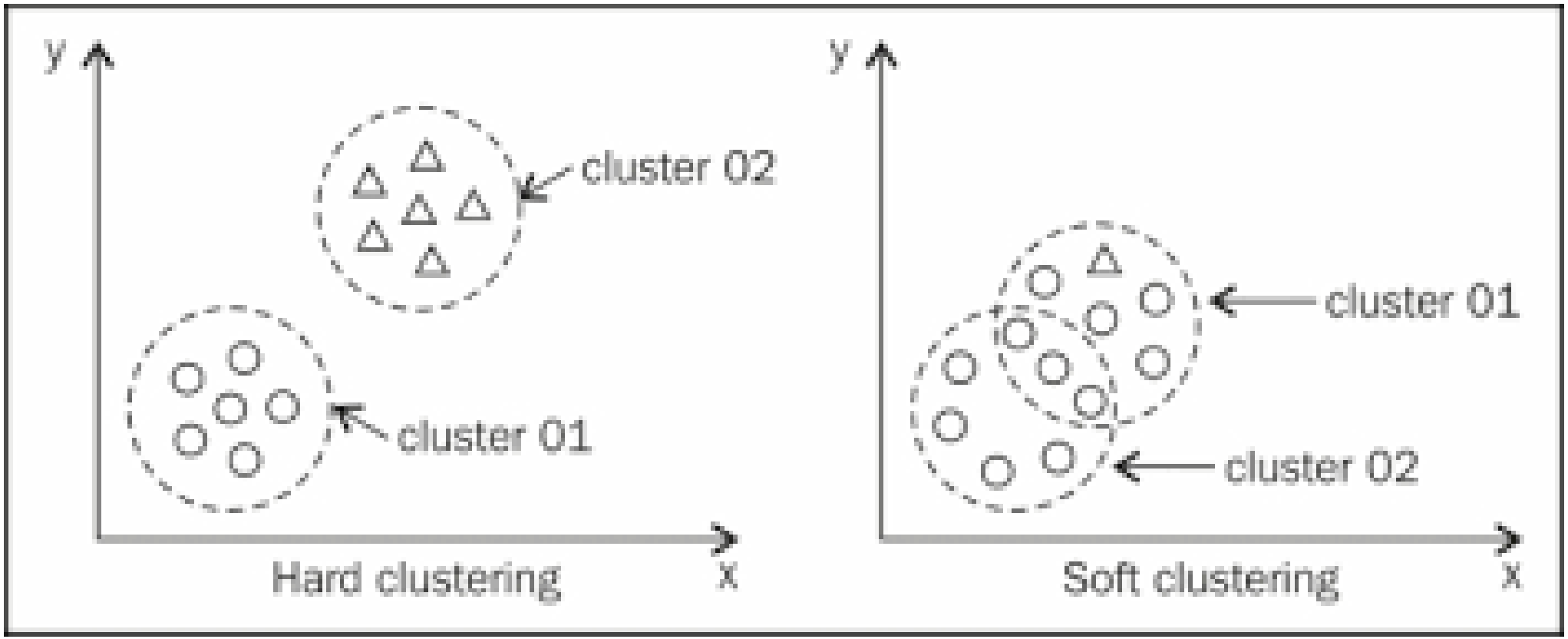




Co-change graph

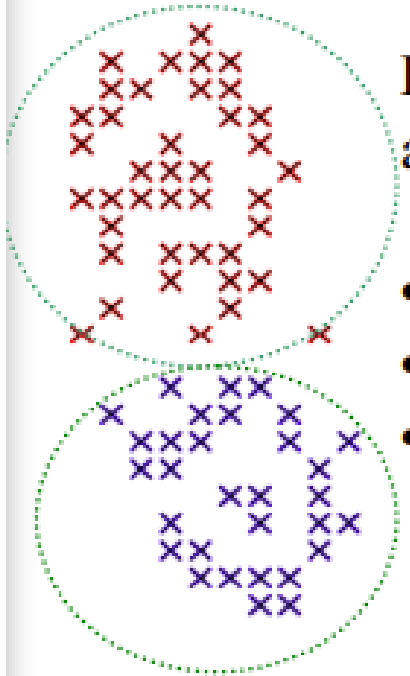


Clustering Types



Graph Clustering

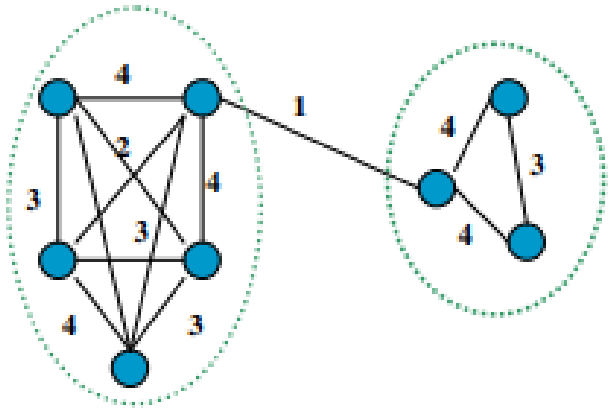
Vector Clustering



