其次,变量类型的归一.由于跨平台编程等原因,很多数据类型经过 *typedef* 的定义,如 *size_t* 就是从 *unsigned int* 定义来的,如果不做处理就可能因为类型名不同使得向量化后不一致,从而漏掉某些相似的函数.根据变量类型表示的数据范围不同,将常见的一些数据类型加以总结并分类统一表示,具体分类见表 1.

Table 1 Clusters of common data types

表 1 常见数据类型的分类

数据类型	变量类型分类
bool	bool, _Bool
char	signed char, s8, __s8, int8_t
unsigned char	unsigned char, u_char, unchar, u8, __u8, uint8_t, __uint8_t, u_int8_t, Byte, Byte_t, byte_t
Short	short int, s16, __s16, int16_t
unsigned short	short unsigned int, umode_t, u_short, ushort, u16, __u16, uint16_t, __uint16_t, u_int16_t
int	int, s32, __s32, int32_t, int32
unsigned int	unsigned int, u_int, uint, u32, __u32, uint32_t, __uint32_t, u_int32_t, uint32
long	long int
unsigned long	long unsigned int, uintptr_t, u_long, ulong
long long	long long int, s64, __s64, int64_t
unsigned long long	Long long unsigned int, u64, __u64, uint64_t, __uint64_t, u_int64_t, u_quad_t
size_t	32bit: unsigned int, 64bit: unsigned long
ssize_t	32bit: int, 64bit: long

另外,函数调用名称也需要进行归一化.事实上,有很多函数实现的功能相似,只是根据上下文环境不同调用者选取了不同的函数进行调用.比如图 1 所举的实例中,已知漏洞函数调用的堆空间分配函数是 av_malloc,而未知漏洞中调用的是 av_mallocz.这两个函数同样用于在堆中分配一定大小的空间,后者是前者的一个包装,只在前者的基础上多加了一步初始化分配空间的操作.如果不进行函数名的归一,这两句调用对应到向量空间中就不一致,在进行相似匹配时也可能会将该未知漏洞漏掉.为了将这种情况也考虑在内,根据经验,开发者在开发类似功能代码时倾向于取相近的函数名,所以我们提取出待测代码集中所有被调用到的函数,利用字符串距离计算对它们进行了一个简单的聚类,将比较相似的字符串归并为同一个表示形式,比如 av_malloc 和 av_mallocz 都用 av_malloc 表示,达到函数名归一的目的.

1.4 特征向量映射

向量化是将函数和切片中提取的特征语句映射到向量空间的过程,是为了简化特征表示以及相似度的计算而采取的方法,相对于利用树和图等结构进行匹配和计算的方法减少了很多计算量.

待测代码集合 $C=\{F_1,...,F_n\}$表示其含有 n 个函数,V 代表 C 所对应的特征向量空间.我们可以将 C 到 V 的映射定义为ϕ.ϕ采用现有哈希算法 hashpjw[17]实现,每一条特征语句都会对应到 V 中的一个哈希值,则向量空间 V 的维度$|V|$就是 C 中所有函数所包含的所有语句的总数.而对于每个函数 $F_i(1\leqslant i\leqslant n)$,其向量的维度均为$|V|$,每一个哈希值 h 表示的维度上的数值根据下面公式计算:

$$\phi_{F_i}(h)=I(s,h)\cdot TF-IDF(s),\text{其中},I(s,h)=\begin{cases}1, & \text{如果哈希值}h\text{对应的语句}s\in F_i\\ 0, & \text{如果哈希值}h\text{对应的语句}s\notin F_i\end{cases}.$$

TF-IDF[18]是信息检索中常用的加权技术,其值等于词频 TF(term frequency)和逆向文件频率 IDF (inverse document frequency)的乘积.引入该权重主要是考虑到语句的频繁程度和重要程度不同,比如赋值特征语句 *int=int+int* 会在代码集中大量存在,需要降低该句对计算结果的影响;而某些语句在某些函数中频繁出现,在其他函数中则不常出现,所以为了评估每个哈希值对不同函数的重要程度,在向量化表示之后又添加了 TF-IDF 权重的计算.但是在本方法中我们将 TF 和 IDF 的计算从文档层面迁移到函数层面,而词的概念则对应于本方法中的哈希值,具体计算方法参考 Salton 等人的文献[18].

1.5 相似度计算与匹配

生成了特征向量之后就是利用相似度计算检测未知漏洞的过程.在本文提出的方法中,相似度计算与排序分为两次,第 1 次是用漏洞特征向量与待测函数的特征向量分别进行相似度的计算,将结果按照相似度排序形

成初步候选结果;第 2 次是对初步候选结果中的函数,利用补丁特征向量再与之进行一次相似度的计算,如果一个候选函数属于无漏洞的误报,那么该函数应该是含有补丁特征的函数,理论上其第 2 次匹配相似度应该高于或者至少是不低于初次计算时的相似度值.进行 2 次计算筛选的目的正是又一次利用补丁信息,除去初步候选集中的不含有漏洞的误报,进一步减轻审计的工作量.

