

# Upholding Ethics & Integrity:

A macro-based approach to detect plagiarism in programming

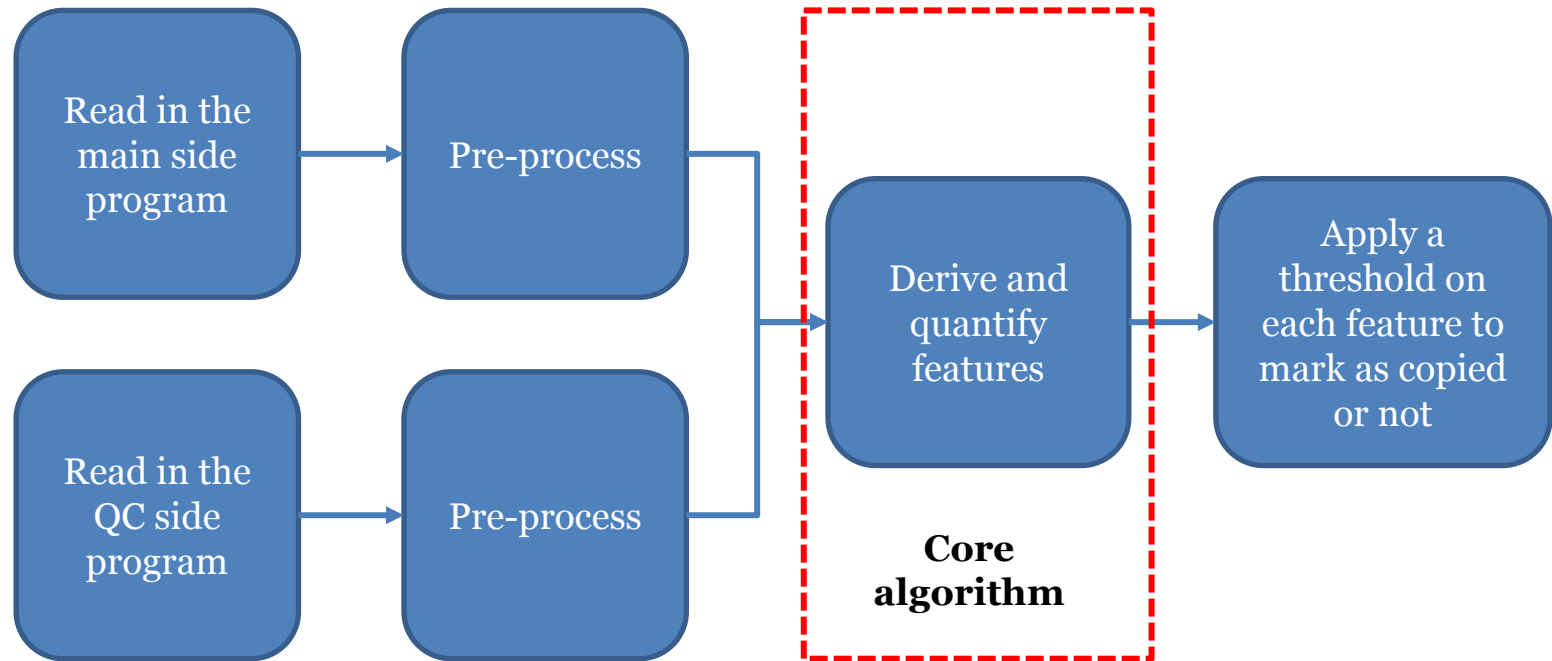
*Sridhar Vijendra & Praveen Kumar*

# Plagiarism in Programming

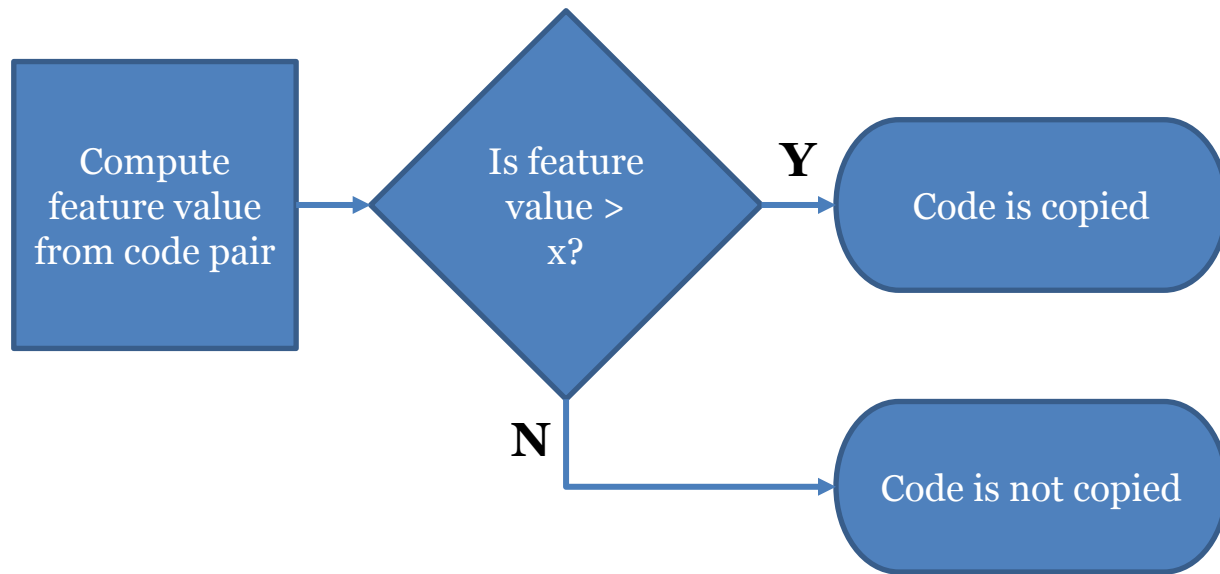
# Identification & Prevention

- *diff* tools
- Finding similarities vs. finding differences
- MOSS (Measure of software similarity)