Each point has a vector, i.e.

- x coordinate
- y coordinate
- color

Graph Clustering

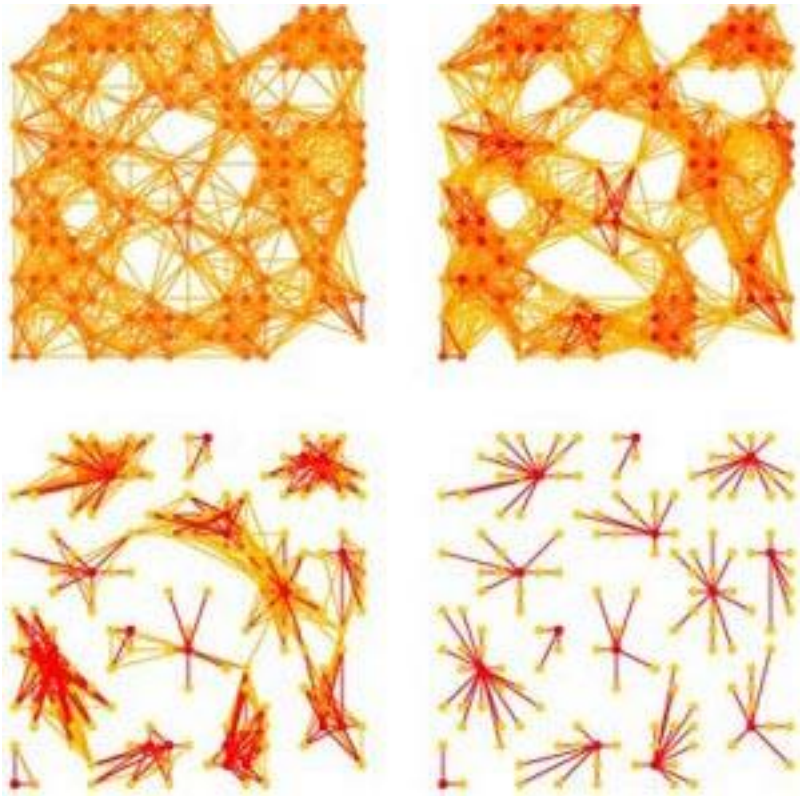


Each vertex is connected to others by (weighted or unweighted) edges.

