



دانشگاه صنعتی شریف  
دانشکده مهندسی کامپیوتر

پایان نامه کارشناسی ارشد  
مهندسی رایانش امن

# بهبود کارایی روش های تشخیص برنامه های اندرویدی باز بسته بندی شده

نگارش

مجتبی موذن

استاد راهنما

دکتر مرتضی امینی

بهمن ۱۴۰۱

به نام خدا  
دانشگاه صنعتی شریف  
دانشکده مهندسی کامپیوتر

## پایان نامه کارشناسی ارشد

این پایان نامه به عنوان تحقق بخشی از شرایط دریافت درجه کارشناسی ارشد است.

عنوان: بهبود کارایی روش های تشخیص برنامه های اندرویدی باز بسته بندی شده

نگارش: مجتبی مودن

## کمیته ممتحنین

استاد راهنما: دکتر مرتضی امینی امضاء:

استاد مشاور: استاد مشاور امضاء:

استاد مدعو: استاد ممتحن امضاء:

تاریخ:

## سپاس

از استاد بزرگوارم که با کمک‌ها و راهنمایی‌های بی‌دریغشان، مرا در به سرانجام رساندن این پایان‌نامه یاری داده‌اند، تشکر و قدردانی می‌کنم. همچنین از همکاران عزیزی که با راهنمایی‌های خود در بهبود نگارش این نوشتار سهیم بوده‌اند، صمیمانه سپاسگزارم.

## چکیده

با گسترش روزافزون استفاده از برنامه‌های اندرویدی در سالیان اخیر حملات موجود بر روی این سیستم عامل با افزایش قابل توجهی همراه بوده است. متن باز بودن برنامه‌های اندرویدی و در نتیجه، دسترسی به کد منبع این دسته از برنامه‌ها، در کنار افزایش حملات بر روی آن‌ها، لزوم توجه به مقابله با حملات مطروحه در این زمینه را افزایش داده است. حملات بازبسته‌بندی روی برنامه‌های اندرویدی، نوعی از حملات هستند که در آن مهاجم، پس از دسترسی به کد منبع برنامه و کپی کردن آن و یا ایجاد تغییراتی که مدنظر مهاجم است، مجدداً آن را بازبسته‌بندی می‌کند. تغییر کدهای برنامه، اهداف متفاوتی نظیر تغییر کتابخانه‌های تبلیغاتی، نقض امنیت کاربر و یا ضربه به شرکت‌های تولید برنامه از تغییر گسترش برنامه‌های جعلی را دنبال می‌کند. بازبسته‌بندی برنامه‌های اندرویدی علاوه بر ماهیت تهدید کاربران و شرکت‌ها، ماهیتی پیشگیرانه نیز دارد. در این حالت توسعه‌دهندگان نرم‌افزار از طریق ایجاد مبهم‌نگاری در برنامه‌های اندرویدی، سعی در پیشگیری از بازبسته‌بندی به وسیله‌ی مهاجمان دارند. تشخیص بازبسته‌بندی در برنامه‌های اندرویدی از آن جهت دارای اهمیت است که هم کاربران و هم شرکت‌های توسعه‌دهنده، می‌توانند از این موضوع ذی‌نفع باشند. تشخیص برنامه‌های بازبسته‌بندی شده، به جهت چالش‌های پیش‌رو، نظیر مبهم‌نگاری کدهای برنامه جعلی به دست مهاجم و همچنین تشخیص و جداسازی صحیح کدهای کتابخانه‌ای مسئله‌ای چالشی محسوب می‌شود. پژوهش‌های اخیر در این زمینه به صورت کلی، از روش‌های تشخیص مبتنی بر شباهت‌سنجی کدهای برنامه و یا طبقه‌بندی برنامه‌های موجود استفاده کرده‌اند. از طرفی برقراری حد واسطی میان سرعت و دقت در تشخیص برنامه‌های جعلی، چالشی است که استفاده از این دست پژوهش‌ها را در یک محیط صنعتی ناممکن ساخته است. در این پژوهش پس از استخراج کدهای برنامه به وسیله‌ی چارچوب سوت و ابزارهای دیس‌اسمبل، در یک روش دو مرحله‌ای کدهای برنامه‌های موجود با یکدیگر مقایسه می‌شود. پس از دیس‌اسمبل کدهای هر برنامه، در طی یک فرایند طبقه‌بندی مبتنی بر ویژگی‌های انتزاعی و دیداری، برنامه‌های کاندید برای هر برنامه مبدا استخراج می‌شود. سپس برای هر کلاس برنامه اندرویدی، امضایی متشکل از مهم‌تری ویژگی‌های کدپایه از آن استخراج و پس از انجام مقایسه با کلاس‌های کتابخانه‌های اندرویدی موجود در مخزن، کتابخانه‌های اندرویدی حذف می‌شوند و در نهایت با مقایسه‌ی کدهای اصلی، برنامه بازبسته‌بندی شده مشخص می‌شود. در قسمت آزمون روش پیشنهادی در این پژوهش، توانستیم روش موجود در این زمینه را با بهبود امضای تولیدشده از هر برنامه و اضافه‌شدن مرحله‌ی پیش‌پردازش، سرعت تشخیص را ۴ برابر افزایش داده و در عین حال دقت روش موجود را نیز حفظ کنیم.

**کلیدواژه‌ها:** پایان‌نامه، حروف چینی، قالب، زی‌پرشین

# فهرست مطالب

۱	مقدمه	۱
۷	مفاهیم اولیه	۲
۷	۱-۲ مبهم سازی	۷
۷	۲-۱-۱ روش های ساده	۷
۸	۲-۱-۲ روش های میانی	۸
۱۰	۲-۱-۳ روش های خاکستری	۱۰
۱۱	۲-۱-۴ روش های ترکیبی	۱۱
۱۱	۲-۱-۵ انواع مبهم نگارها	۱۱
۱۲	۲-۲ ساختار فایل های برنامه های اندرویدی	۱۲
۱۴	۲-۳ کتابخانه های اندرویدی	۱۴
۱۴	۲-۴ طبقه بندی	۱۴
۱۵	۲-۵ بازبسته بندی برنامه های اندرویدی	۱۵
۱۶	۳ تعریف مسئله و مرور کارهای پیشین	۱۶
۱۷	۳-۱ تعریف مسئله	۱۷
۱۸	۳-۲ روند کلی تشخیص برنامه های بازبسته بندی شده	۱۸
۱۸	۳-۲-۱ پیش پردازش برنامه های اندرویدی	۱۸
۱۹	۳-۲-۲ استخراج ویژگی	۱۹

۲۰	تشخیص بازبسته‌بندی ۳-۲-۳
۲۱	روش‌های تشخیص بازبسته‌بندی ۳-۳
۲۱	مبتنی بر تحلیل ایستا ۳-۳-۱
۳۵	مبتنی بر تحلیل پویا ۳-۳-۲
۳۸	سایر روش‌ها ۳-۳-۳
۳۹	پیش‌گیری از بازبسته‌بندی ۳-۴
۳۹	روش‌های مبتنی بر نشان‌گذاری ۳-۴-۱
۴۲	روش‌های مبتنی بر مبهم‌نگاری ۳-۴-۲
۴۵	مقایسه‌ی روش‌ها ۳-۵
۴۸	۴ راهکار پیشنهادی
۴۸	نمای کلی ۴-۱
۴۹	تشکیل امضای کلاس ۴-۲
۵۱	ساخت امضا ۴-۳
۵۲	امضای متد ۴-۳-۱
۵۳	امضای کلاس ۴-۳-۲
۵۴	تشخیص برنامه‌های بازبسته‌بندی شده ۴-۴
۵۴	مؤلفه‌ی تشخیص کدهای کتابخانه‌ای ۴-۴-۱
۵۹	مؤلفه‌ی یافتن نزدیک‌ترین همسایه ۴-۴-۲
۶۳	مقایسه‌ی دودویی و تشخیص برنامه‌های بازبسته‌بندی شده ۴-۴-۳
۶۷	۵ نتیجه‌گیری
۶۸	۶ نتیجه‌گیری
۶۹	مراجع



## فهرست جدول‌ها

۳-۱ ۱۰ منبع پربازدید در پژوهش [۱] ..... ۳۳



## فهرست شکل‌ها

- ۱-۲ نمونه‌ای از مبهم‌نگاری با استفاده از تغییر نام شناسه‌ها . . . . . ۸
- ۲-۲ نمونه‌ای از مبهم‌نگاری با استفاده از قابلیت بازتاب به منظور پنهان‌سازی واسطه فراخوانی شده  
به نام batteryinfo . . . . . ۱۰
- ۳-۲ ساختار پوشه‌ها و فایل‌های بسته‌های Apk [۲] . . . . . ۱۳
- ۱-۳ مراحل تشخیص برنامه‌های بازبسته‌بندی شده در پژوهش آقای آلدینی . . . . . ۲۷
- ۲-۳ هر گره از گراف در پژوهش نگویان شامل لیستی از واسطه‌های فراخوانی شده در آن  
فعالیت است. [۳] . . . . . ۲۸
- ۱-۴ شمای کلی امضای متد در پژوهش . . . . . ۵۲
- ۲-۴ شمای کلی امضای واسطه‌های برنامه‌نویسی . . . . . ۵۳
- ۳-۴ شمای کلی امضای کلاس . . . . . ۵۳
- ۴-۴ شمای کلی هسته‌ی کلاس . . . . . ۵۳
- ۵-۴ نمونه‌ای از تعریف یک فعالیت . . . . . ۶۰
- ۶-۴ نمونه‌ای از تعریف یک حق دسترسی . . . . . ۶۰

# فصل ۱

## مقدمه

سیستم عامل<sup>۱</sup> اندروید<sup>۲</sup> به دلیل سهولت در توسعه<sup>۳</sup> توسط توسعه دهندگان<sup>۴</sup> موبایلی و در نتیجه فراوانی استفاده از آن در تلفن های همراه، تلوزیون های هوشمند و دیگر دستگاه های موجود، حجم بالایی از بازار مصرفی سیستم عامل های موبایلی را به خود اختصاص داده است. بر طبق گزارش پایگاه استاتیسیتا<sup>۵</sup> [۴] سیستم عامل اندروید سهمی معادل ۷۱ درصدی از سیستم عامل های موبایلی را در سه ماهه ی پایانی سال ۲۰۲۲ به خود اختصاص داده است. در سال های اخیر به دلیل گسترش استفاده از این بستر<sup>۶</sup>، فروشگاه های اندرویدی زیادی به جهت ارائه ی خدمات به کاربران به وجود آمده است. برخی از فروشگاه های رسمی مانند فروشگاه اندرویدی گوگل<sup>۷</sup>، از ابزارهایی نظیر پلی پروتکت<sup>۸</sup> [۵] برای بررسی برنامه های اندرویدی موجود در فروشگاه استفاده می کنند. علاوه بر این، در سال های اخیر فروشگاه های متعدد رایگانی به وجود آمده اند که صرفاً برنامه های اندرویدی موجود در سطح وب را غربان<sup>۹</sup> و آن را به کاربران ارائه می دهند. فروشگاه های رایگان غالباً ابزارهای مشخصی را برای حفظ امنیت کاربران استفاده نمی کنند و امنیت کاربران این دسته از فروشگاه های اندرویدی، همواره تهدید می شود. یکی از راه های مورد استفاده توسط مهاجمان برای وارد ساختن بدافزار<sup>۱۰</sup> به تلفن های همراه، بازبسته بندی نرم افزار<sup>۱۱</sup> است. مطابق تعریف، بازبسته بندی شامل

---

Operation System<sup>۱</sup>

Android<sup>۲</sup>

Development<sup>۳</sup>

Developers<sup>۴</sup>

Statista<sup>۵</sup>

Platform<sup>۶</sup>

Google<sup>۷</sup>

Play Protect<sup>۸</sup>

Crawl<sup>۹</sup>

Malware<sup>۱۰</sup>

Software Repackaging<sup>۱۱</sup>

دانلود<sup>۱۲</sup> یک برنامه، دسترسی به محتوای کدهای برنامه اصلی از طریق روش‌های مهندسی معکوس<sup>۱۳</sup> و در نهایت بازبسته‌بندی به همراه تغییر و یا بدون تغییر دادن کدهای برنامه اصلی<sup>۱۴</sup> است. زبان اصلی توسعه در برنامه‌های اندرویدی، زبان جاوا<sup>۱۵</sup> می‌باشد که یک زبان سطح بالا<sup>۱۶</sup> محسوب می‌شود. در طی فرآیند کامپایل<sup>۱۷</sup> برنامه‌های اندرویدی، مجموعه‌ی کدهای منبع در طی فرایندی به بایت‌کدهای دالویک<sup>۱۸</sup> تبدیل می‌شوند و در ادامه ماشین مجازی جاوا<sup>۱۹</sup>، بایت‌کدها را بر روی ماشین مقصد اجرا می‌کند [۶]. فهم و در نتیجه مهندسی معکوس زبان میانی دالویک بایت‌کدها آسان است و به همین علت موجب سهولت در بازبسته‌بندی برنامه‌های اندرویدی می‌شود.