两个特征向量 $A(a_1,...,a_n)$和 $B(b_1,...,b_n)$的距离采用余弦相似度计算,公式如下.距离值应该属于[0,1]区间内,数值越大表明两个向量距离越近,其代表的相应切片和函数也就越相似.数值为 0 时表示两个向量完全不同,没有任何一维特征重合;数值为 1 时表示两个向量完全相同,在所有特征维度上都重合.

$$Simi(A,B)=\frac{A\times B}{|A|\cdot|B|}=\frac{a_1\times b_1+a_2\times b_2+...+a_n\times b_n}{\sqrt{a_1^{\ 2}+a_2^{\ 2}+...+a_n^{\ 2}}\times\sqrt{b_1^{\ 2}+b_2^{\ 2}+...+b_n^{\ 2}}}.$$

2 检测实验与分析

2.1 实验设置

我们将本方法部署在 64 位的 Ubuntu 16.04 平台上,并利用 GCC 4.5 作为编译器.选取了多平台广泛支持的开源音视频处理库 FFmpeg 3.2.4 版本和开源图像浏览软件 Ghostscript 9.21 版本作为实验对象.其中,FFmpeg 包含 1 583 个文件,15 598 个函数,Ghostscript 含有 935 个文件,15 875 个函数.漏洞方面则选取这两个软件在 2017 年最新公布的漏洞作为已知的目标漏洞来进行检测实验.

2.2 性能分析

本方法涉及到 3 个主要的步骤,切片、归一化和向量化我们都直接部署在 GCC 中,在编译过程中同时实现切片和向量映射输出特征向量.相似度计算是在此之后单独进行的,而且相似度计算几乎都能在 1s 内完成,几乎不消耗时间,所以性能实验主要针对切片和向量映射所消耗的时间进行分析.我们首先对实验对象 FFmpeg 和 Ghostscript 进行单独编译记录下时间消耗,然后添加了本方法中切片、归一以及向量映射的处理过程后再次记录下运行时间,重复该过程 10 次,取时间损耗的平均值,结果见表 2.

Table 2 Time consumed by the slicing process and signature extracting process

表 2 本方法切片和获得特征向量过程的时间消耗

实验对象	单独编译时间(s)	加入切片和特征向量映射的时间(s)	切片和向量映射的性能开销(%)
FFmpeg	111.47	210.96	89
Ghostscript	165.59	270.69	63

可见,实际切片一直到特征向量映射步骤结束所需时间小于 1 次编译的时间,对于含 10 000 个以上函数的代码分析仅需几分钟,性能开销完全在可以接受的范围内.

2.3 检测结果

本次对 FFmpeg 和 Ghostscript 进行未知漏洞发现实验的过程中共检测出了 3 个新的未知漏洞,其中对图 1 中的 FFmpeg 的 CVE-2017-5025 漏洞检测出 1 个未知漏洞[19],对 Ghostscript 的 CVE-2017-5951 漏洞检测出了两个相似的未知漏洞[20,21],我们已经提交给相应的开发团队并且已经得到确认.下面将用 FFmpeg 的漏洞 CVE-2017-5025 为例,说明结合了补丁信息进行切片的方法,在降低含有漏洞的函数中的无关语句噪声及其造成的相关误报和漏报方面的有效性,以及 2 次计算利用补丁特征向量过滤无漏洞函数的误报的有效性.

首先,如果不进行切片,单纯使用整个函数 mov_read_hdlr 的所有语句作为特征进行向量化后直接进行相似度计算的结果见表 3(排序结果不包含目标漏洞所在函数).由于函数中漏洞无关的语句比例较高(56 行/61 行),造成最终的相似度计算结果受到了噪声的很大干扰,所有候选函数的相似度都偏低.同时从表 3 中可以发现,根本不含有漏洞相关语句的误报比例非常高,一共有 9 个(表 3 中×号所示),比例占前 15 个的一半以上,检测的准确性很低,误报较高.