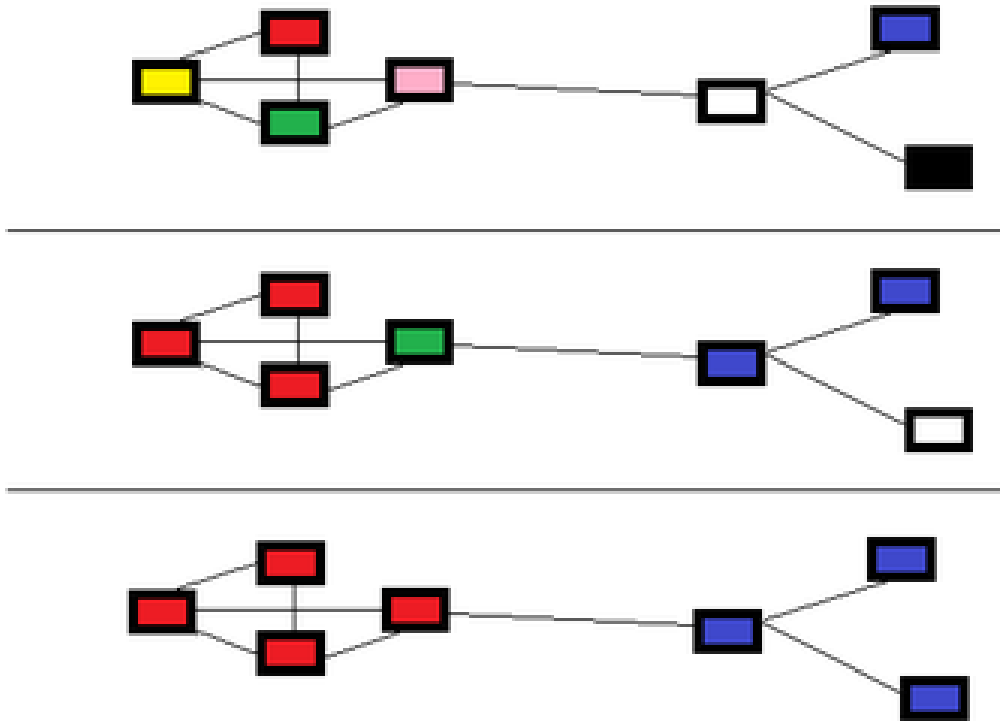
Algorithms

Markov Clustering Algorithm

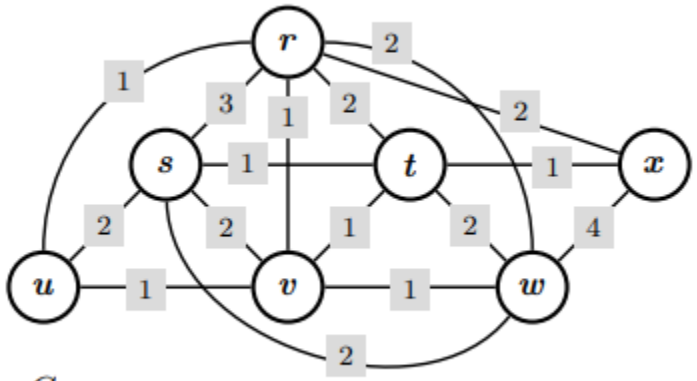
1. Input is an un-directed graph, power parameter *e*, and inflation parameter *r*.
2. Create the associated matrix
3. Add self loops to each node (optional)
4. Normalize the matrix
5. Expand by taking the e^{th} power of the matrix
6. Inflate by taking inflation of the resulting matrix with parameter *r*
7. Repeat steps 5 and 6 until a steady state is reached (convergence).
8. Interpret resulting matrix to discover clusters.



