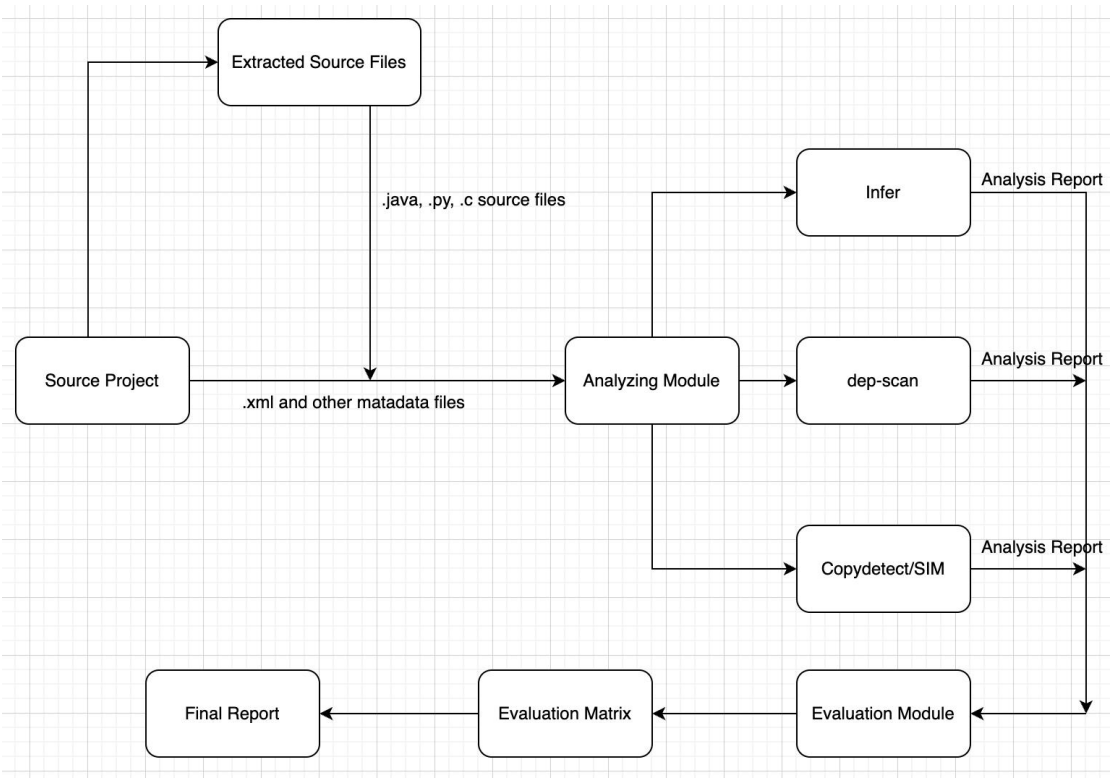
COMSE6156 Project Proposal

Tianzhi Huang (th2888)

Overall:

By revisiting researches done in my midterm paper and searching available workable open source projects, I tend to implement a system that evaluates potential security threats and vulnerabilities. The system is extended based on the analysis results generated from three subsystems that evaluates 1) potential vulnerabilities in external dependencies and Third-Party Libraries included in the project, 2) code errors and vulnerabilities in source files detected by static code analyzing tool, 3) code similarity between source file in the project indicating possible redundant code that does the identical functionality (possibly one is being used and the other is supposed to be deprecated but not, causing the vulnerabilities of inconsistent code). By getting those scanning results from those subsystems, the system will scan through those results, trying to get rid of any obvious false positive cases and then calculating a overall score for each source file based on how many dependencies it includes or imports are vulnerable, how many static code errors or vulnerabilities it has, and how many similar source files (maybe above a threshold) it has. Eventually, based on the scores of those source files, build a model, calculate and output an overall evaluation matrix representing the vulnerability level of the whole project.

Tentative Project Structure Flow:



Tentative Reused Software:

1. Infer: https://github.com/facebook/infer
2. dep-scan: https://github.com/AppThreat/dep-scan
3. Copydetect: https://github.com/blingenf/copydetect
4. SIM: https://github.com/andre-wojtowicz/sim
5. DeepSIM: https://github.com/parasol-aser/deepsim

Tentative Evaluation/Experiment Process:

In order to evaluate the system's performance, I an planning to find two open source projects developed in Java, Python, and one other language as the test input projects. For the sake of accessibility of the code, I may choose some projects that my team or I developed in other classes (including COMS4156). Those two projects in the same language tend to be different in scale and complexity. Then I will run my system on those projects and output the evaluation matrix for each of them. I have not decided what measurements should be included in the final evaluation matrix yet since it depends on what I could have access to and the complexity of retrieving that information. I will probably test those projects using other contemporary and widely used tools in order to better assess the performance of my system. The stats will be included in the final project report.