به طور کلی بازبسته‌بندی را می‌توان از دو جهت مورد بررسی قرار داد، از دید توسعه‌دهندگان، بازبسته‌بندی شامل فرایندی است که توسعه‌دهنده با انجام مبهم‌نگاری<sup>۲۰</sup> در برنامه مورد توسعه، فهم بدنه‌ی اصلی برنامه را برای مهاجم<sup>۲۱</sup> سخت می‌کند. از این دید، بازبسته‌بندی یک روش تدافعی تلقی می‌شود تا مهاجم پس از دسترسی به کد برنامه اصلی، نتواند بدنه‌ی برنامه اصلی برنامه را شناسایی و در نتیجه آن را تغییر دهد. از جهت دیگر، بازبسته‌بندی توسط فردی که برنامه متعلق به او نیست یک عمل تهاجمی محسوب می‌شود. در این حالت، مهاجم پس از دسترسی به کد برنامه اصلی، بسته به هدف او، برنامه را مجدداً بازبسته‌بندی می‌کند و آن را در فروشگاه‌های اندرویدی خصوصاً فروشگاه‌هایی که نظارت کمتری بر روی آن‌ها وجود دارد منتشر می‌کند. در دیدگاه تهاجمی، مهاجم به جهت اهدافی متفاوتی نظیر تغییر کدهای تبلیغاتی<sup>۲۲</sup> در برنامه اصلی، تغییر درگاه‌های پرداخت و یا بازپخش بدافزار، اقدام به بازبسته‌بندی می‌کند. بازبسته‌بندی یکی از راه‌های محبوب مهاجمان برای انتقال بدافزارهای توسعه‌داده‌شده به تلفن همراه قربانی است [۷]. مطابق پژوهش آقای ژو و همکاران [۸] حدود ۸۵ درصد بدافزارهای موجود، از طریق بازبسته‌بندی منتشر می‌شوند. همانطور که گفته شد، برخی فروشگاه‌های اندرویدی نظیر گوگل، سازوکار مشخصی را برای تشخیص<sup>۲۳</sup> بازبسته‌بندی ارائه‌داده‌اند اما بسیاری از فروشگاه‌های اندرویدی فعال و پربازدید، خصوصاً فروشگاه‌های رایگان، یا از هیچ ابزاری استفاده نمی‌کنند و یا در صورت توسعه‌ی نرم‌افزار بومی<sup>۲۴</sup> خود برای شناسایی برنامه‌های بازبسته‌بندی شده، مشخصات و یا دقت آن را گزارش نکرده‌اند.

همانطور که اشاره شد، به دلیل محبوبیت و در نهایت استفاده‌ی زیاد برنامه‌های اندرویدی و همچنین

- Download<sup>۱۲</sup>
- Reverse Engineering<sup>۱۳</sup>
- Orginal Application<sup>۱۴</sup>
- Java<sup>۱۵</sup>
- High Level<sup>۱۶</sup>
- Compile<sup>۱۷</sup>
- Dalvic Byte Code<sup>۱۸</sup>
- Java Virtual Machine<sup>۱۹</sup>
- Obfuscation<sup>۲۰</sup>
- Attacker<sup>۲۱</sup>
- Ad Code<sup>۲۲</sup>
- Detection<sup>۲۳</sup>
- Native<sup>۲۴</sup>

نظارت کم در فروشگاه‌های مرتبط، بازبسته‌بندی یک روش پر استفاده به جهت انتقال بدافزار به تلفن همراه کاربران است. آقای خانمحمدی و همکاران [۹]، پس از بررسی برنامه‌های اندرویدی مجموعه داده‌ی<sup>۲۵</sup> اندروزو<sup>۲۶</sup>، دریافتند که ۵۲/۲۲٪ از برنامه‌های موجود در این مخزن توسط ویروس‌توتال<sup>۲۷</sup>، بدافزار شناسایی شده‌اند. ویروس‌توتال، ابزاری متشکل از ۳۰ ضدبدافزار برای بررسی یک برنامه اندرویدی است. مطابق این پژوهش، ۷۷/۸۴٪ از برنامه‌های این مجموعه داده که بازبسته‌بندی شده‌اند، دارای نوعی از بدافزار ضدتبلیغاتی<sup>۲۸</sup> بوده‌اند که موجب می‌شود تبلیغات موجود در برنامه تغییر کرده و اهداف مالی و امنیتی کاربران و توسعه‌دهندگان مخدوش شود. علاوه بر این، مطابق پژوهشی که توسط ویداس و همکاران [۱۰] انجام شده‌است، پس از پیاده‌سازی ۷ روش پربازدید به جهت تشخیص بازبسته‌بندی، در بهترین حالت، روش‌های موجود قادر به تشخیص ۷۲/۲۲٪ از برنامه‌های بازبسته‌بندی شده‌ی سه فروشگاه مطرح اندرویدی بوده‌اند. بنابراین مشخص است که تشخیص برنامه‌های بازبسته‌بندی شده، به چه میزان می‌تواند اهداف مالی و امنیتی توسعه‌دهندگان و کاربران برنامه‌ها را ارضا کند. در سال‌های اخیر ارائه‌ی یک راهکار پرسرعت به همراه دقت مناسب، همواره یکی از دغدغه‌های مهم پژوهش‌کنندگان در این زمینه بوده‌است.

همانطور که گفته‌شد، بازبسته‌بندی برنامه‌های اندرویدی از دو دیدگاه تهاجمی و تدافعی قابل بررسی است. در حالتی که کاربر متقلب، برنامه اندرویدی اصلی را دچار تغییراتی می‌کند و آن را در اختیار عموم قرار می‌دهد، تشخیص بازبسته‌بندی، با استفاده از مقایسه‌ی برنامه اصلی و برنامه جعلی صورت می‌گیرد. تشخیص بازبسته‌بندی در این حالت را می‌توان به صورت کلی به دو طبقه تقسیم کرد. در حالت اول توسعه‌دهنده روش خود را مبتنی بر تحلیل برنامه مبدا و مقصد پیاده‌سازی می‌کند. عمده‌ی روش‌های موجود در این طبقه مبتنی بر تحلیل ایستا<sup>۲۹</sup>ی جفت برنامه‌ها است و استفاده از تحلیل پویا<sup>۳۰</sup> به جهت سرعت پایین آن، محبوبیت فراوانی ندارد [۱۱]. در سمت دیگر طبقه‌بندی<sup>۳۱</sup> برنامه‌های اندرویدی وجود دارد. روش‌های موجود در این دسته، عمدتاً سرعت بالایی دارند اما در تشخیص جفت بازبسته‌بندی شده دقت پایینی را ارائه می‌دهند.

برنامه‌های اندرویدی متشکل از دو قسمت اصلی کدهای برنامه و منابع<sup>۳۲</sup> هستند. کدهای برنامه، منطق<sup>۳۳</sup> برنامه را تشکیل می‌دهند و رفتار برنامه با توجه به این قسمت مشخص می‌شود. از طرفی منابع

- Data Set<sup>۲۵</sup>
- Androzoo<sup>۲۶</sup>
- Virus total<sup>۲۷</sup>
- AdWare<sup>۲۸</sup>
- Static<sup>۲۹</sup>
- Dynamic<sup>۳۰</sup>
- Classification<sup>۳۱</sup>
- Resources<sup>۳۲</sup>
- Logic<sup>۳۳</sup>

برنامک، رابط کاربری<sup>۳۴</sup> آن را تشکیل می‌دهند. روش‌های مبتنی بر تحلیل برنامک و یا طبقه‌بندی آن، عمدتاً از ویژگی‌های موجود در منابع و یا کد آن استفاده می‌کنند. مهاجم در حالتی که می‌خواهد از محبوبیت برنامک مبدا استفاده کند، سعی در یکسان‌سازی ظاهر برنامک‌های مبدا و مقصد دارد به همین جهت از منابع برنامک مبدا استفاده می‌کند و منطق برنامک را مطابق با اهداف خود تغییر می‌دهد. در حالتی دیگر، متقلب سعی می‌کند که با استفاده از تغییر منابع برنامک و تولید یک برنامک تقلبی و گاهی بدون هیچ تغییری در کد برنامک، ادعای توسعه‌ی یک برنامک جدید را اثبات کند. لازم به ذکر است استفاده از ویژگی‌های کدپایه<sup>۳۵</sup> و منبع‌پایه<sup>۳۶</sup>، به وفور در پژوهش‌های سال‌های اخیر یافت می‌شود که هر کدام معایب و مزایای خود را دارد.

در روش‌های مبتنی بر طبقه‌بندی عمدتاً تعریف تشخیص بازبسته‌بندی محدود به تشخیص دسته‌ی مشکوک که احتمال بازبسته‌بندی بودن جفت‌های داخل این دسته، بیش از سایر دسته‌ها است. بنابراین تشخیص بازبسته‌بندی در این روش‌ها، محدود به تشخیص طبقه‌ی برنامک ورودی می‌باشد و جفت بازبسته‌بندی شده مشخص نمی‌شود. از طرفی در روش‌های مبتنی بر تحلیل برنامک، بررسی دوبه‌دوی برنامک‌های ورودی و مجموعه‌داده مدنظر است. در این روش‌ها تعریف تشخیص بازبسته‌بندی گسترش یافته و یافتن جفت بازبسته‌بندی به صورت مشخص، از اهداف اصلی پژوهش است. تغییر منابع برنامک و همچنین مبهم‌نگاری در برنامک بازبسته‌بندی شده، دوجالش مهم در راستای تشخیص بازبسته‌بندی است. متقلب پس از بازبسته‌بندی برنامک، با استفاده از مبهم‌نگاری سعی می‌کند تغییرات خود و شباهت ساختار منطقی برنامک تقلبی با برنامک اصلی را پنهان کند. به همین جهت، تشخیص بازبسته‌بندی نیازمند ویژگی‌هایی است که مقاومت بالایی مقابل مبهم‌نگاری داشته‌باشد بدین معنا که تغییر و ایجاد ابهام در کد، به راحتی در این ویژگی‌ها قابل انجام نباشد.

در هنگام کامپایل برنامک‌های اندرویدی، کتابخانه‌ها<sup>۳۷</sup>یی که در برنامک مورد استفاده قرار گرفته‌اند به همراه کد مورد توسعه، کامپایل شده و دالویک بایت‌کدهای آن در کنار برنامک قرار می‌گیرد. بر اساس پژوهش آقای زیانگ و همکاران [۱۲] ۵۷٪ از کدهای برنامک‌های مورد بررسی در این پژوهش، شامل کدهای کتابخانه‌ای بودند که دچار مبهم‌نگاری نشده‌اند. بنابراین تشخیص کدهای بازبسته‌بندی شده بدون تشخیص درست و دقیق و جداسازی کدهای کتابخانه‌ای امکان‌پذیر نیست و در صورتی که به درستی جداسازی صورت گیرد، می‌تواند نتایج منفی غلط و مثبت غلط را کاهش دهد. به صورت کلی دو روش برای تشخیص کدهای کتابخانه‌ای استفاده می‌شود، روش مبتنی بر لیست سفید<sup>۳۸</sup> و یا روش تشخیص

<sup>۳۴</sup>User Interface

<sup>۳۵</sup>Code Base

<sup>۳۶</sup>Resource Base

<sup>۳۷</sup>Library

<sup>۳۸</sup>White List

مبتنی بر شباهت سنجی<sup>۳۹</sup>. در روش لیست سفید، لیستی از مشهورترین کتابخانه‌های موجود در مخازن کتابخانه‌ای اندروید نظیر ماون<sup>۴۰</sup> جمع آوری می‌شود و با استفاده از نام کلاس‌ها و بسته‌های موجود، کلاس‌های کتابخانه‌ای تشخیص داده می‌شود. مشخص است که این روش مقاومت بسیار کمی مقابل ساده‌ترین روش‌های مبهم‌نگاری در کتابخانه‌های اندرویدی دارد. در حالت دیگر از روش‌های مبتنی بر شباهت سنجی برای تشخیص کدهای کتابخانه‌ای استفاده می‌شود که در این روش، تحلیل ایستا روی کدهای برنامه‌ی مبدا و مخزن کتابخانه‌های اندروید صورت می‌گیرد و در نهایت از طریق شباهت‌سنجی، کدهای کتابخانه‌ای تشخیص داده می‌شوند. مشخص است که روش‌های مبتنی بر شباهت‌سنجی از دقت بیشتری، خصوصاً در صورت وجود ابهام، برخوردار هستند و تمایز بهتری میان کدهای کتابخانه‌ای و کدهای اصلی قرار می‌دهند اما اینگونه روش‌ها سرعت پایینی دارند.