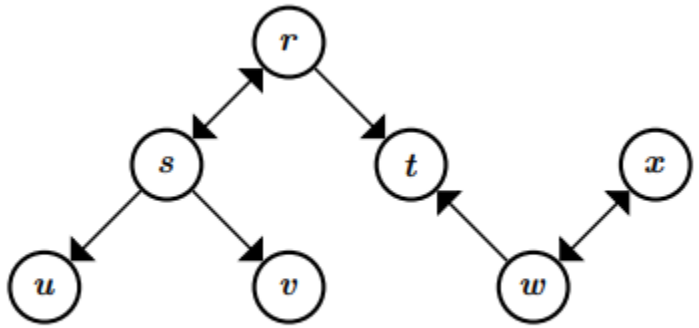
Chinese Whispers



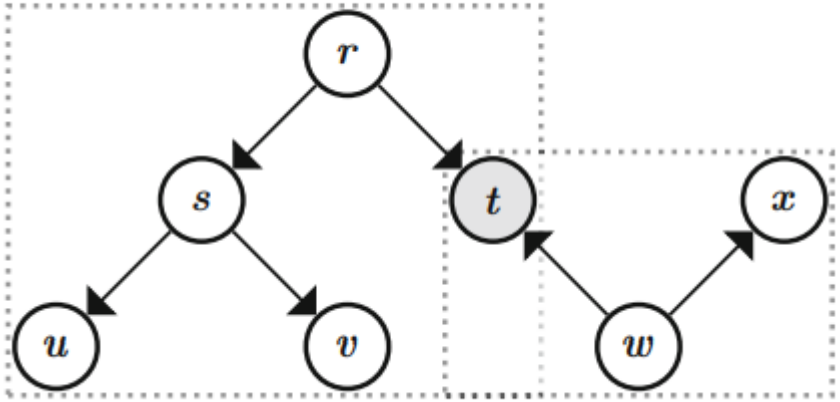
MaxMax



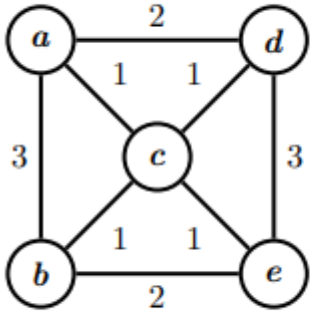
G



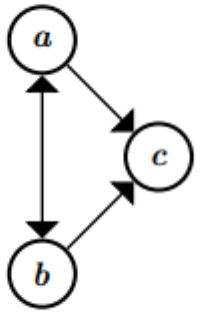
G'



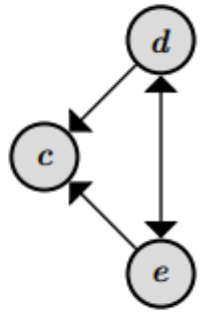
G



C_G

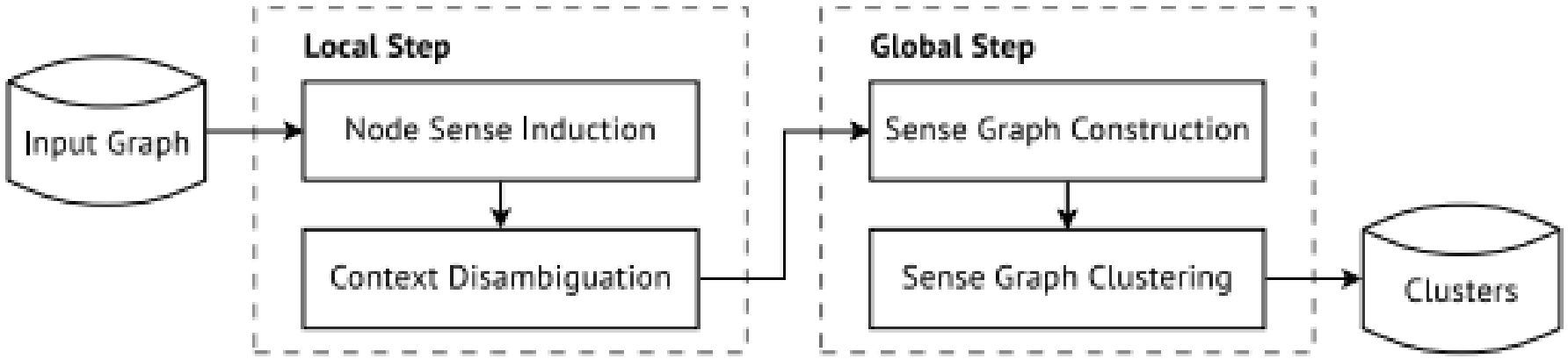


Cluster 1



Cluster 2

Watset


Local Step
Global Step
Input Graph
Node Sense Induction
Context Disambiguation
Sense Graph Construction
Sense Graph Clustering
Clusters

What is pink pony application anyway?

pink pony

is a free software command-line application that suggest functional clusters based on the common changes on git.

https://github.com/Pavlmits/PinkPony

And why I need *pink pony* ?