Table 3 Top 15 most similar functions to CVE-2017-5025 without slicing process

表 3 不切片时前 15 个最相似于 CVE-2017-5025 的函数

排序	相似度	函数名	噪声误报(×)	排序	相似度	函数名	噪声误报(×)
1	0.67	read_header	×	9	0.53	rsd_read_header	×
2	0.63	get_aiff_header	×	10	0.53	mpc_read_header	–
3	0.63	ape_read_header	–	11	0.49	mov_read_ftyp	–
4	0.62	mov_read_frma	×	12	0.47	read_packet	×
5	0.60	mov_read_dref	–	13	0.47	read_header	×
6	0.59	qcp_read_packet	×	14	0.47	read_header	×
7	0.58	smacker_read_header	–	15	0.47	hls_write_header	–
8	0.54	read_ints	×	–	–	–	–

利用补丁信息和程序切片技术获取只含有漏洞特征向量后,进行一次相似度计算后排序的结果见表 4.实验结果表明,引入切片技术后再次计算的相似度结果中没有一个是由于漏洞无关噪声导致的误报,并且相似度普遍提高了.此时再看表 3 中的几个噪声误报的相似度值,原来排序 12 的 read_packet 相似度 0.47,现在相似度 0.34,排序掉到 287 名,其他 8 个因为噪声误报的函数切片后的相似度值均低于 0.2,尤其原来排名为 2 和 4 的 get_aiff_header 和 mov_read_frma 两个函数相似度更是不到 0.1.以上结果充分证明了通过切片确实能够有效去掉噪声影响,减少因为噪声产生的无关误报.此外,我们还发现表 4 候选结果中的第 12 个是一个已知漏洞,并且在我们发现之前刚打了补丁,而排序的第 2 个函数是一个未知漏洞,这两个函数在不切片的相似计算中并没有发现,说明切片降低了无关的噪声后,能够更准确地匹配出相似的函数,提高与已知漏洞相似函数的排序,减少漏报.但是在不经过第 2 次相似计算之前,此时的结果中有 10 个函数(表 4 中“×”号所示)都是已经含有补丁语句的函数,也是不会导致漏洞的误报.

最后,在表 4 的基础上,利用补丁产生的含补丁特征的向量与其中的函数进行 2 次相似度计算,结果见表 5.从表 5 的实验结果可以看出,有 5 个函数(表 5 中“×”号所示)的相似度值高于表 4 中第 1 次相似度计算结果,这说明这 5 个函数是已经包含有所添加的补丁特征的,那么它们就不会再含有该目标漏洞.可以从候选结果中去掉.

Table 4 Top 15 most similar functions to CVE-2017-5025 after slicing (only compute similarity once)

表 4 加入切片方法后的前 15 个最相似于 CVE-2017-5025 的函数(只进行 1 次相似计算)

排序	相似度	函数名	含补丁的误报(×)	排序	相似度	函数名	含补丁的误报(×)
1	0.89	mov_read_udta_string	×	9	0.72	avi_read_tag	×
2	0.80	**w64_read_header***	–	10	0.70	mov_read_wave	×
3	0.78	mpc8_parse_seektable	×	11	0.70	asf_read_stream_properties	×
4	0.78	mov_read_senc	–	12	0.68	**mov_read_keys(patched)**	×
5	0.76	wav_parse_bext_tag	×	13	0.68	mov_read_custom	×
6	0.75	mov_read_dref	–	14	0.67	ape_read_header	–
7	0.75	ff_read_riff_info	×	15	0.65	vmd_read_header	×
8	0.73	read_header	–	–	–	–	–

注:表中*表示未知漏洞

Table 5 Top 15 most similar functions to CVE-2017-5025 after the second time of similarity computing

表 5 进行 2 次相似度计算后的前 15 个最相似于 CVE-2017-5025 的函数

排序	相似度	函数名	含补丁的误报(×)	排序	相似度	函数名	含补丁的误报(×)
1	**0.92**	**mov_read_udta_string**	×	9	0.68	w64_read_header	–
2	**0.85**	**mpc8_parse_seektable**	×	10	0.66	mov_read_senc	–
3	**0.81**	**wav_parse_bext_tag**	×	11	0.64	ff_read_riff_info	–
4	**0.79**	**mov_read_wave**	×	12	0.63	mov_read_dref	–
5	0.76	hls_append_segment	–	13	0.62	read_header	–
6	0.74	open_input_file	–	14	0.61	avi_read_tag	–
7	**0.73**	**asf_read_stream_properties**	×	15	0.61	avi_read_header	–
8	0.72	ogg_read_page	–	–	–	–	–

因此,最终候选集将见表 6,人工审计就可以仅对这些函数进行,减少了审计的工作量.表 6 中仍然有 5 个含

有补丁语句的误报,未能被第 2 次相似度计算筛掉,原因是它们打补丁的方式不同,因而本次实验使用的含补丁特征的向量未能提高这几个函数的相似度.