# Suggested Method



# Use of string features



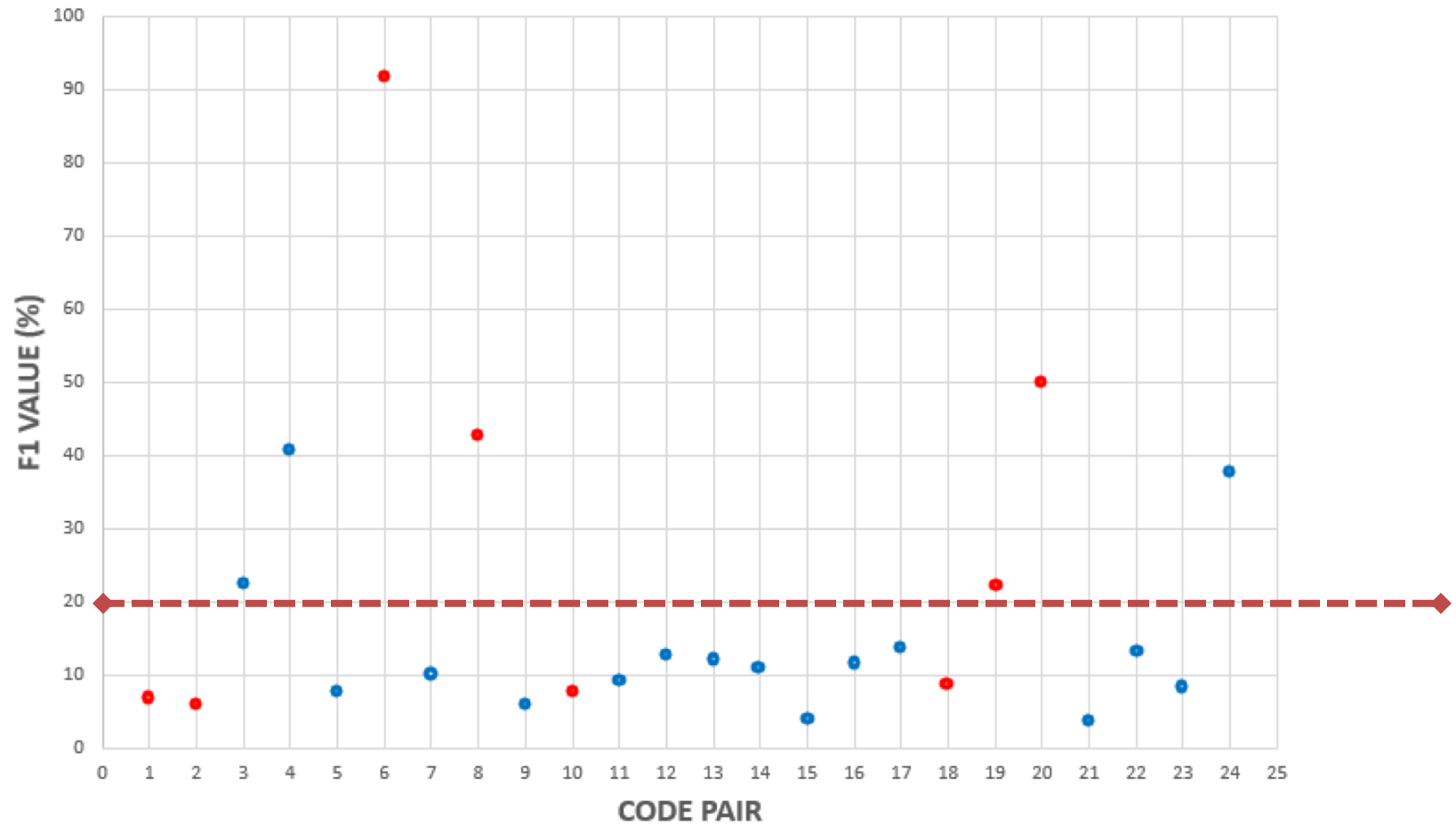
# Test Data

|                 | Identification Set | Test Set  |
|-----------------|--------------------|-----------|
| Plagiarised     | 8                  | 4         |
| Non-plagiarised | 16                 | 8         |
| <b>Total</b>    | <b>24</b>          | <b>12</b> |