“Changes of software systems are less expensive and less error-prone if they affect only one subsystem. Thus, clusters of artifacts that are frequently changed together are subsystem candidates.”

How to use the *pink pony* ?

pink pony

https://github.com/Pavlmits/PinkPony

Clustering algorithms options

- `mr` : Markov Clustering is hard clustering algorithm;
- `ch` : Chinese Whispers is a hard clustering algorithm;
- `max` : MaxMax is a soft clustering algorithm for undirected graphs;
- `watset` : Watset is a *local-global meta-algorithm* for fuzzy graph clustering.

The implementation of the algorithms used from Watset project

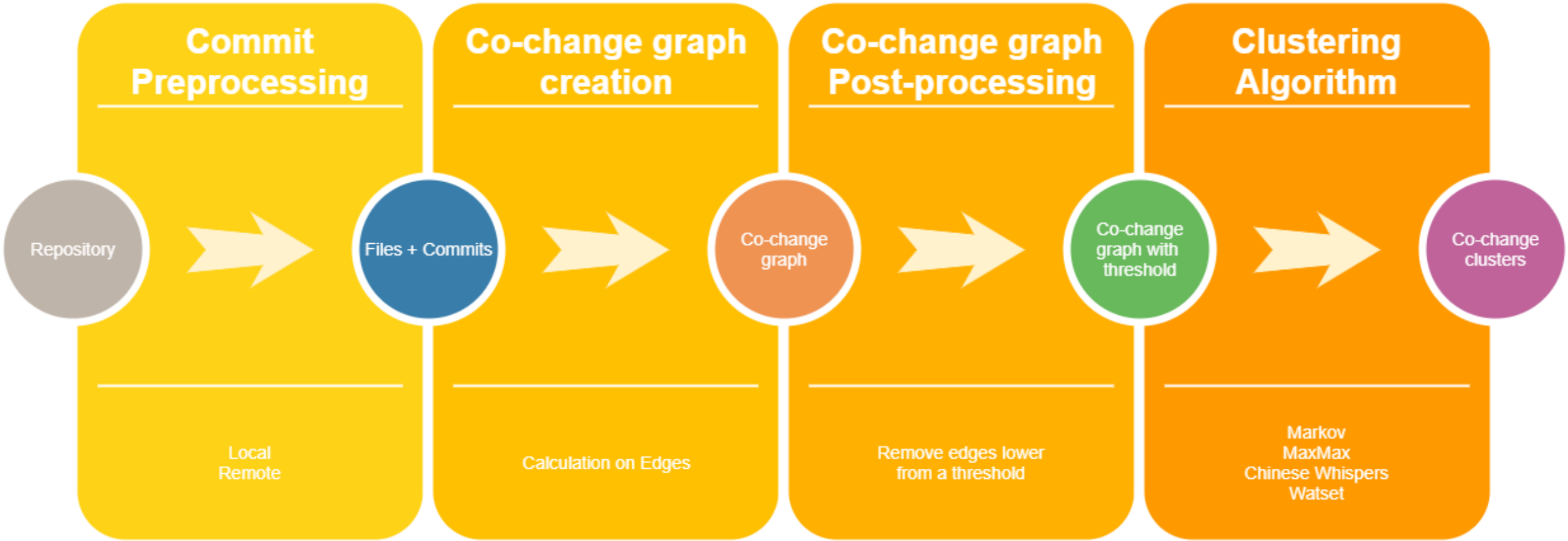
Level of clustering

Depends the investigation that you want to do, the application provides two kind of options.