پژوهش‌های ارائه‌شده در زمینه‌ی تشخیص برنامه‌های بازبسته‌بندی شده در سال‌های اخیر، عمدتاً در تلاش برای بهبود دقت و سرعت روش‌های پیشین بوده‌اند. مبهم‌نگاری باعث می‌شود که دقت روش‌های تشخیص مبتنی بر تحلیل ایستا و شباهت سنجی پایین بیاید و استفاده از ویژگی‌هایی را که مقاومت بالایی مقابل مبهم‌نگاری داشته باشند واجب کند. از طرفی استفاده از ویژگی‌های مقاوم به مبهم‌نگاری، می‌تواند سرعت تشخیص را بسیار پایین آورده تا حدی که عملاً استفاده از این روش‌ها در یک محیط صنعتی را غیر ممکن سازد. در این پژوهش ما با استفاده از ترکیب روش‌های تحلیل ایستا و طبقه‌بندی منابع، به همراه شباهت‌سنجی، روشی را ارائه کرده‌ایم که در حالی که مقاومت بالایی نسبت به مبهم‌نگاری داشته باشد، سرعت روش‌های پیشین را نیز افزایش دهد. در این پژوهش به عنوان پیش‌پردازش، از یک طبقه‌بند<sup>۴۱</sup> نزدیک‌ترین همسایه<sup>۴۲</sup> برای کاهش فضای مقایسه‌ی دودویی<sup>۴۳</sup> و با استفاده از ویژگی‌های مبتنی بر منبع، سرعت تشخیص بهبود داده شده است. با کاهش فضای مقایسه‌ی دودویی و طبقه‌بندی برنامه‌های مشکوک در یک دسته، مقایسه‌ی برنامه‌های موجود در آن دسته آغاز می‌شود. مقایسه‌ی دودویی در هر دسته مبتنی بر تحلیل ایستا و شباهت‌سنجی کدهای برنامه‌ی انجام می‌شود. ابتدا ویژگی‌هایی از هر کلاس و متد<sup>۴۴</sup> در بسته‌های برنامه‌ی استخراج شده و امضا<sup>۴۵</sup>ی هر کلاس ساخته می‌شود به طوری که امضای هر کلاس منحصر به فرد و تا حد امکان مختص همان کلاس باشد. نوآوری روش مطروحه، ترکیب روش‌های مبتنی بر طبقه‌بندی و روش‌های مبتنی بر تحلیل ایستا می‌باشد که در نهایت منجر به افزایش سرعت و در عین حال دقت خوب در تشخیص برنامه‌های بازبسته‌بندی شده است. حذف کدهای کتابخانه‌ای با استفاده از روشی مبتنی بر مقایسه‌ی کدهای موجود در مخزن کتابخانه‌ها و کلاس‌های برنامه‌ی انجام می‌شود. مخزن

Similarity<sup>۳۹</sup>

Maven Repository<sup>۴۰</sup>

Classifier<sup>۴۱</sup>

Nearest Neighbor<sup>۴۲</sup>

Pairwise Comparison<sup>۴۳</sup>

Method<sup>۴۴</sup>

Signature<sup>۴۵</sup>

کتابخانه‌ها متشکل از ۴۵۳ کتابخانه‌ی اندرویدی جمع‌آوری شده از مخزن ماون می‌باشد. در نهایت پس از تشخیص کلاس‌های کتابخانه‌های اندرویدی و حذف آن‌ها از کد برنامه، کدهای مورد توسعه به عنوان ورودی برای مقایسه‌ی دودویی و طبقه‌بندی مورد تحلیل قرار می‌گیرند.

در ادامه‌ی این نگارش، در فصل ۲ به تعریف مفاهیم اولیه مورد نیاز این پژوهش می‌پردازیم. در فصل ۳ به تعریف مسئله می‌پردازیم و همچنین مروری از کارهای پیشین را خواهیم داشت. در ادامه و در فصل ۴ روش مورد استفاده در این پژوهش، شرح داده خواهد شد و در فصل ۵ مقایسه و ارزیابی روش پیشنهادی خود را ارائه می‌دهیم. در نهایت و در فصل ۶ ضمن جمع‌بندی این گزارش علمی، به بررسی نقاط ضعف و قوت این پژوهش و همچنین ارائه‌ی پیشنهاداتی جهت بهبود آن خواهیم پرداخت.

## فصل ۲

# مفاهیم اولیه

در این فصل مفاهیمی را که به صورت مستقیم و غیرمستقیم در این پژوهش از آنها استفاده شده است را شرح می‌دهیم. آشنایی با مفاهیم مطروحه در این فصل، منجر به درک بهتر پژوهش و راه‌حل پیشنهادی در فصل ۴ خواهد شد.

## ۱-۲ مبهم‌سازی

آن‌چنان که در فصل پیشین گفته شد، مبهم‌سازی را می‌توان از دو دیدگاه تهاجمی و تدافعی بررسی کرد. در این قسمت ما با توجه به هدف پژوهش که تشخیص بازبسته‌بندی به جهت دفاع می‌باشد، مبهم‌سازی را فرایندی در نظر می‌گیریم که در آن فرد مهاجم یا به بیان دیگر متقلب، برنامه‌ک اصلی را دانلود کرده و پس از دیکامپایل<sup>۱</sup> کردن، به نوعی تغییر می‌دهد که منطق کلی برنامه، تغییری نمی‌کند. مبهم‌سازی یکی از ارکان اصلی در فرایند بازبسته‌بندی است و هدف اصلی آن این است که ابزارهای تشخیص بازبسته‌بندی، خصوصاً در مواردی که از تحلیل ایستا استفاده می‌کنند را به اشتباه بیاندازد. روش‌های مبهم‌سازی را از نظر میزان سختی در تشخیص به ۳ دسته‌ی کلی می‌توان تقسیم کرد[۱۳]:

## ۱-۱-۲ روش‌های ساده

راهکارهای موجود در این دسته عمدتاً بدون تغییر در برنامه‌ک اصلی رخ می‌دهد. در این روش متقلب پس از آن‌که به کدهای برنامه‌ک اصلی دسترسی پیدا کرد، آن را بدون هیچ گونه تغییری کامپایل و بسته‌بندی می‌کند.

---

<sup>۱</sup>Decompile



باز بسته‌بندی تنها موجب تغییر در امضاء توسعه‌دهنده‌ی برنامه‌ی و جمع‌آزمای<sup>۲</sup> می‌شود. بنابراین روش‌هایی که مبتنی بر این دو خصوصیت هستند در این سطح دچار مشکل می‌شوند.

## ۲-۱-۲ روش‌های میانی

این دسته از روش‌های مبهم‌سازی، شامل روش‌هایی است که در آن بیشتر ویژگی‌های مبتنی بر معناسازی<sup>۳</sup> تغییر می‌کند و ویژگی‌های مبتنی بر نحو<sup>۴</sup> ثابت باقی می‌مانند. بنابراین، روش‌هایی که بیشتر مبتنی بر معناسازی برنامه‌های اندرویدی هستند، دچار خطای بیشتری در این سطح از مبهم‌نگاری می‌شوند. در ادامه به معرفی مختصری از انواع روش‌های مبهم‌نگاری مطابق با پژوهش [۱۴] در این دسته می‌پردازیم:

- **تغییر نام شناسه‌ها<sup>۵</sup>:** تغییر نام شناسه‌های موجود در برنامه شامل نام کلاس‌ها، متدها و یا متغیرها<sup>۶</sup>ی موجود [۱۳]

```
1 public class a{
2     private Integer a;
3     private Float b;
4     public void a(Integer a, Float b){
5         this.a = a + Integer.valueOf(b)
6     }
7 }
```

شکل ۲-۱: نمونه‌ای از مبهم‌نگاری با استفاده از تغییر نام شناسه‌ها

- **تغییر نام بسته:** در این روش مبهم‌نگاری با استفاده از تغییر نام بسته‌های برنامه صورت می‌گیرد.
- **رمزنگاری رشته‌ها:** استفاده از رمزنگاری در رشته‌های<sup>۷</sup> مورد استفاده از فایل‌های دکس<sup>۸</sup>، موجب کاهش سطح معناسازی می‌شود.
- **فراخوانی غیرمستقیم:** یکی از روش‌های ساده‌ی تغییر گراف فراخوانی<sup>۹</sup>، استفاده از یک تابع واسط به عنوان تابع فراخوانده‌ی<sup>۱۰</sup> تابع اصلی است. در این حالت تابع اولیه یک تابع واسط و تابع واسط به صورت زنجیره‌ای تابع اصلی را فراخوانی می‌کند. بدنه‌ی تابع واسط در این حالت، بسیار ساده و شامل یک فراخوانی تابع اصلی است.

---

Checksum<sup>۲</sup>  
Semantic<sup>۳</sup>  
Syntax<sup>۴</sup>  
Identifier<sup>۵</sup>  
Variable<sup>۶</sup>  
String<sup>۷</sup>  
Dex Files<sup>۸</sup>  
Call Graph<sup>۹</sup>  
Caller<sup>۱۰</sup>

فقره جابه‌جایی دستورات: جابه‌جایی دستورات موجود در برنامه اصلی، یکی از روش‌های پرکاربرد توسط ابزارهای مبهم‌نگاری است. جابه‌جایی دستورات به شکلی انجام می‌شود که استقلال هر قسمت حفظ گردد.

- **جابه‌جایی ساختار سلسله‌مراتبی:** در این روش، ساختار سلسله‌مراتبی کلاس‌های برنامه به نوعی تغییر می‌کند که منطق کلاس‌ها دچار تغییر نشود.
- **ادغام و شکستن:** می‌توان توابع و یا کلاس‌های موجود در برنامه‌های اندرویدی را ادغام کرد. برای مثال می‌توان هر جایی که یک تابع صدا زده شده بود، فراخوانی تابع با بدنه‌ی تابع جایگزین شود. از طرفی می‌توان بدنه‌ی چند تابع را تحت یک تابع با یکدیگر ادغام کرد. این کار ساختار توابع فراخواننده را نیز تغییر می‌دهد. از طرفی می‌توان یک تابع را به چندین تابع مشخص شکست و بدین صورت گراف جریان برنامه را تغییر داد.
- **وارد ساختن کدهای بیهوده:** کدهای بیهوده، کدهایی هستند که اجرا می‌شوند ولی تاثیری در ادامه‌ی روند اجرایی برنامه، ندارند. کدهای بیهوده عموماً دارای ساختارهای کنترلی<sup>۱۱</sup> و حلقه‌های نپ<sup>۱۲</sup> هستند که تاثیری در روند اجرای برنامه ندارند. ذکر این نکته حائز اهمیت است که در صورتی که در ساختار کدهای بیهوده از شروط کنترلی مبتنی بر متغیرهای پویا<sup>۱۳</sup> استفاده شود آنگاه دیگر تحلیل ایستای برنامه‌های اندرویدی قادر به تشخیص این نوع از مبهم‌نگاری‌ها نیست.
- **وارد ساختن کدهای مرده:** یکی دیگر از روش‌های تغییر گراف‌های برنامه از جمله گراف جریان<sup>۱۴</sup>، اضافه‌کردن کدهای مرده‌ای است که در ساختار گراف جریان برنامه‌های اندرویدی هیچ‌گاه اجرا نمی‌شوند اما به عنوان یک گره در گراف حضور دارند.
- **روش‌های دیگر:** روش‌های دیگری نظیر تغییر نام منابع مورد استفاده در برنامه‌های اندرویدی و حذف فایل اشکال‌زدایی<sup>۱۵</sup> از روش‌های دیگری است که در این سطح به وفور مورد استفاده قرار می‌گیرد.

---

<sup>۱۱</sup> Control's Statement

<sup>۱۲</sup> Nop

<sup>۱۳</sup> Dynamic Variables

<sup>۱۴</sup> Flow Graph

<sup>۱۵</sup> Debug File

## ۳-۱-۲ روش‌های خاکستری

روش‌های موجود در این دسته، مبتنی بر نحوه برنامه‌های اندرویدی و خصوصاً زبان جاوا به وجود آمده‌است. عمده‌ی روش‌های مورد استفاده در این سطح، از خصوصیات مهم زبان جاوا به عنوان زبان اصلی در پیاده‌سازی برنامه‌های اندرویدی، استفاده می‌کنند. در ادامه به بررسی مهم‌ترین روش‌های موجود در این دسته می‌پردازیم.

- **بازتاب<sup>۱۶</sup>**: بازتاب یکی از ویژگی‌های مهم و پیچیده‌ی زبان جاوا می‌باشد [۱۵] که امکان فراخوانی متدها و ارتباط با کلاس‌های برنامه را به صورت پویا فراهم می‌سازد. مهاجمان با استفاده از فراخوانی متدها به وسیله‌ی قابلیت بازتاب، می‌توانند نام واسط فراخوانی‌شده را پنهان سازند و بدین وسیله سطح جدیدی از مبهم‌نگاری را در برنامه‌های اندرویدی ایجاد سازند. استفاده از قابلیت بازتاب و رمزنگاری<sup>۱۷</sup> رشته‌ی واسط مورد نظر، به طور کامل واسط فراخوانی‌شده را پنهان می‌کند.

```
1 Object object = new Object();
2 Method getService = Class.forName("android.os.
  ServiceManager").getMethod("getService",
  String.class);
3 Object obj = getService.invoke(object, new
  Object[]{new String("batteryinfo")});
```

شکل ۲-۲: نمونه‌ای از مبهم‌نگاری با استفاده از قابلیت بازتاب به منظور پنهان‌سازی واسط فراخوانی‌شده به نام **batteryinfo**

- **رمزنگاری دالویک بایت‌کدها**: در این روش، مهاجم در حین ساختن برنامه بازبسته‌بندی شده، قسمتی مهمی از کدهای برنامه را رمزنگاری کرده و در هنگام اجرا با استفاده از یک رویه‌ی رمزگشایی<sup>۱۸</sup>، کدهای اصلی را بارگیری<sup>۱۹</sup> می‌کند. این روش عمدتاً زمانی استفاده می‌شود که مهاجم نیاز به فراخوانی توابع واسط‌های برنامه‌نویسی داشته‌باشد و قسمتی را که واسط‌ها فراخوانی می‌شوند را رمزنگاری می‌کند.

