LERO Software for a better world

UNIVERSITY OF LIMERICK OLLSCOIL LUIMNIGH

A World Leading SFI Research Centre

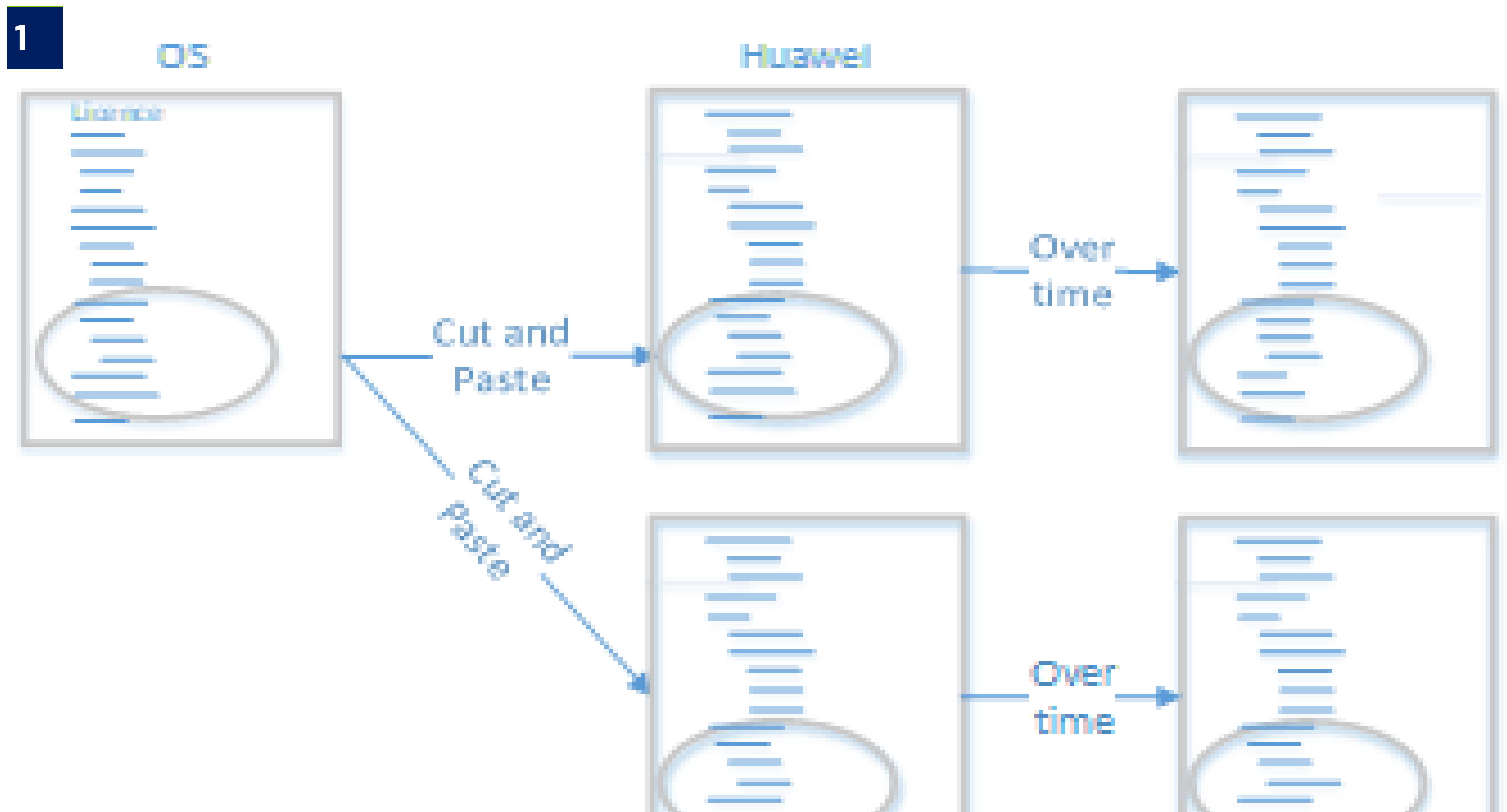
Science Foundation Ireland For what's next