Table 6 Final candidate functions after filtering by the second similarity computing

表 6 经过第 2 次相似度计算的筛选后得到的最终候选集

排序	相似度	函数名	含补丁的误报(×)	排序	相似度	函数名	含补丁的误报(×)
1	0.80	**w64_read_header***	–	6	0.72	avi_read_tag	×
2	0.78	mov_read_senc	–	7	0.68	**mov_read_keys(patched)**	×
3	0.75	mov_read_dref	–	8	0.68	mov_read_custom	×
4	0.75	ff_read_riff_info	×	9	0.67	ape_read_header	–
5	0.73	read_header	–	10	0.65	vmd_read_header	×

注:表中*表示未知漏洞

2.4 对比实验

我们将本方法与相关工作进行了对比实验,通过对比进一步说明本方法确实能够有效地降低与漏洞无关的噪声带来的误报,并能利用补丁进一步过滤结果中与补丁特征相似的误报.由于无法直接获得这些相关工作的原本实现,本文重新实现了最具代表性的 Yamaguchi 等人的"漏洞外推"方法[11]作为对照.我们按照相关论文的描述,还原了其每一个步骤,包括特征选取、向量映射、主成分分析以及相似度计算等步骤.并且在特征选取和参数设置等方面都与原文保持一致.我们使用所实现的漏洞外推方法在同一个代码集 FFmpeg 3.2.4 上,对同一个漏洞 CVE-2017-5025 进行检测,得到的结果中前 15 个最相似的函数见表 7.

Table 7 Top 15 most similar functions to CVE-2017-5025 using the vulnerability extrapolation approach

表 7 漏洞外推方法检测到的前 15 个最相似于 CVE-2017-5025 的函数

排序	相似度	函数名	误报类型	排序	相似度	函数名	误报类型
1	0.88	mov_read_saiz	II	9	0.81	mov_metadata_hmmt	I
2	0.85	mov_read_sbgp	I	10	0.80	mov_read_pasp	I
3	0.84	mov_read_trex	II	11	0.79	mov_read_stss	II
4	0.83	mov_read_mdhd	I	12	0.79	mov_read_stts	I
5	0.82	mov_read_stsc	I	13	0.78	mov_read_trun	II
6	0.82	mov_read_tfhd	I	14	0.78	mov_read_tkhd	I
7	0.82	mov_read_elst	I	15	0.77	mov_read_stco	I
8	0.82	mov_read_ctts	II	**1 5542**	**0.17**	**w64_read_header**	–

通过分析,表 7 中所示函数均为误报,类型 I 表示缺少漏洞形成条件的函数,比如根本不含有堆分配操作或者分配空间大小是一个常量,也就是根本不会导致漏洞发生的误报;类型 II 表示已经含有对分配空间大小进行检查的操作,也就是含有补丁特征的误报.这些函数均与该漏洞所在函数处于同一个 C 语言文件中,均是解析处理特定类型文件的函数.本文方法与之相对比,可以明显看到,表 7 中的结果没有一个出现在利用本方法进行检测实验的候选集中.此外,本方法检测到的未知漏洞函数 w64_read_header 用漏洞外推的方法计算相似度仅为 0.17,排名 15542/15598,显然也是无法检测出的.

检测失败的主要原因是,Yamaguchi 等人的方法在进行漏洞外推时实际上依赖的是函数间的相似度而非漏洞相关代码间的相似度.虽然 Yamaguchi 等人引入了主成分分析来发现用以比较的主要维度,但无助于减低函数中的漏洞无关代码所引入的噪声,这些噪声代码既参与到主成分的计算中,又最终影响到了最终的比较.虽然该方法计算出的候选函数相似度值普遍较高,但多数排序靠前的候选是在语义和功能上与漏洞所在函数有所相似的函数,并不一定是在漏洞特征上相似,所以检测结果存在很多误报.

对比实验结果表明,与最具代表性的漏洞外推方法相比,本方法利用补丁信息和切片技术能够有效减少漏洞无关语句的噪声影响,从而减少其引发的大量误报,切实提高漏洞检测工作的效能,减少了人工审计的工作量.

3 讨论与展望

3.1 代码向量空间

本文直接在代码向量空间中实施相似度比较,而没有使用主成分分析、潜在语义分析等降维的方法,原因在于本方法通过引入了漏洞补丁的信息,利用程序切片技术将待比较的代码直接聚焦于与漏洞关系密切的语句上,天然地达到了降维的效果,避免了主成分分析等复杂计算的开销.而在此过程中,本文实际上也考虑到了代码作为结构化文本所具有的语义信息,那就是切片时依据的控制依赖和数据依赖关系.这使得在代码向量中保留有漏洞相关代码的关键结构信息.虽然本方法在代码向量空间中尽量去除了影响比较的噪声代码,但无可避免还会存在漏报误报.但实验结果表明,本方法切实提高了漏洞代码相似性检测的有效性,能够减少大量的人工审计的工作.