- **بارگذاری پویای کلاس‌ها<sup>۲۰</sup>**: زبان جاوا از قابلیت مهمی به نام بارگیری پویای کد پشتیبانی می‌کند که اجازه می‌دهد تکه کدی را که پیش از این در کد مورد توسعه‌ی یک برنامه موجود نبوده را در حین اجرا به برنامه اضافه کنیم. مهاجم با استفاده از این قابلیت زبان جاوا می‌تواند قسمت‌هایی

<sup>۱۶</sup> Reflect  
<sup>۱۷</sup> Encryption  
<sup>۱۸</sup> Decryption  
<sup>۱۹</sup> Load  
<sup>۲۰</sup> Dynamic Class Loading

از برنامه‌ک را در حین اجرای آن تغییر دهد که عملاً تشخیص آن‌ها با استفاده از تحلیل‌های ایستا امکان‌پذیر نیست.

## ۴-۱-۲ روش‌های ترکیبی

هر ترکیبی از روش‌های گفته‌شده در سطوح مختلف را می‌توان برای مبهم‌نگاری استفاده کرد. به صورت کلی روش‌های میانی ۲-۱-۲ و روش‌های خاکستری ۳-۱-۲ را می‌توان دو دسته‌ی مهم از انواع مبهم‌نگاری به حساب آورد که به صورت گسترده در مبهم‌نگارهای رایگان و یا تجاری مورد استفاده قرار می‌گیرد.

## ۵-۱-۲ انواع مبهم‌نگارها

در قسمت پیشین، دریافتیم که مبهم‌نگاری، سطوح متفاوتی دارد که متقلبان برای تولید برنامه‌های بازبسته‌بندی شده از آن‌ها استفاده می‌کنند. از آنجایی که بسیاری از برنامه‌های تقلبی با استفاده از مبهم‌نگار<sup>۲۱</sup>ها توسعه یافته‌اند و علاوه بر این برای ابداع یک روش مفید جهت تشخیص برنامه‌های بازبسته‌بندی شده ابتدا باید انواع مبهم‌نگارهای موجود را بررسی کرد. در پژوهشی که توسط ژانگ و همکاران [۱۶] انجام شده، ۴۳٪ از برنامه‌های بازبسته‌بندی شده‌ی مورد بررسی در این پژوهش، از مبهم‌نگاری‌های بسیار ساده‌ای نظیر تغییر نام و با استفاده از مبهم‌نگارهای رایگان، انجام شده‌است. در ادامه به بررسی چند مبهم‌نگار رایگان و تجاری<sup>۲۲</sup> می‌پردازیم.

### • پروگارد

پروگارد<sup>۲۳</sup> یک نرم‌افزار متن‌باز رایگان به جهت بهینه‌سازی و مبهم‌نگاری در برنامه‌های جاوا مورد استفاده قرار می‌گیرد. بهینه‌سازی از طریق حذف کدهای مرده و منابع بلااستفاده انجام می‌شود و مبهم‌نگاری عمدتاً با استفاده از روش‌های مشروحه در بخش ۳-۱-۲ انجام می‌شود. [۱۷]

### • آلاتوری

آلاتوری<sup>۲۴</sup> یک مبهم‌نگار رایگان تولیدشده توسط شرکت روسی اسماردک<sup>۲۵</sup> می‌باشد که سطوح مختلفی از مبهم‌نگاری را با توجه به فایل‌های پیکربندی<sup>۲۶</sup> پوشش می‌دهد. این مبهم‌نگار از تغییرنام،

<sup>۲۱</sup> Obfuscator

<sup>۲۲</sup> Commercial

<sup>۲۳</sup> Proguard

<sup>۲۴</sup> Allatori

<sup>۲۵</sup> Smardec

<sup>۲۶</sup> Configuration

مبهم‌نگاری مبتنی بر تغییر گراف‌های جریان، مبهم‌نگاری فایل‌های اشکال‌زدایی و رمزنگاری داده‌های رشته‌ای پشتیبانی می‌کند. [۱۸، ۱۹]

#### • دکس‌گارد

این مبهم‌نگار نسخه‌ی تجاری نرم‌افزار پروگارد است که توسط شرکت گارداسکووار<sup>۲۷</sup> تولید شده‌است. دکس‌گارد<sup>۲۸</sup> را می‌توان مشهورترین و یکی از پیچیده‌ترین مبهم‌نگارهای موجود به حساب آورد. آخرین نسخه‌ی این نرم‌افزار انواع مبهم‌نگاری‌های سطح خاکستری نظیر بارگیری پویای کد و همچنین رمزنگاری کلاس‌ها و توابع را به صورت کامل انجام می‌دهد.

## ۲-۲ ساختار فایل‌های برنامه‌های اندرویدی

هر برنامه‌ی اندرویدی یک فایل فشرده‌شده با پسوند APK<sup>۲۹</sup> است که به اختصار شامل ۴ پوشه‌ی مهم و ۳ فایل می‌باشد. برای درک بهتر از روش پیشنهادی این پژوهش، در ادامه هر کدام از این قسمت‌ها را معرفی و کارکرد آن را بررسی خواهیم کرد. [۲۰]

• **پوشه‌ی res:** این پوشه شامل منابع برنامه‌های اندرویدی است که مربوط به رابط کاربری برنامه می‌شود. این پوشه در نهایت به فایل‌های R. نگاشت شده و هر کدام از منابع با یک شناسه<sup>۳۰</sup> مشخص می‌گردد.

• **پوشه‌ی lib:** فایل‌های کامپایل‌شده‌ی بومی در این پوشه قرار می‌گیرند که شامل کتابخانه‌های اندرویدی و جاوایینیز می‌گردد. استفاده از فایل‌هایی که کامپایل شده‌اند سرعت اجرای برنامه‌های اندرویدی را بالا می‌برد لذا استفاده از آن‌ها به عنوان بسته‌های<sup>۳۱</sup> از پیش آماده محبوبیت دارد.

• **فایل Classes.dex:** فایل‌های با پسوند dex فایل‌های دودویی<sup>۳۲</sup> هستند که اطلاعات را در سطر و ستون‌های خود ذخیره می‌کنند. این فایل در برنامه‌های اندرویدی حاوی بایت‌کدهای دالویک است که توسط ماشین مجازی دالویک<sup>۳۳</sup> اجرا می‌شود.

---

<sup>۲۷</sup> Guardsquare

<sup>۲۸</sup> DexGuard

<sup>۲۹</sup> Android Package

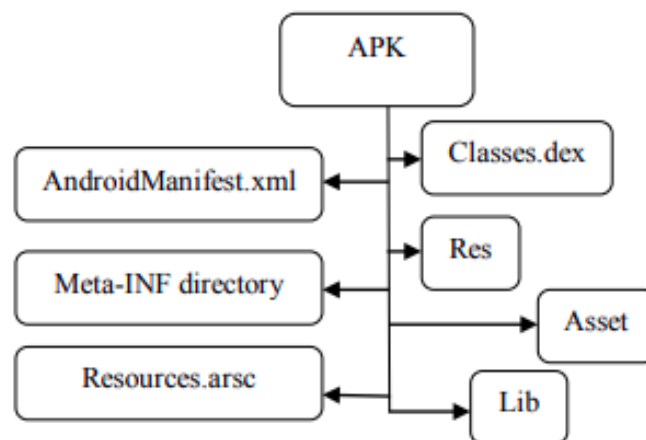
<sup>۳۰</sup> Resource Id

<sup>۳۱</sup> Module

<sup>۳۲</sup> Binary

<sup>۳۳</sup> Dalvik Virtual Machine

- فایل **AndroidManifest.xml**: پیکربندی‌های مهم فایل‌های APK از جمله لیست مجوزهای مورد نیاز، لیست مولفه‌ها<sup>۳۴</sup> و نام بسته‌ی برنامه در این فایل نوشته می‌شود.
- پوشه‌ی **assets**: این پوشه همانند پوشه‌ی res برای منابع ایستا مورد استفاده قرار می‌گیرد با این تفاوت همه توسعه‌دهندگان در این پوشه می‌توانند عمق زیرپوشه‌ها را به تعداد نامتناهی افزایش دهند تا ساختار بهتری را فراهم سازند.
- پوشه‌ی **META-INF**: این پوشه شامل اطلاعات کلیدهای عمومی<sup>۳۵</sup> کاربر توسعه‌دهنده‌ی برنامه است که برنامه با کلید خصوصی متناظر آن امضا شده‌است. امضای موجود در این پوشه، خاصیت صحت‌سنجی<sup>۳۶</sup> دارد اما اطلاعاتی را از توسعه‌دهنده نشر نمی‌دهد و به صورت خودامضا ساخته می‌شود.
- فایل **resources.arsc**: این فایل برای انجام نگاشت<sup>۳۷</sup> میان منابع موجود در پوشه‌ی resources و شناسه‌ی هر منبع استفاده می‌شود تا بتوان در حین اجرای برنامه‌ها، هر شناسه را به منبع آن ترجمه کرد.



شکل ۲-۳: ساختار پوشه‌ها و فایل‌های بسته‌های Apk [۲]

<sup>۳۴</sup>Component  
<sup>۳۵</sup>Public Key  
<sup>۳۶</sup>Integrity  
<sup>۳۷</sup>Mapping

## ۳-۲ کتابخانه‌های اندرویدی

کتابخانه‌های اندرویدی، بسته‌های از پیش توسعه‌یافته هستند که توسط توسعه‌دهندگان نوشته شده و توسعه‌دهندگان اندروید به جهت سهولت در پیاده‌سازی و کمک به تسریع توسعه‌ی نرم‌افزار به وفور از این نمونه‌ها استفاده می‌کنند. کتابخانه‌های اندرویدی به صورت کلی به دو بخش کتابخانه‌های مختص برنامه‌نویسی اندرویدی و کتابخانه‌های زبان جاوا تقسیم می‌شوند. در هنگام کامپایل، تمامی کتابخانه‌هایی که توسعه‌دهنده هنگام توسعه‌ی برنامه، آن‌ها را استفاده کرده‌است به همراه کدهای مورد توسعه، کامپایل شده و در ساختار سلسله‌مراتبی تحت فایل‌های `classes.dex` قرار می‌گیرد [۲۱]. ذکر این نکته قابل توجه است که تشخیص برنامه‌های بازبسته‌بندی شده بدون شناسایی کتابخانه‌های برنامه اندرویدی مورد نظر امکان‌پذیر نیست. واضح است که در صورتی که نتوانیم کتابخانه‌های اندرویدی شاخص را از جفت برنامه‌های مورد بررسی جدا کنیم، آنگاه بخش زیادی از شباهت دو برنامه ناشی از کتابخانه‌های اندرویدی و اشتراکات موجود در آن‌ها است چرا که بسیاری از کتابخانه‌ها خصوصاً کتابخانه‌های زبان جاوا، در هر برنامه اندرویدی موجود است. از طرفی، به دلیل عدم وجود مرز مشخصی میان کدهای کتابخانه‌ای و کدهای مورد توسعه توسط توسعه‌دهندگان، شناسایی کتابخانه‌های اندرویدی تبدیل به یک چالش در زمینه‌ی تشخیص برنامه‌های بازبسته‌بندی در این حوزه شده‌است.

## ۴-۲ طبقه‌بندی

طبقه‌بندی اطلاعات ورودی یکی از روش‌های مرسوم در هوش مصنوعی و یادگیری ماشین است که توسط الگوریتم‌های طبقه‌بند انجام می‌شود. یک طبقه‌بند شامل مجموعه‌ای از الگوریتم<sup>۳۸</sup>‌ها است که برای طبقه‌بندی و یا مرتب‌سازی<sup>۳۹</sup> داده‌های ورودی مورد استفاده قرار می‌گیرد [۲۲]. یکی از ساده‌ترین مثال‌های موجود برای طبقه‌بندی، جداسازی هرزنامه<sup>۴۰</sup>‌ها در سرویس‌های ایمیل است. روش‌های طبقه‌بندی نیازمند مجموعه‌ای از ویژگی‌های اطلاعات مورد بررسی به عنوان ورودی مسئله می‌باشند تا پس از اجرای الگوریتم، اطلاعات مسئله را بر اساس آن‌ها طبقه‌بندی کنند.