- `file` : Suggest functional clusters on file level.
- `module` : Suggest functional clusters on module level.
- This will suggest functional clusters from files.

```
$ java -jar pinkpony.jar path\to\.git file max
```

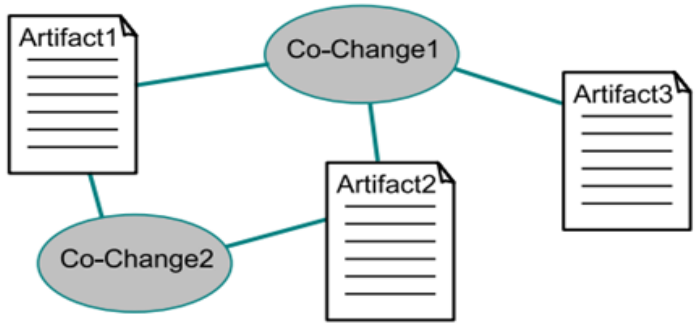
How did I make?


Commit Preprocessing
Co-change graph creation
Co-change graph Post-processing
Clustering Algorithm
Repository
Files + Commits
Co-change graph
Co-change graph with threshold
Co-change clusters
Local
Remote
Calculation on Edges
Remove edges lower from a threshold
Markov
MaxMax
Chinese Whispers
Watset

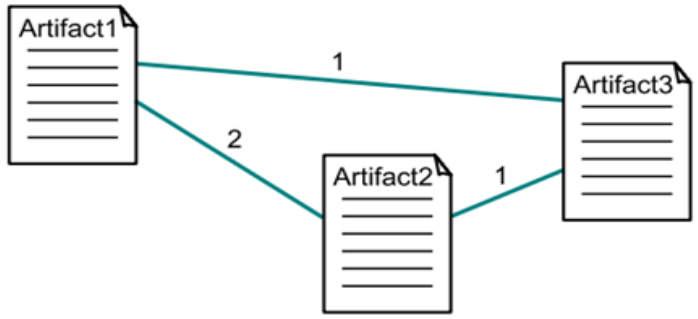
Pre-processing tasks

- Removing commits not associated to maintenance issues
- Removing commits not changing classes
- Merging commits related to the same maintenance issue
- Removing highly scattered commits