# Feature 1

$$\text{F1} = \frac{\text{Number of lines of exactly matching codes}}{\text{Maximum number of lines in both codes}} \times 100$$

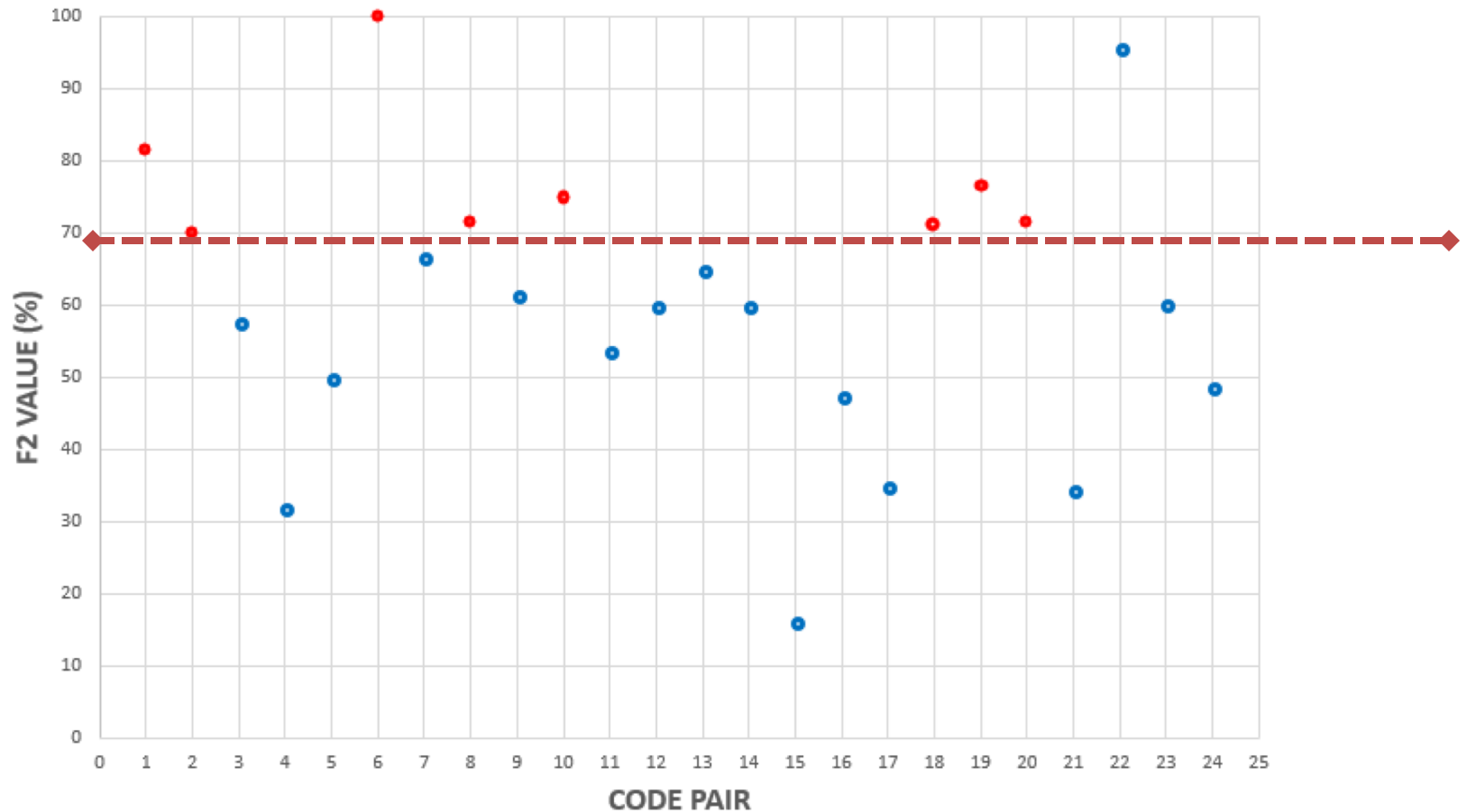
# Feature 1 – Scatter plot (24 pairs)