---

Algorithm<sup>۳۸</sup>  
Sorting<sup>۳۹</sup>  
Spam<sup>۴۰</sup>

## ۵-۲ بازبسته‌بندی برنامه‌های اندرویدی

با پوشش عمیق در پژوهش‌های مرتبط با این حوزه در سالیان اخیر متوجه می‌شویم که تعاریف متنوعی برای بازبسته‌بندی در نظر گرفته شده است. برخی از پژوهش‌ها نظیر [۲۳، ۲۴] بازبسته‌بندی را در تغییر منابع و ظاهر برنامه‌ها در نظر می‌گیرند و در نهایت ویژگی‌های مبتنی بر ظاهر آن‌ها را با یکدیگر مقایسه می‌کنند در حالی که برخی از پژوهش‌های اخیر دیگر نظیر [۲۵، ۲۶] بازبسته‌بندی مبتنی بر تغییر ویژگی‌های کدپایه تعریف شده است. البته که نمی‌توان به هیچ کدام از تعاریف بالا خرده گرفت چرا که هر دو تعریف از نظر مهاجم و اهداف تعریف شده توسط اون قابل استناد است. علاوه بر این یکی دیگر از اختلافات موجود در تعریف بازبسته‌بندی، وجود مبهم‌نگاری در برنامه‌های بازبسته‌بندی شده است. برخی از پژوهش‌ها نظیر [۲۷] بازبسته‌بندی را منوط به تغییر در امضای برنامه می‌دانند اما بسیاری از پژوهش‌های به‌روزتر، نظیر [۲۸، ۲۹] بازبسته‌بندی را تنها به تغییر منابع و یا کدهای برنامه تقلبی نسبت به برنامه اصلی می‌دانند. همانطور که مشاهده شد، هنوز تعریف مشخصی از بازبسته‌بندی در پژوهش‌ها ارائه نشده است اما به طور کلی می‌توان گفت که برنامه A بازبسته‌بندی یک برنامه دیگر است اگر تغییرات آن نسبت به برنامه مبدا محدود و با حفظ کارکرد و منابع برنامه اصلی باشد. این تعریف در این پژوهش نیز به عنوان تعریف مبنای بازبسته‌بندی در نظر گرفته شده است.



## فصل ۳

### تعریف مسئله و مرور کارهای پیشین

پژوهش‌های اخیر در حوزه‌ی تشخیص برنامه‌های اندرویدی بازبسته‌بندی شده نشان می‌دهد که تشخیص این دسته از برنامه‌ها تحت تاثیر دو عامل مبهم‌نگاری و جداسازی صحیح کتابخانه‌های اندرویدی قرار دارد. برخی از پژوهش‌های اخیر انجام‌شده در این حوزه، تشخیص کتابخانه‌های بسته‌ی تقلبی را با فرض عدم مبهم‌نگاری کتابخانه‌ها انجام داده‌اند که مشخصاً این فرضی نادرست است چرا که بسیاری از مبهم‌نگارهای ابتدایی نیز این کار را در کتابخانه‌های اندرویدی انجام می‌دهند. در اکثر روش‌های پیشنهادی قسمتی از روش، مختص تشخیص و جداسازی کتابخانه‌های اندرویدی است. شناسایی کدهای کتابخانه‌ای از آن جهت اهمیت دارد که تشخیص درست آن‌ها می‌تواند مثبت غلط و منفی غلط را کاهش دهد. در بیشتر مواقع، خصوصاً در ابزارهای مبهم‌نگاری، متقلب هنگام بازبسته‌بندی اقدام به مبهم‌نگاری در کتابخانه‌های اندرویدی می‌کند و بدین صورت سعی در افزایش منفی غلط در ابزارهای تشخیص دارد. در صورتی که کدهای کتابخانه‌ای به درستی تشخیص و جداسازی نشوند، شباهت‌های موجود میان برنامه‌های مورد بررسی، خصوصاً در روش‌های مبتنی بر تحلیل ایستا، ناشی از کدهای کتابخانه‌ای خواهد بود. از سوی دیگر، تشخیص مبهم‌نگاری در کدهای مورد توسعه توسط متقلب، نیازمند ویژگی‌هایی از برنامه‌هاست که مقاومت بالایی در برابر مبهم‌نگاری داشته باشند. بدین معنا که متقلب برای تغییر این دسته از ویژگی‌ها ناچار به پرداخت هزینه‌ی زمانی و فنی باشد و در نهایت از تغییر این دست از ویژگی‌ها، پرهیز کند. در بسیاری از روش‌های ارائه‌شده در سال‌های اخیر، تشخیص برنامه‌های بازبسته‌بندی شده مبتنی بر ویژگی‌هایی صورت گرفته است که در عین مقاومت در مقابل مبهم‌نگاری، هزینه‌ی محاسباتی تشخیص برنامه‌های بازبسته‌بندی شده را افزایش می‌دهد به طوری که استفاده از این روش‌ها را عملاً در یک محیط صنعتی غیر ممکن ساخته‌است.

با توجه به اهمیت تشخیص مبهم‌نگاری و در نهایت تشخیص برنامه‌های بازبسته‌بندی شده و همچنین، در

نظر گرفتن سرعت تشخیص به عنوان یک عامل مهم، در این فصل به بررسی و مرور کارهایی می‌پردازیم که روش‌های گوناگونی را برای تشخیص برنامه‌های بازبسته‌بندی استفاده کرده‌اند و مزایا و معایب هر کدام را به صورت جدا بررسی خواهیم کرد. از آنجایی که هدف این پژوهش بهبود کارایی روش‌های تشخیص برنامه‌های بازبسته‌بندی شده است و تمرکز پژوهش بر روی تشخیص کدهای کتابخانه‌ای نبوده است، در ابتدا روند کلی تشخیص برنامه‌های بازبسته‌بندی شده را در پژوهش‌های مرتبط بیان کرده و به اختصار، روش‌های جداسازی کتابخانه‌های اندرویدی از کدهای مورد توسعه را توضیح می‌دهیم و از مرور کارهای پیشین انجام‌شده در این حوزه عبور خواهیم کرد.

در ادامه ابتدا به روند کلی تشخیص برنامه‌های بازبسته‌بندی شده می‌پردازیم و مسئله‌ی تشخیص برنامه‌های بازبسته‌بندی شده را از دیدگاه این پژوهش، شرح می‌دهیم. همچنین، دسته‌بندی انواع روش‌های تشخیص را با توجه به پژوهش‌های سال‌های اخیر بیان می‌کنیم و از هر دسته، چند پژوهش انجام‌شده را بررسی خواهیم کرد. برای درک بهتر روش پیشنهادی، در هر قسمت به بیان مزایا و معایب هر روش خواهیم پرداخت و علاوه بر این روش تشخیص کدهای کتابخانه‌ای در هر پژوهش را مشخص خواهیم کرد.

### ۳-۱ تعریف مسئله

علازم پژوهش‌های متعدد صورت‌گرفته در این زمینه، همانند تعریف بازبسته‌بندی، هنوز تعریف مشخصی نیز برای تشخیص بازبسته‌بندی ارائه نشده است. پژوهش‌های سال‌های اخیر در حالت کلی تشخیص بازبسته‌بندی را به دو صورت تعریف می‌کنند:

**تعریف ۳-۱ (تشخیص بازبسته‌بندی مبتنی بر برنامه‌ی مبدا)** / تشخیص بسته‌ی بازبسته‌بندی شده، یعنی تشخیص جفتی از برنامه‌های درون مخزن که دقیقاً جفت مشابه برنامه‌ی ورودی باشد. به بیان دیگر در این تعریف مشخص می‌شود که برنامه‌ی ورودی بازبسته‌بندی شده است یا خیر و در صورتی که بود، جفت برنامه‌ی آن درون مخزن نیز مشخص می‌شود.

**تعریف ۳-۲ (تشخیص بازبسته‌بندی مبتنی بر تصمیم‌گیری برنامه‌ی مقصد)** تشخیص بسته‌ی بازبسته‌بندی شده، یعنی مشخص کنیم برنامه‌ی ورودی بازبسته‌بندی شده است یا خیر. در این حالت تشخیص برنامه‌ی اصلی اهمیت ندارد و مسئله، تصمیم‌گیری<sup>۱</sup> درباره‌ی بازبسته‌بندی بودن یک برنامه‌ی ورودی است.

در سال‌های اخیر، اکثر پژوهش‌ها از یکی از تعاریف بالا برای تشخیص بازبسته‌بندی استفاده کرده‌اند. برای پاسخ به تعریف ۲، پژوهش‌هایی نظیر [۳۰، ۳۱، ۳۲] از روش‌های مبتنی بر مدل‌های یادگیری ماشین<sup>۲</sup>

<sup>۱</sup> Decision  
<sup>۲</sup> Machine Learning

برای تشخیص برنامه‌های بازسته‌بندی شده استفاده کرده‌اند. حال آن‌که پژوهش‌های مرتبط با تعریف ۱، نظیر [۳۳، ۹] بیشتر از روش‌های مقایسه‌ی دودویی و مبتنی بر شباهت‌سنجی استفاده کرده‌اند. تعریفی که در این پژوهش مورد استفاده قرار گرفته است، تعریف ۱ است. یعنی تشخیص بازسته‌بندی منوط به تشخیص جفت برنامه اصلی در مخزن برنامه‌های پژوهش می‌باشد. بنابراین در طی فرایند تشخیص به ۲ سوال اساسی پاسخ می‌دهیم:

- آیا برنامه ورودی بازسته‌بندی شده‌ی یک برنامه‌ی دیگر است؟
- در صورتی که برنامه مورد بررسی، بازسته‌بندی شده‌ی برنامه دیگری بود، آنگاه جفت بازسته‌بندی شده‌ی برنامه ورودی مشخص گردد.

## ۳-۲ روند کلی تشخیص برنامه‌های بازسته‌بندی شده

با بررسی پژوهش‌های صورت‌گرفته در حوزه‌ی تشخیص برنامه‌های بازسته‌بندی شده، درمی‌یابیم که به طور مشخص عمده‌ی این روش‌ها مراحل مشابهی را برای حل این مسئله، دنبال کرده‌اند. به طور کلی عمده‌ی روش‌های تشخیص، به عنوان ورودی، یک برنامه اندویدی شامل یک فایل با پسوند Apk را دریافت کرده و پس از گذر از سه مرحله، مسئله را حل می‌کنند. در ادامه به بررسی این سه مرحله می‌پردازیم:

### ۳-۲-۱ پیش‌پردازش برنامه‌های اندرویدی

یکی از مراحل مهم در تشخیص برنامه‌های بازسته‌بندی شده، مرحله‌ی پیش‌پردازش<sup>۳</sup> است که تاثیر به سزایی در سرعت و دقت روش تشخیص خواهد داشت. حذف کدهای کتابخانه‌ای، حذف کدهای مرده و یا بیهوده و اعمال فیلترهای ساختاری<sup>۴</sup> از موارد نمونه در قسمت پیش‌پردازش است. در این قسمت روش‌های کلی مورد استفاده توسط پژوهش‌های اخیر جهت حذف کدهای کتابخانه‌ای را توضیح می‌دهیم. با توجه به مرور کارهای پیشین انجام‌شده در این حوزه، به صورت کلی دو دیدگاه در مورد تشخیص و جداسازی کتابخانه‌های اندرویدی وجود دارد:

- **مبتنی بر لیست سفید:** در این روش، لیستی از نام بسته‌ای مشهور کتابخانه‌ای در برنامه‌های اندرویدی در دسترس است و با استفاده از نام بسته‌های موجود در برنامه، کدهای کتابخانه‌ای از

---

<sup>۳</sup> Pre process  
<sup>۴</sup> Structural

کدهای مورد توسعه جدا می‌شوند. راه حل‌های مبتنی بر این روش، عموماً در مقابل مبهم‌نگاری‌های ساده‌ای نظیر تغییر نام بسته نیز مقاوم نیستند و به راحتی می‌توان آن‌ها را دور زد. مزیت این روش آن است که سرعت بالایی دارد چرا که فقط نام بسته‌ها با یکدیگر مقایسه می‌شوند اما دقت خوبی را ارائه نمی‌دهند. غالب پژوهش‌های مبتنی بر استفاده از لیست سفید، فرض کرده‌اند که تنها کدهای مورد توسعه توسط متقلب مبهم‌نگاری شده‌است و ابهام در کدهای کتابخانه‌ای را نادیده گرفته‌اند.

• **مبتنی بر شباهت‌سنجی و کدهای تکراری:** در این روش، ابتدا مخزن بزرگی از کتابخانه‌های اندرویدی تهیه می‌شود و به روش‌های گوناگون کدهای کلاسی برنامه‌ها و کدهای کتابخانه‌ای موجود در مخزن، با یکدیگر مقایسه می‌شوند و بدین طریق کتابخانه‌های اندرویدی از کدهای مورد توسعه در برنامه، جدا می‌شود. روش‌های مبتنی بر شباهت‌سنجی، بسته به این‌که از چه روشی برای یافتن کدهای تکراری استفاده می‌کنند، دقت‌های متفاوتی دارند اما به صورت کلی می‌توان گفت که مقاومت آن‌ها در مقابل مبهم‌نگاری بسیار بیشتر از روش‌های مبتنی بر لیست سفید است چرا که در صورتی که ویژگی‌های منتخب مقابل مبهم‌نگاری مقاوم باشند، آن‌گاه می‌توان گفت که درصد بالایی از کتابخانه‌های اندرویدی را می‌توان از کد اصلی برنامه جدا کرد.