Extract Co-change graph



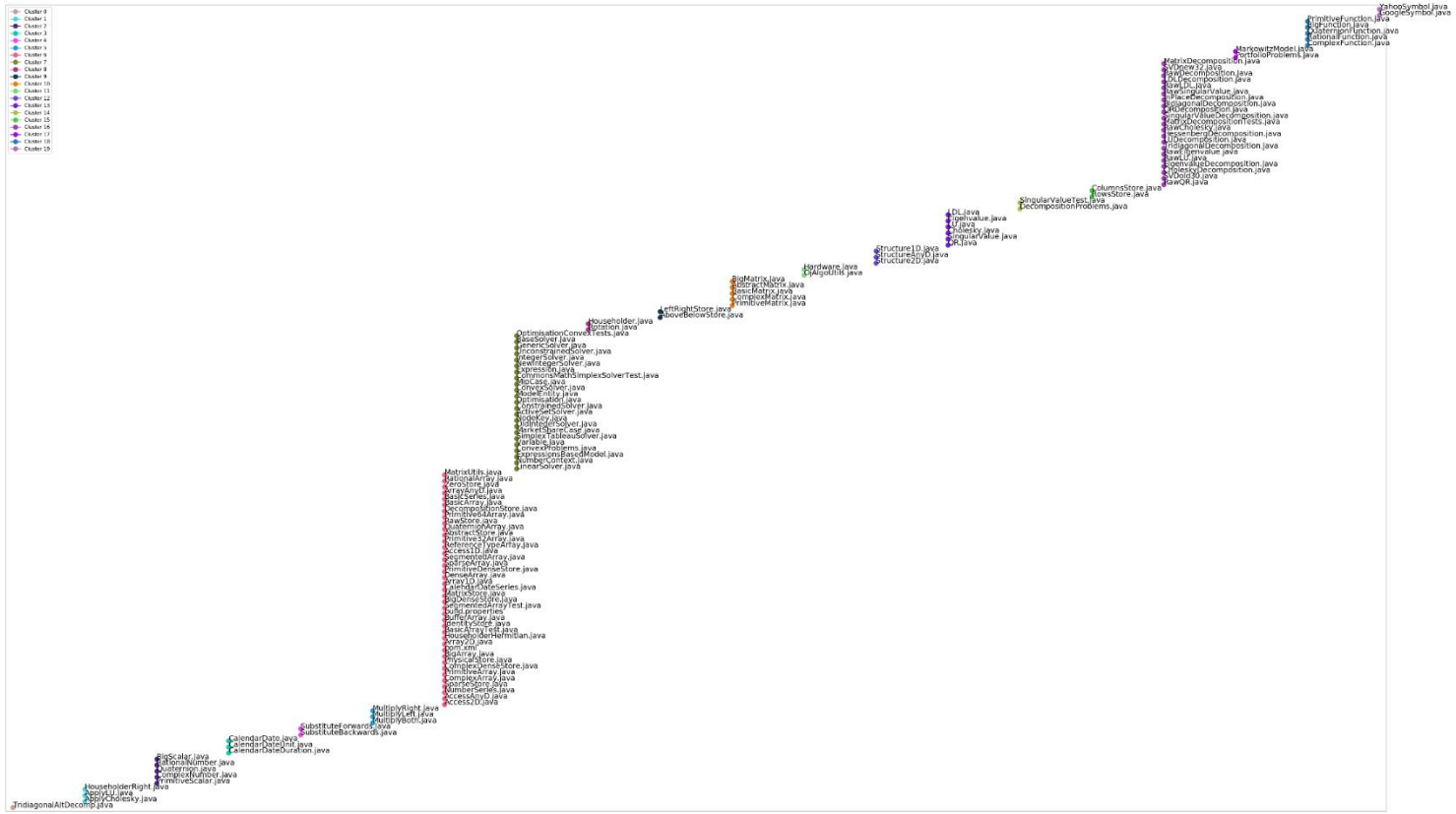
(a) Co-change graph



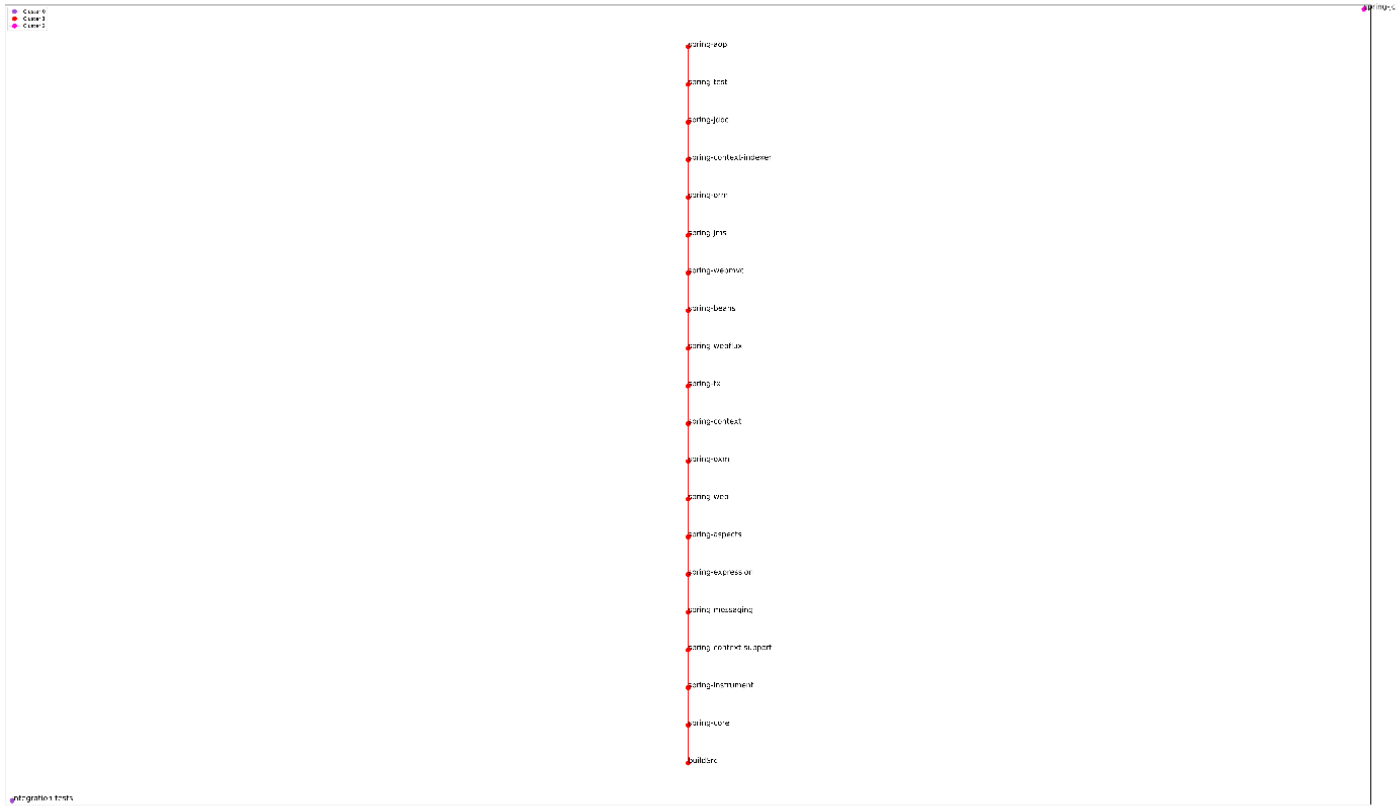
(b) Condensed co-change graph

Experiments

ojAlgo



Spring


spring-aop
spring-test
spring-jdbc
spring-jms
spring-beans
spring-tx
spring-context
spring-core
spring-expression
spring-instrument
spring-aspects
integration-tests

