

Code Genome

IBM Research Security

—
Doug Schales
Dhilung Kirat
Jiyong Jang
Ian Molloy
Ted Habeck
JR Rao

Code Genome = Code Gene + Knowledge Graph

```

; section: __text
; function: _f1 at 0x10000c20 -- 0x10000c35
0x10000c20: 55      push rbp
0x10000c21: 48 89 e5  mov rbp, rsp
0x10000c24: 89 7d fc  mov dword ptr [rbp - 4], edi
0x10000c27: 8b 7d fc  mov edi, dword ptr [rbp - 4]
0x10000c2a: 83 c7 20  add edi, 0x20
0x10000c2d: 89 7d f8  mov dword ptr [rbp - 8], edi
0x10000c30: 8b 45 f8  mov eax, dword ptr [rbp - 8]
0x10000c33: 5d      pop rbp
0x10000c34: c3      ret
    
```

```

int f1(int a){
    int x;
    x = a + 32;
    return x;
}
    
```

Compile

Lift

```

define i64 @_f1(i32 %arg1) local_unnamed_addr {
dec_label_pc_10000c20:
    %tmp5 = zext i32 %arg1 to i64
    %v2_10000c2a = add nuw nsw i64 %tmp5, 32
    %v17_10000c2a = and i64 %v2_10000c2a, 4294967295
    ret i64 %v17_10000c2a
}
    
```

Canonicalize

```

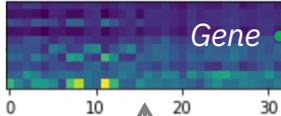
define i64 @_F(i32 %a1) local_unnamed_addr #0 {
b1:
    %v0 = add i32 %a1, 32
    %v1 = zext i32 %v0 to i64
    ret i64 %v1
}
    
```