## ۳-۲-۲ استخراج ویژگی

پس از حذف کدهای کتابخانه‌ای در قسمت قبلی و انجام پیش‌پردازش‌های مورد نیاز، کدهای منبع برنامه هدف، به یک طرح کلی مدل می‌شود. به صورت کلی می‌توان روش‌های تشخیص برنامه‌های بازبسته‌بندی شده را در پژوهش‌های سالیان خیر، ناشی از تفاوت در دیدگاه در مرحله‌ی استخراج ویژگی<sup>۵</sup> دانست. همانطور که در شکل — مشاهده می‌شود، روش‌های تشخیص برنامه‌های بازبسته‌بندی به صورت کلی به دو بخش تحلیل ایستا و تحلیل پویا تقسیم می‌شود. از آنجایی که هدف ما در این پژوهش، تنها بررسی پژوهش‌هایی است که روش‌های تشخیص بازبسته‌بندی ارائه داده‌اند بنابراین روش‌هایی که توسعه‌دهندگان و شرکت‌های توسعه‌دهنده جهت جلوگیری از انجام بازبسته‌بندی پیاده‌سازی می‌کنند را توضیح نمی‌دهیم. به صورت کلی، می‌توان روش‌های تشخیص برنامه‌های بازبسته‌بندی شده را به دو بخش روش‌های تحلیل پویا و یا روش‌های تحلیل ایستا تقسیم کرد که در ادامه به بررسی هر کدام از این روش‌های می‌پردازیم.

• **روش‌های مبتنی بر تحلیل ایستا:** روش‌های مبتنی بر تحلیل ایستا، در مقابل مبهم‌نگاری‌های ایستا که در هنگام بازبسته‌بندی و انجام دی‌کامپایل انجام می‌شود مقاوم هستند. اما همانطور که می‌توان

<sup>۵</sup> Feature Extracting

حدس زد، این دسته از روش‌ها مقابل روش‌های مبهم‌نگاری همانند بازتاب مقاومتی ندارند و ممکن است دچار خطا شوند. همچنین روش‌های مبهم‌نگاری مبتنی بر رمزنگاری پویا نیز این روش‌ها را دچار خطا می‌کند. یکی از مزایای مهم روش‌های مبتنی بر تحلیل ایستا آن است که در صورت پیاده‌سازی درست و استفاده از ویژگی‌های مقاوم، می‌توانند طیف وسیعی از برنامه‌های بازبسته‌بندی شده را تشخیص دهند.

• **روش‌های مبتنی بر تحلیل پویا:** ارائه‌ی روش‌های مبتنی بر تحلیل پویا، به هدف جلوگیری از مبهم‌نگاری‌های در لحظه‌ی اجرا<sup>۶</sup> که در برنامه‌های اندرویدی صورت می‌گیرد، می‌باشد. به همین علت روش‌های موجود در این حوزه، عمدتاً برنامه‌ها را در هنگام اجرا بررسی و استخراج ویژگی عمدتاً در هنگام اجرا انجام می‌گیرد. به طول کلی، روش‌های مبتنی بر تحلیل پویا از مقاومت بیشتر در مقابل استفاده از راهکارهای مبهم‌نگاری برخوردار هستند. استفاده از شبیه‌سازهای جعبه‌شن<sup>۷</sup> به وفور در پژوهش‌های این حوزه، یافت می‌شود. یکی از چالش‌های اصلی در تشخیص برنامه‌های اندرویدی بازبسته‌بندی شده، چگونگی پیاده‌سازی شبیه‌سازها<sup>۸</sup> است. بسیار از شبیه‌سازها توانایی شبیه‌سازی تمامی خدمات موجود در برنامه را ندارند و برای تحلیل دقیق‌تر نیازمند استفاده از کاربران واقعی در شبیه‌سازی و استفاده از خدمات برنامه هستند. عامل دیگری که تشخیص با استفاده از تحلیل پویا را مشکل می‌کند، این است که بسیاری از بدافزارهای توسعه‌یافته، توانایی تشخیص محیط اجرای شبیه‌سازی شده را دارند و ممکن است تمامی قابلیت‌های خود و یا بخشی از آن را به جهت دور زدن سیستم‌های تشخیص پویا، پنهان کنند.

### ۳-۲-۳ تشخیص بازبسته‌بندی

در این مرحله با توجه به معیارها و ویژگی‌هایی که از قسمت قبل به دست آمده است و با استفاده از روش‌های گوناگون برنامه بازبسته‌بندی شده مشخص می‌شود. به صورت کلی، روش‌های پیاده‌سازی شده در این قسمت، مبتنی بر مقایسه‌ی دودویی و یا طبقه‌بندی و یادگیری ماشین هستند.

• **مقایسه‌ی دودویی:** روش‌های مبتنی بر مقایسه‌ی دودویی، مدل استخراج شده در قسمت قبلی را با استفاده از شباهت‌سنجی با برنامه‌های موجود در مخزن مقایسه می‌کند و در نهایت برنامه بازبسته‌بندی شده را مشخص می‌کند. اکثر روش‌های مبتنی بر مقایسه‌ی دودویی، جفت برنامه اصلی را نیز مشخص می‌کنند و از تعریف ۱-۳ استفاده می‌کنند بنابراین یکی از مزیت‌های این

Execution Time<sup>۶</sup>  
Sand Box<sup>۷</sup>  
Simulator<sup>۸</sup>

روش‌ها پوشش گسترده‌تر از تعریف تشخیص بازبسته‌بندی است ولی در کنار آن اکثر روش‌های موجود در این زمینه، محاسبات بالایی دارند که باعث می‌شود سرعت آن‌ها کاهش یابد.

• **مبتنی بر طبقه‌بندی و یادگیری ماشین:** یکی دیگر از روش‌های تشخیص بازبسته‌بندی با استفاده از ویژگی‌های مستخرج از مرحله‌ی قبل، استفاده از طبقه‌بند ها و مدل‌های یادگیری ماشین است. اکثر پژوهش‌های موجود در این زمینه از تعریف ۲-۳ برای تشخیص برنامه‌ک بازبسته‌بندی شده استفاده می‌کنند. بنابراین، تنها تصمیم‌گیری در مورد بازبسته‌بندی بودن یا نبودن برنامه‌ک ورودی را انجام می‌دهند. یکی از مزایای مهم این روش‌ها، سرعت بالای آن است چرا که تنها در زمان مرحله‌ی یادگیری، نیازمند محاسبات بالایی هستند و در صورتی که مدل این روش‌ها به درستی عمل کند، سرعت تشخیص به صورت قابل توجهی بالاتر از روش‌های مبتنی بر مقایسه‌ی دودویی است.

### ۳-۳ روش‌های تشخیص بازبسته‌بندی

همانطور که در شکل — مشاهده می‌شود، اکثر پژوهش‌های تشخیص بازبسته‌بندی از روش‌های مقایسه‌ای مبتنی بر تحلیل ایستا و پویا استفاده می‌کنند. در ادامه‌ی این قسمت ابتدا روش‌های ایستا و همچنین پژوهش‌های اخیر مرتبط با این حوزه را بررسی خواهیم کرد و در ادامه روش‌های مبتنی بر تحلیل پویا شرح داده می‌شود.

#### ۱-۳-۳ مبتنی بر تحلیل ایستا

در این قسمت، روش‌های مبتنی بر تحلیل ایستا و پژوهش‌های مرتبط با آن را بررسی خواهیم کرد. همانطور که گفتیم تحلیل ایستا، روشی محبوب در میان پژوهش‌های اخیر موجود در این حوزه است چرا که پیچیدگی‌های روش‌های پویا را ندارد و می‌توان به کمک آن‌ها طیف وسیعی از تشخیص مبهم‌نگاری‌ها را در برنامه‌ک‌های اندرویدی بازبسته‌بندی شده پشتیبانی کرد.

#### روش‌های مبتنی بر آپکد و دستورات

استفاده از آپکد<sup>۹</sup>‌های موجود در فایل‌های دالویک، یکی از روش‌های تشخیص برنامه‌ک‌های بازبسته‌بندی شده است. هدف از پژوهش آقای ژو [۳۴] و همکاران، توسعه‌ی ابزاری به نام درویدمس<sup>۱۰</sup> بوده است

Opcodes<sup>۹</sup>  
DroidMoss<sup>۱۰</sup>

که توسط آن مشخص شود چه تعدادی از برنامه‌های موجود در فروشگاه‌های اندرویدی غیررسمی، بازبسته‌بندی شده‌ی برنامه‌های موجود در فروشگاه‌های رسمی هستند. همانطور که گفته شد نظارت کافی‌ای بر روی فروشگاه‌های غیر رسمی وجود ندارد، بنابراین متقلبین از این فروشگاه‌ها به عنوان یک راه امن و دردسترس برای پخش کردن برنامه‌های بازبسته‌بندی شده استفاده می‌کنند. برای استخراج امضای برنامه در این پژوهش از کدهای دالویک موجود در `Classes.dex` و امضای دیجیتال برنامه‌نویس در فراداده<sup>۱۱</sup> استفاده شده است. پس از جداسازی کدهای کتابخانه‌ای به وسیله‌ی لیست سفید و استخراج آپکدها از فایل‌های دالویک، از یک پنجره‌ی لغزات<sup>۱۲</sup> روی آپکدها استفاده شده و در نهایت چکیده‌ی Hashی آپکدها به همراه امضای دیجیتال برنامه‌نویس، موجود در پوشه‌ی META-INF تشکیل امضای برنامه را می‌دهند. همانطور که می‌توان فهمید، فرض پژوهش این بوده است که کلید خصوصی توسعه‌دهنده لو نرفته است. در نهایت برای قسمت شباهت‌سنجی، از الگوریتم فاصله ویراشی<sup>۱۳</sup> استفاده شده است. در قسمت شباهت‌سنجی از ۲۹۰۶ برنامه موجود در فروشگاه‌های رسمی استفاده شده و نتایج پژوهش نشان می‌دهد که ۵ تا ۱۳ درصد از برنامه‌های موجود در فروشگاه‌های غیر رسمی، بازبسته‌بندی شده‌ی برنامه‌های فروشگاه‌های رسمی است. در پژوهش دیگری که توسط آقای ژو [۲۶] ارائه شده است، هدف پژوهش، افزایش سرعت پژوهش قبلی با استفاده از نمونه‌های n تایی از آپکدها بوده است. در این پژوهش امضای هر برنامه متشکل از قسمتی از فراداده‌ی آن شامل فایل‌های منیفست<sup>۱۴</sup> و اطلاعاتی در مورد تعداد فایل‌های برنامه، توصیفات آن و چکیده‌ی آپکدهای دستورات برنامه است. این پژوهش با استفاده از یک مرحله پیش‌پردازش شامل بررسی فایل فراداده‌ی برنامه‌های موجود، فضای جست‌وجوی دودویی برنامه‌های مورد مقایسه را کاهش می‌دهد. دنوز و همکاران [۳۵] روش دیگری را مبتنی بر شباهت‌سنجی روی آپکدها با استفاده از فاصله‌ی فشرده‌سازی نرمال شده ارائه کرده‌اند. در این پژوهش ابتدا برای هر متد با توجه به دنباله‌ی دستورات موجود امضای مشخصی تولید می‌شود و در مرحله‌ی بعد متدهایی که بکتاب هستند از هر دو برنامه، بر اساس معیار فاصله‌ی فشرده‌سازی نرمال شده با یکدیگر مقایسه و بدین ترتیب متدهای مشابه استخراج می‌شود. در پژوهش [۳۶، ۳۷] پس از استخراج هیستوگرام‌های مربوط به تکرار آپکدها در قسمت‌های مختلف برنامه، هیستوگرام‌ها با استفاده از معیار فاصله‌ی مینی‌کاوسکی که یک معیار فاصله‌ی مبتنی بر هیستوگرام‌ها است مقایسه می‌شوند و در نهایت برنامه‌های بازبسته‌بندی شده مشخص می‌شوند. جرومه و همکاران [۳۸] در پژوهش خود با استفاده از آپکدها و تکرار آن‌ها و روش‌های مبتنی بر یادگیری ماشین برنامه‌های بازبسته‌بندی شده را تشخیص می‌دهند. در پژوهشی که توسط [۳۹] و همکاران، ارائه شده است، از نمونه‌برداری مبتنی بر n-gram در ۴ اندازه‌ی متفاوت ۱ تا ۴ استفاده شده است.

<sup>۱۱</sup> MetaData

<sup>۱۲</sup> Sliding Window

<sup>۱۳</sup> Edit Distance

<sup>۱۴</sup> Manifest

برای شباهت‌سنجی از روش‌های طبقه‌بندی مبتنی بر درخت‌تصمیم، شبکه‌های عصبی و بردار ماشین استفاده شده‌است.