3.2 方法的适用性

到目前为止,本文主要是针对一般性的补丁情况也就是增改代码的补丁进行讨论的.但在实际中,虽然情况比较少见,但确实存在少数漏洞补丁只减少代码.虽然本方法在设计时主要考虑的是常见的具有代码增改情况的补丁,但实际上对于只减少代码的补丁也是同样适用的,只是需要对最终候选集的生成条件作出一点调整,方法的其他部分保持不变.前面提到最终候选集需要在第 1 次相似计算结果中,去掉第 2 次相似度计算的值大于或者至少等于第 1 次计算结果的那些函数.这是因为对于增改代码的补丁,补丁特征向量将比漏洞特征向量含有更多信息,与补丁更相似的那些函数相似度值则自然不小于其与漏洞特征的相似度值.而对于只减少代码的补丁,情况正好相反,漏洞特征向量包含的信息要多于其相应的补丁特征向量,因此最终候选集需要由第 1 次相似度计算值大于或等于第 2 次计算结果的那些函数组成.

我们以 Linux 内核代码为目标调查了 300 个最新漏洞的补丁,找到了一个只减少代码的编号为 CVE-2016-4557 的漏洞来验证方法的适用性,其补丁如图 4(左)所示.该漏洞出现的原因是函数__bpf_map_get 在实现时已经在错误处理路径上调用过 fdput 函数,而在 replace_map_fd_with_map_ptr 中又再次调用了这个函数,从而造成了 Double Free 型漏洞.因此该漏洞的补丁就是去掉 replace_map_fd_with_map_ptr 中对 fdput 的调用语句.由于在实际检测中并没有检测到新的未知漏洞,为了体现实验效果,我们选择了一个与漏洞特征有所相似的函数 map_delete_elem 并在其中添加了一条对 fdput 的调用语句来制造出一个漏洞,然后通过实验检验本方法是否能够检测出该函数.

```
    //kernel/bpf/verifier.c
2864 static int replace_map_fd_with_map_ptr(
                struct bpf_verifier_env *env)
2904 { […]
2905   f = fdget(insn->imm);
2906   map = __bpf_map_get(f);
2907   if (IS_ERR(map)) {
2908     verbose("fd %d is not pointing to
            valid bpf_map\n", insn->imm);
2909 -   fdput(f);
2910     return PTR_ERR(map);
       } […]
2946   fdput(f); […]
2957   return 0;
2958 }
```

```
    //kernel/bpf/syscall.c
423 static int map_delete_elem(union bpf attr
                                  *attr)
424 { […]
435   f = fdget(ufd);
436   map = __bpf_map_get(f);
437   if (IS_ERR(map)) {
438     fdput(f);  //添加这条语句,人为地制造一个漏洞
439     return PTR_ERR(map);
440   }
      […]
463   fdput(f); […]
464   return 0;
465 }
```

Fig.4 Patch for CVE-2016-4557 in Linux kernel (left) and a function with a faked vulnerability (right)

图 4 Linux 内核漏洞 CVE-2016-4557 的补丁(左)及一个模拟了相似漏洞的函数(右)

具体实验结果见表 8,根据调整方法最终候选集就应该由第 1 次相似度计算值大于等于第 2 次的那些函数组成,排序在第 2~第 5 的函数就是已经含有补丁特征的误报,无需再进行审计.最终候选集中的第 1 个就是目标函数 map_delete_elem,可以很容易被检测出.排序在第 6~第 10 的几个函数是不含有漏洞的误报,误报原因是它

们没有调用函数__bpf_map_get 而使用了其他函数,而那些函数的调用不会导致 Double Free 漏洞.

Table 8 Top 10 most similar functions to the vulnerability vector after the first time of similarity computing and their results of the second similarity computing

表 8 根据漏洞特征向量进行第 1 次相似度计算后的前 10 个最相似的函数及其第 2 次相似度计算值

排序	函数名	第 1 次相似度计算值	第 2 次相似度计算值	最终候选集(√)
1	**map_delete_elem**	**0.99**	**0.95**	√
2	map_update_elem	0.93	0.98	–
3	map_get_next_key	0.92	0.98	–
4	bpf_map_get_with_uref	0.91	0.95	–
5	map_lookup_elem	0.90	0.96	–
6	memcg_write_event_control	0.87	0.72	√
7	vfio_virqfd_enable	0.87	0.74	√
8	ib_uverbs_open_xrcd	0.85	0.72	√
9	perf_cgroup_connect	0.82	0.78	√
10	ucma_migrate_id	0.82	0.79	√