Convert

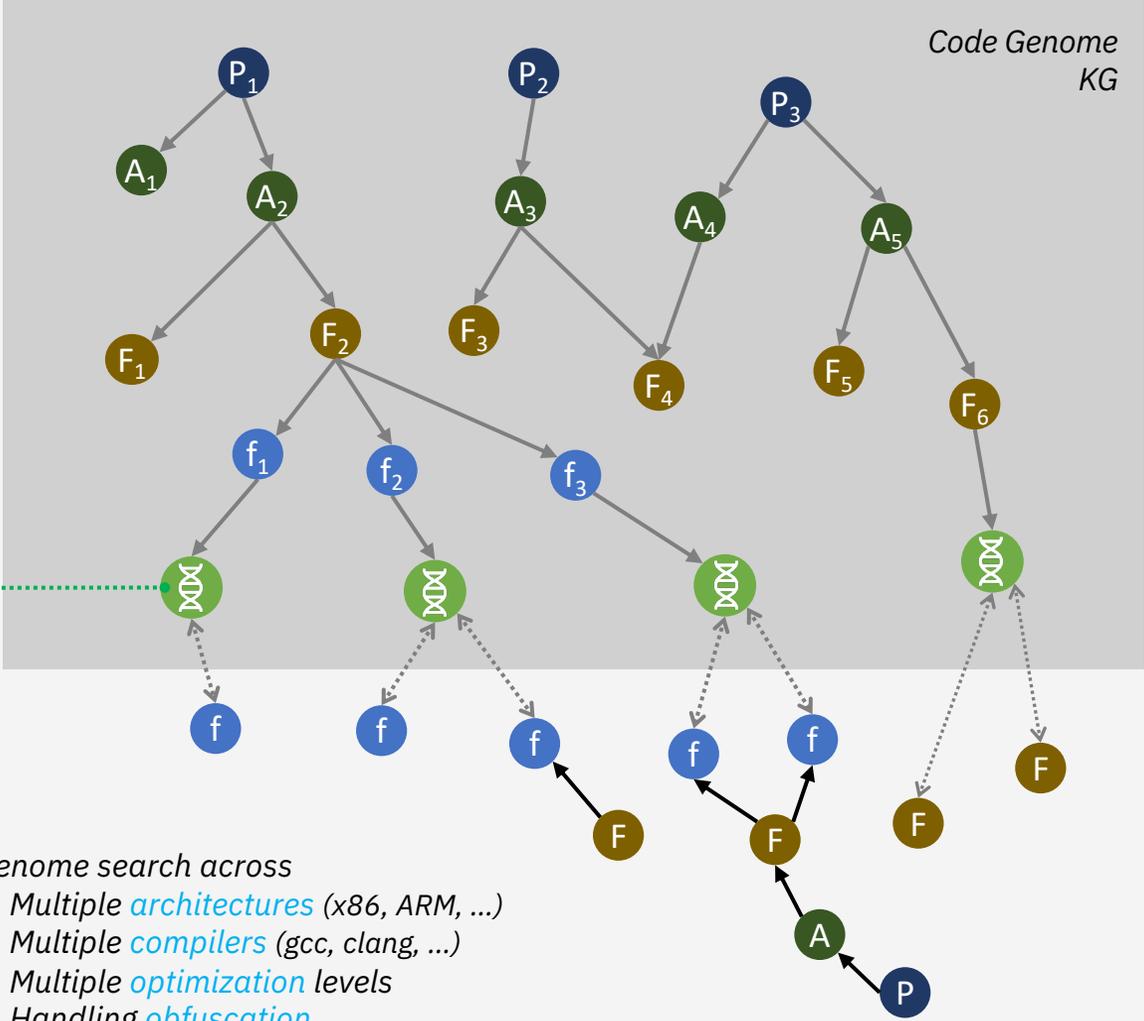
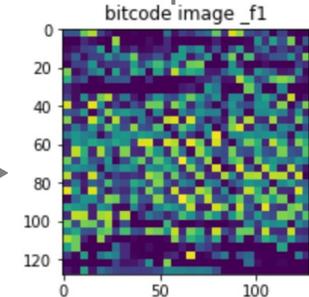
```

Offset: 00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F
00000000: 42 43 C0 DE 35 14 00 00 05 00 00 00 62 0C 30 24  BC@*5.....b.0s
00000010: 49 59 0E 26 EF D3 3E 20 44 01 32 05 00 00 00 00  IY>6oS=-0.2...
00000020: 21 0C 00 00 D5 00 00 00 0B 02 21 00 02 00 00 00  |...U.....1...
00000030: 16 00 00 00 07 81 23 91 41 C8 04 49 06 10 32 39  .....#.AH.I..29
00000040: 92 01 84 0C 25 05 08 19 1E 04 88 62 80 0C 45 02  .....b..E.
00000050: 42 92 0B 42 64 10 32 14 38 08 18 4B 0A 32 32 88  B..Bd.2.8..A.22.
00000060: 48 70 C4 21 23 44 12 87 8C 10 41 92 02 64 C8 08  HpDI#D...A..dH.
00000070: B1 14 20 43 46 88 20 C9 01 32 32 84 18 2A 28 2A  1..CF..I.22..*(*
00000080: 90 31 7C B0 5C 91 20 C3 C8 00 00 00 89 20 00 00  .1|0\..CH.....
00000090: 0C 00 00 00 32 22 C8 08 20 64 85 04 93 21 A4 84  ..d...1s.
000000a0: 04 93 21 E3 84 A1 90 14 12 4C 86 8C 0B 84 64 4C  .s..0C..E0m.#.
000000b0: 10 18 73 04 A0 30 47 00 06 45 40 48 03 01 23 00
    
```

Convert



Embedding



Code Genome Use Cases

Gene similarity: 96

File Name: coreutils-8.29_gcc-4.9.4_x86_64_03_touch.elf	File Name: coreutils-8.29_gcc-4.9.4_arm_64_03_touch.elf
File Hash: d79d11381f9be54a1585567c331813423f96712270af33368f6f...	File Hash: 4c9299905a3bc1e3be98b92578864c2fd14477ba2a20c62b02eaf7047b3f6...
File Type: abi-sysv, elf-file, elf-exec, arch-x86_64	File Type: abi-sysv, arch-arm64, elf-file, elf-exec
Last updated: 2023-05-08T09:01:41.000Z	Last Updated: 2023-05-08T08:17:38.000Z
File size: 314864	File size: 274264
Gene count: 200	Gene count: 192