آقای لین و همکاران [۴۰] در این پژوهش، با استفاده از فراخوانی‌های سیستمی<sup>۱۵</sup> صدا زده‌شده توسط برنامه، رفتار آن را طبقه‌بندی می‌کنند. به عقیده‌ی این پژوهش، از آنجایی که اکثر بدافزارهای هم‌خانواده، در بازبسته‌بندی برنامه‌های اندرویدی، رفتار مشابه یکدیگر دارند، بنابراین استفاده از فراخوانی‌های سیستم و استخراج آن‌ها از سطح دالویک بایت‌کدها و سطح نخ<sup>۱۶</sup>، می‌تواند امضاء یکتایی از هر برنامه تولید کند. پس از استخراج بردار ویژگی<sup>۱۷</sup> از فراخوانی‌های موجود با استفاده از آپکدهایر برنامه، از یک طبقه‌بند نیز برای شباهت‌سنجی استفاده شده‌است. با توجه به روش پژوهش، شناسایی و طبقه‌بندی بدافزارهای بازبسته‌بندی شده‌ای که رفتار مشخصی ندارند و در مخزن بدافزارها موجود نیستند، یکی از ویژگی‌های مفید پژوهش ارائه‌شده است. فروکی [۴۱] و همکاران در پژوهش خود یک راه حل مبتنی بر استفاده از بلاک‌های ۶۴ بیتی بر روی فایل‌های دودویی برنامه‌های اندرویدی ارائه کرده‌اند. در روش ارائه‌شده پس از استخراج بلاک‌های ۶۴ بیتی از فایل‌ها و با استفاده از چکیده خلاصه تشابه و استخراج آنتروپی<sup>۱۸</sup> برای هر بلاک، بلاک‌هایی که کوچکتر و بزرگتر از یک حد کمینه و آستانه باشند حذف می‌شوند. سپس به هر بلاک با توجه آنتروپی آن، یک اولویت<sup>۱۹</sup> اختصاص پیدا کرده که نشان می‌دهد بلوک مورد نظر دارای آنتروپی با احتمال بیشتر است. در نهایت پس از حذف بلوک‌هایی که احتمال رخداد پایین‌تری دارند نرخ مثبت غلط پژوهش کاهش یافته و از یک روش مبتنی بر بلوم‌فیلتر<sup>۲۰</sup> برای مقایسه و شباهت‌سنجی استفاده می‌شود.

آقای کو و همکاران [۴۲] از یک راه‌حل مبتنی بر استفاده از k-gram برای تشخیص بسته‌های بازبسته‌بندی شده استفاده کرده‌اند. نویسندگان، از حذف عملوندهای<sup>۲۱</sup> موجود در کدهای دودویی، به جهت کاهش مثبت‌های غلط در تشخیص بسته‌های بازبسته‌بندی شده استفاده شده‌است. در پژوهشی [۴۳]، کیشو و همکاران از یک راه‌حل مبتنی بر ترکیبی از دستورات کلاسی و متدهای برنامه استفاده کرده‌اند. در این پژوهش، در دو مرحله، ابتدا کلاس‌های مشابه با یکدیگر مشخص می‌شود و سپس در داخل کلاس‌های مشابه، متدهایی که یکسان هستند یافت می‌شود. شباهت‌سنجی میان کلاس‌ها، با استفاده از سه ویژگی، شامل لیست تمامی متدهای کلاس شامل ورودی و خروجی، لیست متغیرهای کلاسی و لیست کلاس‌هایی که داخل این کلاس فراخوانی شده‌اند، انجام می‌شود. پس از استخراج کلاس‌های مشابه، برای یافتن متدهای مشابه میان دو کلاس، از یک امضای مشترک شامل توصیف متدها به همراه نوع ورودی و

<sup>۱۵</sup>System calls

<sup>۱۶</sup>Thread

<sup>۱۷</sup>Feature Vector

<sup>۱۸</sup>Entropy

<sup>۱۹</sup>Priority

<sup>۲۰</sup>Bloom filter

<sup>۲۱</sup>Operand



خروجی آن‌ها و همچنین نام آن‌ها استفاده می‌شود. شباهت‌سنجی با استفاده از فاصله‌ی فشرده‌سازی<sup>۲۲</sup> انجام شده‌است.

راهول و همکاران [۴۴]، روشی را پیشنهاد کرده‌اند که در آن استخراج ویژگی مبتنی بر درخت نحو انتزاع<sup>۲۳</sup> انجام می‌شود. پژوهش پس از دستیابی به کد میانی برنامه‌های اندرویدی و تبدیل آن به مجموعه‌ای از قوانین نحوی، که به صورت مجموعه‌ای از عبارات منظم<sup>۲۴</sup> هستند، درخت نحو انتزاع را در سطح تابع تشکیل می‌دهد و سه ویژگی تعداد ورودی هر تابع، نام توابع صدا زده شده به صورت مستقیم و مجازی<sup>۲۵</sup> و متغیرهای شرطی را استخراج می‌کند. سپس برای طبقه‌بندی از الگوریتم نزدیک‌ترین همسایه به جهت تشخیص بازبسته‌بندی استفاده شده‌است. نرخ منفی غلط بسیار پایین از ویژگی‌های مورد توجه این پژوهش است. همچنین برای ذخیره‌سازی درخت نحو انتزاع، از یک ساختار مبتنی بر درخت B+ و پایگاه داده‌ی MySQL استفاده شده‌است.

به صورت کلی می‌توان گفت که روش‌های مبتنی بر دستورات، خصوصاً روش‌هایی که به صورت مستقیم از آپکد برای تشخیص برنامه‌های بازبسته‌بنده شده استفاده می‌کنند، توانایی بالایی را در تشخیص برنامه‌های بازبسته‌بندی شده ارائه نمی‌دهند. این روش‌ها هم‌اکنون مقابل ساده‌ترین مبهم‌نگاری‌ها نظیر تغییر نام بسته‌ها و کلاس‌ها مقاوم نیستند و بخش زیادی از پژوهش‌های این حوزه، بازبسته‌بندی را بدون تغییر در کدهای برنامه اصلی تعریف کرده‌اند که به نظر با توجه به وجود مبهم‌نگارهای امروزه، این فرضی غلط و غیر قابل اتکا است.

## روش‌های مبتنی بر گراف

در این قسمت به بررسی روش‌های مبتنی بر استفاده از گراف در تشخیص برنامه‌های بازبسته‌بندی شده می‌پردازیم. به صورت کلی می‌توان پژوهش‌های صورت گرفته در این دسته را از دو جنبه بررسی کرد، دیدگاه اول پژوهش‌هایی هستند که در نهایت امضاء هر برنامه را با استفاده از یک مدل گرافی نشان می‌دهند. در این دسته از پژوهش‌ها، برای مقایسه‌ی امضاء، ناچاراً از الگوریتم‌های تشابه گراف نظیر الگوریتم‌های تشخیص گراف‌های هم‌ریخت استفاده می‌شود و به علت سربار محاسباتی بسیار بالای این پژوهش‌ها، روش‌های این دسته بسیار کند هستند. از دیدگاه دیگر، پژوهش‌های این دسته، صرفاً با بررسی ویژگی‌های مبتنی بر گراف‌های جریان و داده‌ای میان قسمت‌های مختلف، ویژگی‌های هر برنامه را استخراج کرده و در نهایت امضاء هر برنامه را تشکیل می‌دهند. همانطور که می‌توان حدس زد، دسته‌ی دوم از سرعت بالاتری در تشخیص برخوردار است اما چگونگی مدل‌سازی با استفاده از ویژگی‌های گرافی، بخش مهمی

<sup>۲۲</sup> Compression Distance

<sup>۲۳</sup> Abstract Syntax Tree

<sup>۲۴</sup> Regular Expressions

<sup>۲۵</sup> Virtual

در پژوهش‌های این دسته است که باید به دقت پیاده‌سازی شود.

در پژوهشی که توسط آقای کروسل [۴۵] و همکاران، ابزاری مبتنی بر گراف وابستگی توسعه داده شده است. در این ابزار ابتدا، برنامه‌های موجود در مخزن با استفاده از یک ابزار شباهت‌سنجی در سطح فراداده<sup>۲۶</sup>ی برنامه، به جهت افزایش سرعت، کاهش می‌یابد. پس از حذف کدهای کتابخانه‌ای به روش لیست سفید، امضاء هر برنامه با استفاده از گراف وابستگی استخراج شده تشکیل می‌گردد. گراف وابستگی توابع، وابستگی اجزای یک تابع از دو منظر کنترلی و داده‌ای را معرفی می‌کند. وابستگی کنترلی<sup>۲۷</sup> در این پژوهش، الزام اجرای یک دستور خاص پیش از دستور دیگری است و وابستگی داده‌ای<sup>۲۸</sup>، الزام مقداردهی متغیر پیش از اجرای دستور مرتبط با آن است. در قسمت شباهت‌سنجی پس از ساخت گراف وابستگی داده‌ای، با استفاده از الگوریتم تشخیص گراف‌های هم‌ریخت VF2<sup>۲۹</sup> شباهت‌سنجی انجام شده و بسته‌های بازبسته‌بندی شده مشخص می‌شوند. در پژوهش دیگری که توسط آقای سان [۸] انجام شده است، هدف پژوهش افزایش دقت تشخیص برنامه‌های بازبسته‌بندی شده با تاکید بر شبیه‌سازی رفتار برنامه بوده است. در این پژوهش، واسطه‌های برنامه‌نویسی برنامه‌های اندرویدی مشخص کننده رفتار اصلی برنامه در نظر گرفته شده است. برای ساخت گراف هر برنامه، از گراف جریان مبتنی بر فراخوانی واسطه‌های اندرویدی استفاده شده و در نهایت پس از استخراج گراف، هر گراف نمایانگر امضاء یک برنامه می‌باشد. در گراف حاصل هر گره گراف حاوی اطلاعات یک واسطه و یال‌های گراف شامل جریان کنترلی بین واسطه‌های اندرویدی است. برای شباهت‌سنجی، از الگوریتم VF2 جهت تشخیص گراف‌های هم‌ریخت استفاده شده است. در مرحله آخر برنامه‌هایی که امضاء مشابه و یکسانی در تشخیص هم‌ریختی دریافت کرده‌اند به عنوان برنامه‌های بازبسته‌بند شده در نظر گرفته می‌شوند.

پژوهش آقای هو و همکاران [۴۶] شامل دو مرحله ساخت گراف فراخوانی متدهای<sup>۳۰</sup> برنامه و ماژول تشخیص بازبسته‌بندی است. پس از دیس‌اسمیل کردن فایل‌های برنامه، گراف فراخوانی متدهای برنامه تشکیل شده و تشکیل جنگلی از گراف‌های متصل و جدا از هم می‌دهند. سپس با استفاده از فراخوانی واسطه‌های اندرویدی موجود در هر متد، گراف به دو بخش فراخوانی‌های حساس<sup>۳۱</sup> و غیرحساس<sup>۳۲</sup> تقسیم می‌شود و با توجه به میزان حساسیت واسطه‌های فراخوانی شده، امتیاز اولویت<sup>۳۳</sup> به هر گراف نگاشت می‌شود و در نهایت با استفاده از مقایسه‌ی گرافی مبتنی بر امتیاز اولویت، شباهت‌سنجی انجام می‌شود.