当然,没有一种检测方法是能够覆盖所有漏洞类型的,本方法也有一些使用效果不佳的情况.本方法目前暂未考虑补丁中不涉及语句内容修改和数据类型变化的漏洞.比如常量的数值错误引起的漏洞.由于本方法进行归一化时会忽略变量的值,因此即使有与漏洞特征向量极其相似的候选函数,也并不代表它一定含有未知漏洞.对于这种情况,本方法无法正确区分漏洞和正常代码,也无法区分漏洞和补丁特征.再比如,语句执行顺序错误的漏洞,其相应的补丁未改变语句内容和数据类型,仅调换了语句次序,那么利用本方法得到的漏洞特征向量和补丁特征向量很可能是一样的.向量空间模型不考虑维度之间的顺序,则本方法计算出的与漏洞特征相似的函数中,可能含有大量与补丁特征一致,但实际并没有漏洞的函数.

3.3 未来工作

目前来说,本方法最需要改进的地方是引入类比的功能.比如 FFmpeg 中的两个同样是将数据从原始数据结构 AVIOContext 中读取到缓冲数组中的函数 ffio_read_size 和 avio_get_str,如果能够由前者类比到后者,就能更加提高本方法的精度.未来我们准备借鉴 Word2Vec[22]和 API2VEC[23]一类的方法来解决这个问题.

另外,本方法目前仅包含过程内的分析,相似度的计算以函数为对象.因为从经验的角度分析,一个函数是程序代码中的一个比较独立的功能单位,并且很多漏洞形成的原因及漏洞影响范围都局限在一个函数内部.另外,也是出于对切片和相似度计算性能的考虑,本方法仅实现了函数级别的相似度检测.所以对于需要进行过程间分析的相关漏洞,暂时不在本方法的适用范围之内.未来工作可能考虑引入过程间分析.

最后,其实有关克隆漏洞(clone bugs)检测的研究工作在某些思想上与本文的相似性检测方法具有一定程度的相关性,未来可能考虑研究克隆检测的相关方法是否能够被利用到本工作中.

4 相关工作

代码分析和软件漏洞检测工作一直都是软件安全领域的重要组成部分.尤其是静态分析在检测软件漏洞方面,已经有很多的方法和技术.有些工作是依赖于先验知识的,需要提前给出规则相关的一些知识,再静态分析代码生成规则,用以进行检测.如 Tan 等人提出的 iComment[3],是事先确定好要发现的规则类型并给出了模板,然后利用自然预言处理和机器学习等方法从源码中的注释提取规则填充模板内容,利用模板规则检查代码是否遵循注释的要求.另外 Tan 等人还提出了 AutoISES[4],关注的是 Linux 中的安全检查函数,也是静态分析代码抽取这些函数频繁保护的数据以生成规则,检测违例.另外有一些工作聚焦于研究如何自动提取代码规则[2,5−10].典型代表有 PR-Miner[2]以及 AntMiner[9],都是利用数据挖掘的思想,将代码潜在规则的抽取转化为频繁项集的挖掘.APISan[10]也是自动提取编程规则检测违例的工具,它无需人工模板和 API 使用文档,使用一种放宽条件的符号执行方法提取 API 模式,检查 API 使用过程中返回值、参数条件、API 相关性以及 API 使用前后的检查条件等是否有违背.这些自动提取代码规则并用于检测的方法的有效性依赖于提取规则的正确性,但是自

动提取很可能会出现错误.此外,此类工作也非常依赖于人工审计,审计的成本较大.

近年来,已有一些研究者提出利用漏洞相似性检测未知漏洞的方法.如 Yamaguchi 等人则从漏洞外推[11]的角度给静态分析检测漏洞提供了新的思路,在提取特征时由仅仅提取 API 符号作为特征变为利用函数的抽象语法树映射到特征向量.而 discovRE[12]方法应用于在二进制代码上查找与已知漏洞函数相似的函数,首先利用量化特征如指令条数和基本块个数等组成的特征向量计算出一系列相近的候选函数,再从它们的结构特征也就是 CFG 图的相似度上进行比较筛选.但是该方法仍需要进行图匹配的计算,有一定的性能开销.同样是在二进制代码上进行漏洞检测,Feng[13]等人设计实现的漏洞搜索引擎 Genius 为了进行跨平台的漏洞检测,将控制流图抽象成平台无关的具有普遍适用性的特征,以实施一种可扩展的漏洞检测.此类方法大多采用漏洞所在函数整体作为特征提取的主体,会导致含有漏洞的函数中的无关语句形成相似计算时的噪声,从而产生相应的误报和漏报,这正是本方法被提出的原因.