coreutils-8.29_gcc-4.9.4_x86_64_03_touch.elf (d79d1138) Functions	coreutils-8.29_gcc-4.9.4_arm_64_03_touch.elf (4c929990) Functions	Score	op
argmatch	argmatch	99	~
debug_strftime.constprop.11	debug_strftime.constprop.12	99	~
emit_bug_reporting_address	emit_bug_reporting_address	99	~
fdutimensat	fdutimensat	99	~
mktime_ok	mktime_ok	99	~
mktime_2	mktime_2	99	~
posix2_version	posix2_version	99	~
posixtime	posixtime	99	~
process_long_option	process_long_option	99	~
rpl_fclose	rpl_fclose	99	~
rpl_fflush	rpl_fflush	99	~
rpl_fseeko	rpl_fseeko	99	~
set_tz	set_tz	99	~
time_zone_hhmm.isra.4	time_zone_hhmm.isra.4	99	~
tzalloc	tzalloc	99	~
x2nrealloc	x2nrealloc	99	~
xrealloc	xrealloc	99	~
__do_global_ctors_aux	get_quoting_style	98	~
__libc_csu_init	__libc_csu_init	98	~
__xargmatch_internal	__xargmatch_internal	98	~

Semantic search

Same source code compiled in different environments and setups generates the same gene.

Compare

Gene similarity: 99

File Name: openssl_1.0.1f	File Name: openssl_1.0.1g
File Hash: b0cf5cbcb674a8c6935c7ea248450c485fbc6f4a44bc3e4d4ff30c5cdfff...	File Hash: 5f521e31e493829d35a31e23e6ed10c778ccdf0af5015fe239f29230f335...
File Type: abi-sysv, elf-shared-object, elf-file, arch-x86_64	File Type: abi-sysv, elf-shared-object, elf-file, arch-x86_64
Last updated: 2023-05-11T21:27:51.000Z	Last Updated: 2023-05-11T21:29:37.000Z
File size: 3115016	File size: 3115048
Gene count: 4649	Gene count: 4649

openssl_1.0.1f (b0cf5cbb) Functions	openssl_1.0.1g (5f521e31) Functions	Score	op
CMS_decrypt_set1_password	CMS_decrypt_set1_password	98	
tlst_heartbeat	tlst_heartbeat	98	
dtls_process_heartbeat	dtls_process_heartbeat	98	
dtls_heartbeat	dtls_heartbeat	98	
tlst_process_heartbeat	tlst_process_heartbeat	98	
function_eccd87	function_eccd87	0	+
function_eccd47	function_eccd47	98	-
ACCESS_DESCRIPTION_free	ACCESS_DESCRIPTION_free	100	=
ACCESS_DESCRIPTION_new	ACCESS_DESCRIPTION_new	100	=
AES_bi_ige_encrypt	AES_bi_ige_encrypt	100	=
AES_decrypt	AES_decrypt	100	=
AES_encrypt	AES_encrypt	100	=
AES_ige_encrypt	AES_ige_encrypt	100	=
AES_options	AES_options	100	=

Code evolution

Examine the changes between versions and analyze how code has evolved over time.

Unknown package

Genome KG

Component	Version	License
arping	20160308	BSD and GPLv2+
clockdiff	20160308	BSD and GPLv2+
ifenslave	20160308	BSD and GPLv2+
iputils	20160308	BSD and GPLv2+
ping	20160308	BSD and GPLv2+
rdisc	20160308	BSD and GPLv2+
tracepath	20160308	BSD and GPLv2+
tracepath6	20160308	BSD and GPLv2+

Attribution & Forensics

Genome KG built on semantic fingerprinting allows identifying unknown binary code and understanding provenance and evolution.

- Cloud-native process engine and knowledge graph
- Currently supported
 - Binaries: ELF, PE, Mach-0
 - Architectures: x86, x86_64, arm, aarch64, mips, ppc
 - Packages: deb, rpm, ipa
 - Archives: ar, cpio, tar, bzip2, gzip, zstd, xz, rar, 7zip
- Demo is available at <https://youtu.be/1JtaPY9TRfA>