GLOCC: Giga-line Location of Code Clones

Muslim Chochlov, Gul Ahmed, James Patten, Jim Buckley, David Gregg

HUAWEI

1



SOFTWARE CLONES:

- Pieces of code that are similar (syntactically or semantically)…
- May be a result of a cut and paste

THE PROBLEM:

- They can cause copyright issues;
- They can cause inherited maintenance issues;
- They can be difficult to find, given that evolve independently (diverge) over time;

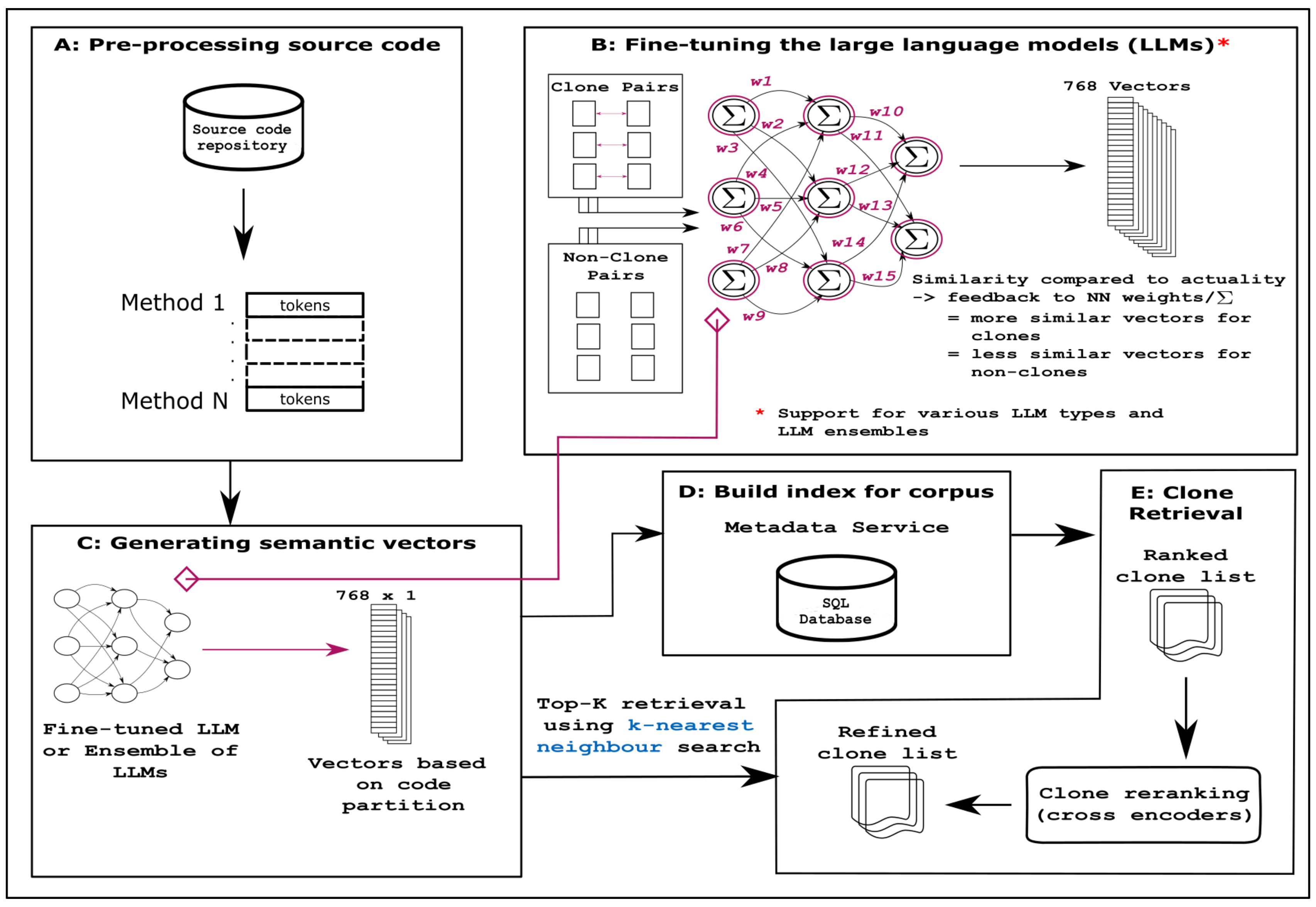
THE SOLUTION:

- We need to detect these diverging clones, at 1000MLOC+ scale

2

- Approaches for finding Type I/II clones are already quite accurate and quite efficient;
- State-of-the-art NN approaches are showing more promise for Type III/IV clones;
- But they rely on pairwise comparison of code segments and do not scale well as that involves ~O(n^2) comparisons
  - Oreo took nearly 1 day, 21 hours to work its way through the standard benchmark in the field: BCB - 250 MLOC in Java
- We tried a Nearest Neighbour approach.

3 SSCD:


A: Pre-processing source code
Source code repository
Method 1
tokens
Method N
tokens
B: Fine-tuning the large language models (LLMs)*
Clone Pairs
Non-Clone Pairs
w1
w2
w3
w4
w5
w6
w7
w8
w9
w10
w11
w12
w13
w14
w15
768 Vectors
Similarity compared to actuality -> feedback to NN weights/∑
= more similar vectors for clones
= less similar vectors for non-clones
* Support for various LLM types and LLM ensembles
C: Generating semantic vectors
768 x 1
Fine-tuned LLM or Ensemble of LLMs
Vectors based on code partition
D: Build index for corpus
Metadata Service
SQL Database
E: Clone Retrieval
Ranked clone list
Top-K retrieval using k-nearest neighbour search
Refined clone list
Clone reranking (cross encoders)

LLM-encoded vectors, where nearness of vectors reflects code similarity (clones)

4 SSCD, with/without Remove-White-Space/Variable Anonymization Pre-processings, with Active Learning and with state-of-the-art LLMs

Max Recall at a precision of >=0.2 (across a Huawei code-base and 8 OS systems: 362 MLOC):

>=0.2 Precision	Clones Identified by SSCD Variants (and CCFinderX – After filtering)							CCFinderX Totals (after filter)
	Candidates	True clones	Ranked Precision	T1	T2	T3	T4	
SSCD	664	133		18 (15)	14 (14)	98 (14)	3 (2)	45
SSCD (RWS)	1390	279		21 (15)	66 (13)	180 (20)	12 (6)	54
SSCD (RWS + VA)	1496	300		20 (15)	55 (13)	205 (11)	20 (5)	44

Max Recall at a precision of >=0.2 (across the Huawei code-base and the 4 OS systems not used for Active Learning):

>=0.2 Precision	Clones Identified by SSCD Variants (and CCFinderX – After filtering)							CCFinderX Totals (after filter)
	Candidates	True clones	Ranked Precision	T1	T2	T3	T4	
SSCD (RWS)	415	84		4 (3)	28 (4)	47 (6)	5 (2)	15
SSCD (AL)	895	179		4 (3)	43 (11)	121 (6)	11 (4)	24

Preliminary results of trialling newer LLMs on the Huawei-provided Benchmark

Newer LLMs trialled on Huawei's 480KLOC dataset			
LLM	Size (Parameters)	C F-score	C++ F-score
Code T5	220M	85.04	90.45
CodeBERT	125M	77.16	83.33
GraphCodeBERT	125M	80.29	88.31
CuBERT	345M	97.14	95.6
Code T5+	110M	99.29	97.04
SPT-Code	262M	97.84	92.77

- Overall (and in Type 3 particularly), SSCD shows substantial improvement over CCFinderX;
- And pre-processings show further significant improvement;
- Preliminary indications suggest that incorporation of newer LLMs will improve things further.

HOST INSTITUTION

UNIVERSITY OF LIMERICK OLLSCOIL LUIMNIGH

PARTNER INSTITUTIONS

Ollscoil na Gaillimhe University of Galway
Maynooth University
Trinity College Dublin
University College Dublin
UCC
MTU
DCU
SETU
Dundalk
TUS
ATU

FUNDED BY:

Ireland's European Structural and Investment Funds Programmes 2014-2020
European Union European Regional Development Fund
Science Foundation Ireland For what's next