另外,克隆检测的思想在某种程度上与本类方法具有相似性.如 Yamaguchi 等人[24]提出过利用源代码生成的代码属性图,用图查询方式进行克隆代码检测的方法,相对于我们的方法来说,利用图进行检测的开销较大.CP-Miner[25]则是基于符号特征的,生成符号序列检查其他代码中的复制序列,但需要挖掘频繁序列,可能出现很多误报.DECKARD[26]利用抽象语法树作为特征生成向量,计算向量相似度来进行克隆代码检测.这与 Yamaguchi 等人的漏洞外推方法在思路上很相似,只是相对于检测克隆代码,相似性漏洞检测需要研究如何得到抽象程度更高、更能应对代码变化的特征.Kim 等人在 2017 年提出的 VUDDY[27]则根据已有漏洞所在函数提取特征指纹,检测是否有函数因为代码克隆而与该段代码含有相同漏洞.但这种方法中所有变量和函数全部用同一种类型和函数名称替换,粒度较粗.且每个函数体计算一个哈希值,只能检测精确克隆造成的相似漏洞,不能检测出代码结构有所改变的克隆函数,而本文提出的方法则考虑了这一点.

最后,值得一提的是有一些工作已经意识到补丁对于漏洞发现和利用的意义,其中比较有代表性的是 Brumley 等人提出的 APEG[28]方法.该方法从攻击者的角度考虑,如果提前得到漏洞补丁就可以在未及时打补丁的时间空隙进行漏洞利用.该方法主要关注未合理验证输入的漏洞类型,且注意到了补丁的意义所在并加以利用:一是通过对比软件含漏洞的版本和加了补丁的版本可以定位新加入的输入检查语句,二是通过求解到达新检查语句的路径条件获得使检查失败的非法输入以攻破已有漏洞.本文方法同样利用了补丁的价值,但是使用的目的和方法都与 APEG 有天壤之别.APEG 从攻击的角度着眼于漏洞本身的利用,使用补丁是为了获得非法输入.本文方法则是从已知漏洞出发去检测发现相似的未知漏洞,并且利用补丁帮助去掉可疑候选结果中的部分误报.另外,本文方法并未特地区分漏洞的类型.

5 总 结

本文分析了当前一种由含有漏洞的代码段出发静态检测未知漏洞的相似性检测思路及其局限性,并提出了一种利用补丁的未知漏洞发现方法.该方法结合了漏洞的补丁信息以及程序切片的技术,准确去除已知漏洞函数中无关语句带来的噪声,减少了因这部分噪声带来的部分误报和漏报;并用两次相似度计算的方法分别用含有漏洞函数和打过补丁的函数的特征向量与待测函数的特征向量分别进行相似计算,将计算结果中已经含有补丁语句特征的函数过滤掉,减少了这部分误报,更加便于后续审计.实验结果表明,本文方法对于前期噪声的减少是有效的,对 FFmpeg 和 Ghostscript 两个开源代码库的实验还检测出了 3 个新的未知漏洞,说明本文方法在未知漏洞发现方面是有效的.

References:

[1] Hopcroft J, Motwani R, Ullmann J. Introduction to Automata Theory, Languages, and Computation. 2nd ed., New York: Addison-Wesley, 2001.

[2] Li ZM, Zhou YY. PR-Miner: Automatically extracting implicit programming rules and detecting violations in large software code. In: Proc. of the European Software Engineering Conf. Held Jointly with ACM SIGSOFT Int'l Symp. on Foundations of Software Engineering. ACM Press, 2005. 306−315. [doi: 10.1145/1081706.1081755]

[3] Tan L, Yuan D, Krishna G, Zhou YY. iComment: Bugs or bad comments. In: Proc. of ACM Symposium on Operating Systems Principles. ACM Press, 2007. 145−158. [doi: 10.1145/1294261.1294276]

[4] Tan L, Zhang XL, Ma X, Song WW, Zhou YY. AutoISES: Automatically inferring security specification and detecting violations. In: Proc. of the USENIX Security Symp. USENIX Association Press, 2008. 379−394.

[5] Tan L, Zhou YY, Padioleau Y. aComment: Mining annotations from comments and code to detect interrupt related concurrency bugs. In: Proc. of the Int'l Conf. on Software Engineering. ACM Press, 2011. 11−20. [doi: 10.1145/1985793.1985796]

[6] Pradel M, Jaspan C, Aldrich J, Gross TR. Statically checking API protocol conformance with mined multi-object specifications. In: Proc. of the Int'l Conf. on Software Engineering. IEEE Computer Society Press, 2012. 925−935. [doi: 10.1109/ICSE.2012.6227127]

[7] Yamaguchi F, Wressnegger C, Gascon H, Rieck K. Chucky: Exposing missing checks in source code for vulnerability discovery. In: Proc. of the ACM SIGSAC Conf. on Computer & Communications Security. ACM Press, 2013. 499−510. [doi: 10.1145/2508859.2516665]