## Feature 2

$$\text{F2} = \frac{\text{Number of PROC and DATA steps in main side}}{\text{Number of PROC and DATA steps in QC side}} \times 100$$

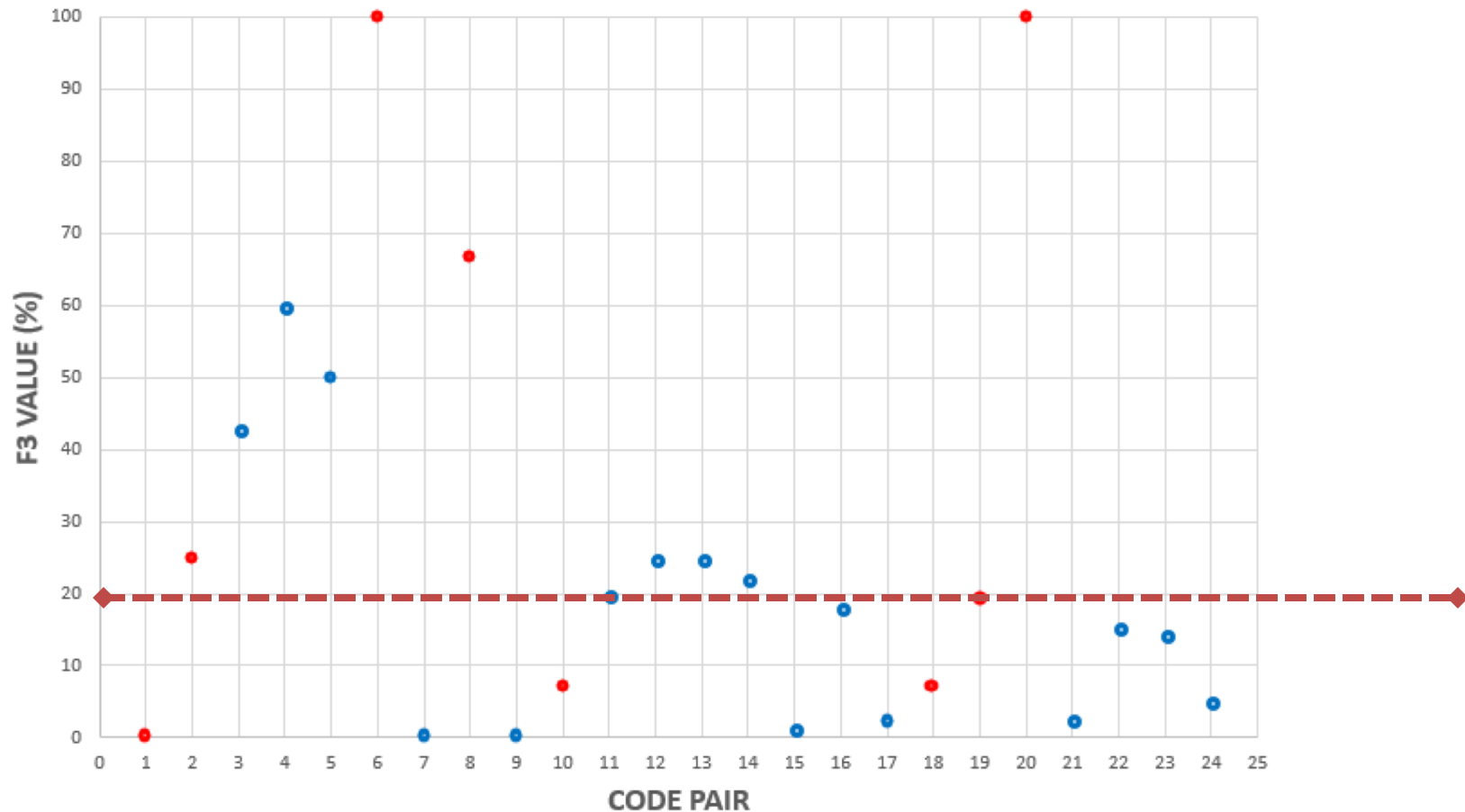
## Feature 2 – Scatter plot (24 pairs)



# Feature 3

$$\mathbf{F3} = \frac{\text{Number of common intermediate data set names}}{\text{Maximum number of intermediate data sets in both codes}} \times 100$$

## Feature 3 – Scatter plot (24 pairs)



# Results with Test Data – Feature 2 stands out

|                                   | Code copied<br>(as per Manual check) | Code Not Copied<br>(as per Manual Check) |
|-----------------------------------|--------------------------------------|------------------------------------------|
| Code Copied<br>(as per macro)     | 3                                    | 0                                        |
| Code Not Copied<br>(as per macro) | 1                                    | 8                                        |
| Total                             | 4                                    | 8                                        |

# Summary

- **Small test set but promising results**
- **More complex algorithms to be tried**
- **Promising tool for screening of codes**

**Thank you**