Real experiment

pink pony

Was it so easy?

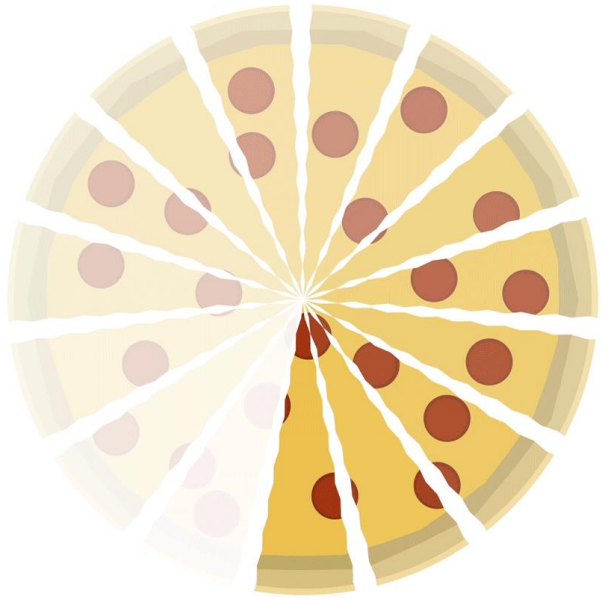
Of course NOT!

Filter data

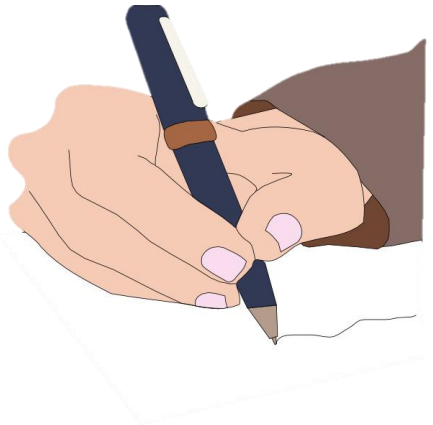
Performance issues

Visualization issues

Evaluation issues



Next steps…



Medium

Pavlina Mitsou

https://medium.com/@pavlinamitsou/the-most-important-lesson-that-i-learned-as-an-intern-4f2f0c265d7a