[8] Yamaguchi F, Golde N, Arp D, Rieck K. Modeling and discovering vulnerabilities with code property graphs. In: Proc. of the IEEE Symp. on Security and Privacy. IEEE Computer Society Press, 2014. 590−604. [doi: 10.1109/SP.2014.44]

[9] Liang B, Bian P, Zhang Y, Shi WC, You W, Cai Y. AntMiner: Mining more bugs by reducing noise interference. In: Proc. of the Int'l Conf. on Software Engineering. ACM Press, 2016. 333−344. [doi: 10.1145/2884781.2884870]

[10] Yun I, Min C, Si X, Jang Y, Kim T, Naik M. APISan: Sanitizing API usages through semantic cross-checking. In: Proc. of the USENIX Security Symp. USENIX Association Press, 2016. 363−378.

[11] Yamaguchi F, Lottmann M, Rieck K. Generalized vulnerability extrapolation using abstract syntax trees. In: Proc. of the Annual Computer Security Applications Conf. ACM Press, 2012. 359−368. [doi: 10.1145/2420950.2421003]

[12] Eschweiler S, Yakdan K, Padilla EG. discovRE: Efficient cross-architecture identification of bugs in binary code. In: Proc. of the 23nd Annual Network and Distributed System Security Symp. The Internet Society Press, 2016.

[13] Feng Q, Zhou RD, Xu CC, Cheng Y, Testa B, Yin H. Scalable graph-based bug search for firmware images. In: Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. ACM Press, 2016. 480−491. [doi: 10.1145/2976749.2978370]

[14] FFmpeg. http://ffmpeg.org/

[15] Ghostscript. https://www.ghostscript.com/

[16] GNU Compiler Collections (GCC) Internals. https://gcc.gnu.org/onlinedocs/gccint

[17] Aho AV, Sethi R, Ullman JD. Compilers: Principles, Techniques, and Tools. World Student Series ed., New York: Addison-Wesley, 1986.

[18] Salton G, McGill MJ. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1984.

[19] The Commit of FFmpeg. https://git.ffmpeg.org/gitweb/ffmpeg.git/commitdiff/a4fb44723dcaa56416173bc3d0ff41e9cda25067?hp=25a592e5d45672bdfdac35bf0119907cdcddd1b7

[20] Bug 698066 of Ghostscript. https://bugs.ghostscript.com/show_bug.cgi?id=698066

[21] Bug 698073 of Ghostscript. https://bugs.ghostscript.com/show_bug.cgi?id=698073

[22] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proc. of the Advances in Neural Information Processing Systems 26: The 27th Annual Conf. on Neural Information Processing Systems. 2013. 3111−3119.

[23] Nguyen TD, Nguyen AT, Phan HD, Nguyen TN. Exploring API embedding for API usages and applications. In: Proc. of the Int'l Conf. on Software Engineering. IEEE/ACM Press, 2017. 438−449. [doi: 10.1109/ICSE.2017.47]

[24] Yamaguchi F, Maier A, Gascon H, Rieck K. Automatic inference of search patterns for taint-style vulnerabilities. In: Proc. of the IEEE Symp. on Security and Privacy. IEEE Computer Society Press, 2015. 797−812. [doi: 10.1109/SP.2015.54]

[25] Li ZM, Lu S, Myagmar S, Zhou YY. CP-Miner: Finding copy-paste and related bugs in large-scale software code. IEEE Trans. on Software Engineering, 2006,32(3):176−192. [doi: 10.1109/TSE.2006.28]

[26] David Y, Yahav E. Tracelet-Based code search in executables. In: Proc. of the ACM SIGPLAN Conf. on Programming Language Design and Implementation. ACM Press, 2014. 349−360. [doi: 10.1145/2594291.2594343]

[27] Kim SB, Woo SH, Lee HJ, Oh HJ. VUDDY: A scalable approach for vulnerable code clone discover. In: Proc. of the IEEE Symp. on Security and Privacy. IEEE Computer Society Press, 2017. 595−614. [doi: 10.1109/SP.2017.62]

[28] Brumley D, Poosankam P, Song D, Zheng J. Automatic patch-based exploit generation is possible: Techniques and implications. In: Proc. of the IEEE Symp. on Security and Privacy. IEEE Computer Society Press, 2008. 143−157. [doi: 10.1109/SP.2008.17]

李赞(1993－),女,天津人,硕士生,主要研究领域为软件安全分析.

边攀(1987－),男,博士生,主要研究领域为程序静态分析.

石文昌(1964－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为系统安全,数字取证,基础软件.

梁彬(1973－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为软件系统安全性分析,系统软件安全机制,信息安全攻防对抗.