---

Meta Data <sup>۲۶</sup>
Control Dependency <sup>۲۷</sup>
Data Dependency <sup>۲۸</sup>
Graph Isomorphism <sup>۲۹</sup>
Method Invocation Graph <sup>۳۰</sup>
Sensitive <sup>۳۱</sup>
Non-Sensitive <sup>۳۲</sup>
Priority Score <sup>۳۳</sup>

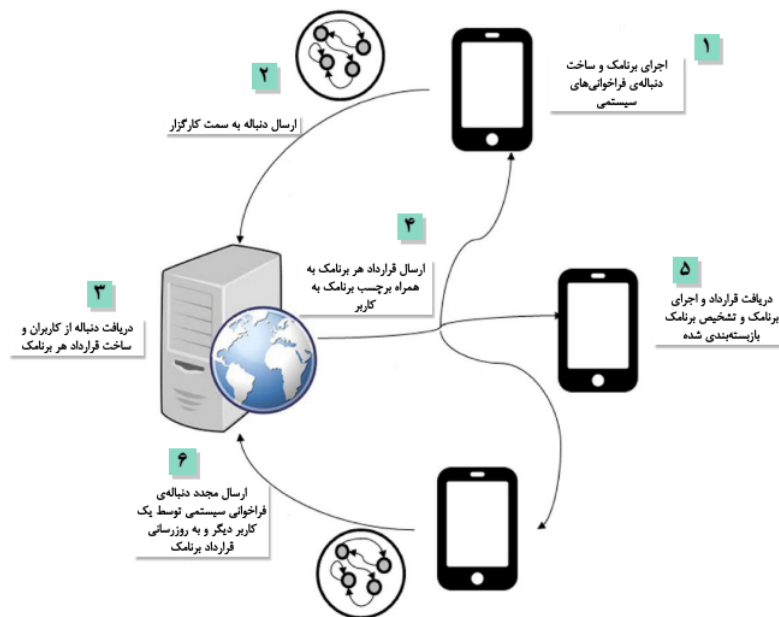
از آنجایی که پژوهش‌های مبتنی بر گراف در تشخیص برنامه‌های بازبسته‌بندی، عمدتاً به دلیل استفاده از روش‌های تشخیص گراف‌های هم‌ریخت‌کن هستند، ژو و همکاران [۴۷] روشی برای افزایش سرعت در تشخیص ارائه کرده‌اند. در این پژوهش در ابتدا ماژول‌های اصلی برنامه که رفتار اصلی آن را شکل می‌دهند شناسایی می‌شود. برای شناسایی ماژول‌های اصلی برنامه، از یک گراف جهت‌دار مبتنی بر ارتباط بسته<sup>۳۴</sup>‌های برنامه با یکدیگر استفاده شده و در نهایت یال‌های گراف بر اساس میزان ارتباطات بین بسته‌ها، مقداردهی می‌شود. با تشکیل گراف وزن‌دار ابتدایی، بسته‌هایی که ارتباط آن‌ها بر اساس وزن یال بین دو بسته از یک مقدار آستانه بیشتر باشد، با یکدیگر ادغام می‌شوند و رویه‌ی ادغام بسته‌ها در یک روند بازگشتی تا زمانی که هیچ بسته‌ای را نتوان با یکدیگر ادغام کرد تکرار می‌شود. در این حالت بسته‌ی نهایی شامل بسته‌ی اصلی برنامه است که منطق برنامه در این قسمت پیاده‌سازی شده است. رویه‌ی ساخت گراف ارتباطی بین بسته‌ها و ایده‌ی استفاده شده در قسمت ادغام بسته‌های اصلی با یکدیگر، ایده‌ای نو در این حوزه است که منجر به افزایش سرعت تشخیص نسبت به تمامی روش‌های گرافی شده است. برای مقایسه‌ی میان ماژول‌های اصلی برنامه، ابتدا اصلی‌ترین ماژول شامل بیشترین تعداد فعالیت، مشخص می‌شود و مقایسه میان ماژول‌های اصلی برنامه‌های اندرویدی، با استفاده از یک بردار ویژگی متشکل از فراخوانی واسط‌ها و مجوزهای<sup>۳۵</sup> درخواستی انجام می‌شود. برای مقایسه از درخت VP<sup>۳۶</sup> استفاده شده است که منجر به افزایش سرعت در کنار دقت مناسب شده است. برخلاف روش‌های معمول گرافی و در نهایت مقایسه‌ی دودویی، روش پیاده‌سازی شده در پژوهش ژو، با مرتبه‌ی زمانی  $N \log n$  یکی از پر سرعت‌ترین روش‌های مبتنی بر تشکیل گراف می‌باشد.

در پژوهش دیگری که به جهت افزایش سرعت در دسته پژوهش‌های مبتنی بر گراف ارائه شده است، چن و همکاران [۴۸] با استفاده از مدل کردن گراف به یک فضای سه‌بعدی، سرعت تشخیص و مقایسه‌ی گراف‌های هم‌ریخت را به مراتب افزایش داده‌اند. ایجاد کدهای مرده<sup>۳۷</sup> در کدهای برنامه، منجر به تغییر گراف جریان برنامه‌های اندرویدی می‌شود به همین جهت در این پژوهش تمامی گره‌های گرافی که نشان‌دهنده‌ی متدهای برنامه هستند به یک فضای سه‌بعدی نگاشت شده و مرکز جرم هر گراف با توجه به مختصات گره‌های گرافی تعیین می‌شود. در قسمت مقایسه‌ی گرافی، مرکز جرم گراف‌های متناظر با یکدیگر مقایسه شده و کاندیدهای بازبسته‌بندی مشخص می‌شود. در مرحله‌ی بعد برای مقایسه‌ی گره‌های هر گراف و تطبیق گراف‌های کاملاً متناظر، از مقایسه‌ی فاصله‌ی ویرایشی گره‌های متناظر استفاده می‌شود. روش ارائه‌شده علاوه بر مقاومت بالا مقابل مبهم‌نگاری، ناشی از مدل‌کردن برنامه در یک فضای گرافی، به دلیل استفاده از روشی نو در مقایسه‌ی گراف‌های مخزن برنامه‌ها، سرعت بسیار بالاتری از روش‌های

Package<sup>۳۴</sup>  
 Premissions<sup>۳۵</sup>  
 Vantage-Point Tree<sup>۳۶</sup>  
 Dead Code<sup>۳۷</sup>

پیشین دارد. برای حذف کتابخانه‌های اندرویدی از روش لیست سفید مبتنی بر اندازه‌ی بسته‌های مورد مقایسه استفاده شده‌است.

بر خلاف پژوهش‌های رایج در حوزه‌ی تشخیص بازبسته‌بندی برنامه‌های اندرویدی، در پژوهش آقای آلدینی و همکاران [۴۹]، از یک معماری کارخواه-کارگزار<sup>۳۸</sup> برای تشخیص برنامه‌های اندرویدی بازبسته‌بندی شده استفاده شده‌است. در این معماری یک برنامه‌ی بر روی دستگاه اندرویدی کاربران نصب می‌شود و شروع به ثبت و ارسال فراخوانی‌های سیستمی به کارگزار می‌کند.



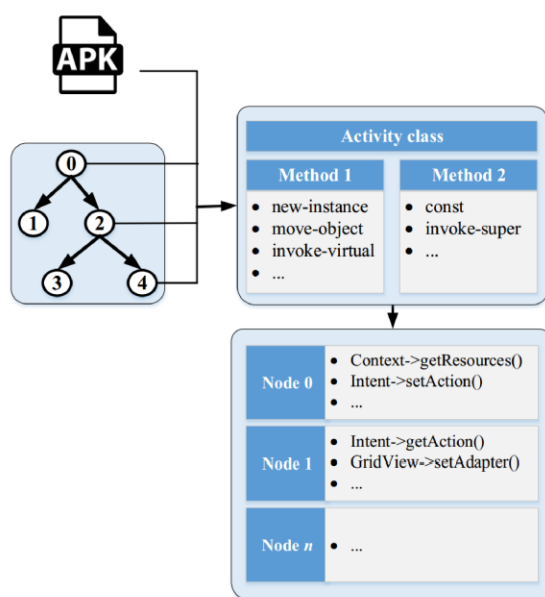
شکل ۳-۱: مراحل تشخیص برنامه‌های بازبسته‌بندی شده در پژوهش آقای آلدینی

ایده‌ی پژوهش، استفاده از فراخوانی‌های سیستمی برای شبیه‌سازی ایستای رفتار برنامه‌های اندرویدی بوده است. اثرانگشت برنامه توسط گراف فراخوانی‌های سیستمی ارسالی از سمت کاربران در سمت کارگزار، تشکیل می‌شود سپس با استفاده از اثرانگشت موجود، یک مدل برنامه تحت عنوان قرارداد ساخته شده و این مدل به برنامه‌ی کارخواه فرستاده می‌شود. در سمت کارخواه، برنامه نصب شده در مرحله‌ی اول، فراخوانی‌های سیستمی برنامه موجود را با قرارداد فرستاده شده مطابقت می‌دهد و در صورتی که فاکتورهایی نظیر نوع و تعداد فراخوانی‌های اندرویدی برنامه، با مدل پیش‌بینی شده یکسان نباشد، هشدار لازم از طریق برنامه نصب شده روی کارخواه به کاربر داده می‌شود. در کنار استفاده از گراف فراخوانی‌های سیستمی، پیاده‌سازی یک معماری کارخواه و کارگزار یکی از ایده‌های نو در این پژوهش بوده‌است. همچنین به دلیل استفاده از این معماری، پردازش سمت کارخواه به حداقل خود رسیده است و تحلیل برنامه نیازمند هیچ دانش اولیه‌ای از سمت کارخواه نمی‌باشد. علاوه بر این، تعداد برنامه‌های موجود در مخزن کارگزار،

<sup>۳۸</sup> Client-Server

به صورت مرتب افزایش پیدا کرده و این موجب پویایی مخزن برنامه‌های اندرویدی پژوهش می‌شود.

پژوهش دیگری در زمینه‌ی روش‌های تشخیص بازبسته‌بندی مبتنی بر گراف توسط جنگ و همکاران [۵۰] ارائه شده‌است. در این پژوهش، پس از استخراج گراف ارتباط بین کلاس‌ها از داخل دالویک بایت‌کدهای برنامه اندرویدی، امضای هر برنامه شامل یک بردار ویژگی، متشکل از فراخوانی‌های کلاسی برنامه می‌باشد. مقایسه‌ی بردارهای برنامه‌های موجود به وسیله‌ی الگوریتم یافتن بزرگ‌ترین زیردنباله‌ی مشترک<sup>۳۹</sup> انجام شده و تعیین یک حد آستانه در زیردنباله‌های مشترک، برنامه بازبسته‌بندی شده شناسایی می‌شود. روش مورد نظر را می‌توان به نوعی یک روش مبتنی بر گراف و دنباله‌ی آپکدهای برنامه توصیف کرد چرا که از هر دو ایده‌ی مدنظر استفاده نموده‌است. یکی دیگر از آخرین پژوهش‌های موجود در این دسته، در سال ۲۰۲۱ توسط نگویان [۳] مطرح شده‌است. پژوهش مورد نظر مبتنی بر استخراج گراف فعالیت<sup>۴۰</sup> در برنامه‌های اندرویدی است. گره‌های گراف مورد نظر شامل لیستی از واسطه‌های فراخوانی شده در آن فعالیت و یال‌های گراف، نشان‌دهنده‌ی یک انتقال از یک فعالیت به فعالیت دیگر است. یکی از معایب این پژوهش استفاده از الگوریتم VF۲ به عنوان الگوریتم اصلی برای مقایسه و یافتن گراف‌های هم‌ریخت است که باعث کاهش سرعت پژوهش شده‌است.



شکل ۳-۲: هر گره از گراف در پژوهش نگویان شامل لیستی از واسطه‌های فراخوانی شده در آن فعالیت است. [۳]

به عنوان آخرین پژوهش صورت گرفته در این قسمت، پژوهش [۵۱، ۵۲] را بررسی خواهیم کرد. پژوهش [۵۲] که الهام‌دهنده‌ی پژوهش [۵۱] می‌باشد با استفاده از پیمایش گرافی بر روی گراف جریان

<sup>۳۹</sup> Longest common Subsequence  
<sup>۴۰</sup> Activity Graph

برنامک‌های اندرویدی و استخراج ویژگی‌های متعدد نظیر واسط‌های اندرویدی، فراخوانی توابع و استفاده از فیلترهای ساختاری نظیر اندازه‌ی طول امضا، امضای هر کلاس را تشکیل می‌دهد و در نهایت با استفاده از الحاق تمامی کلاس‌های برنامک (مرتب شده به صورت الفبایی) امضاء مخصوص هر برنامک ساخته می‌شود. تفاوت این دو پژوهش بیشتر در ساختار امضاء هر برنامک می‌باشد که در فصل ۴ به تفصیل بیشتر این دو پژوهش و شرح تفاوت‌های آن خواهیم پرداخت.

به طور کلی می‌توان گفت که روش‌های گرافی تشخیص برنامک‌های بازبسته‌بندی شده از دقت بالایی در تشخیص برخوردار هستند. از طرفی اکثر روش‌های ارائه‌شده در این دسته، خصوصاً آن‌هایی که در نهایت هر برنامک را به یک طرح گرافی مدل می‌کنند، روش‌های تشخیص گراف‌های هم‌ریخت را در قسمت مقایسه به کار گرفته‌اند که این موضوع باعث می‌شود تا سربار محاسباتی سنگینی به پژوهش‌های مطرح وارد و سرعت تشخیص را کند سازد. علاوه بر این همانطور که بررسی شد روش‌های ارائه‌شده، در صورتی که مدل‌ها گرافی را در فضای دیگری بررسی کنند، سرعت تشخیص پژوهش بالاتر رفته و می‌توان از دقت در تشخیص مسائل گرافی نیز استفاده نمود.

### روش‌های مبتنی بر تحلیل ترافیک شبکه

در این قسمت به بررسی پژوهش‌های صورت‌گرفته در حوزه‌ی تشخیص برنامک‌های بازبسته‌بندی شده مبتنی بر تحلیل ترافیک شبکه<sup>۴۱</sup> می‌پردازیم. از آنجایی که ترافیک عبوری از برنامک‌هایی که مبتنی بر فضای مجازی عمل می‌کنند، می‌تواند رفتار خوبی از برنامک را مدل کند بنابراین استفاده از ویژگی‌های مبتنی بر این کارکرد، از محبوبیت بالایی میان توسعه‌دهندگان، برخوردار است. به عنوان اولین پژوهش مورد بررسی، پژوهش ارائه‌شده توسط وو[۵۳] و همکاران را بررسی خواهیم کرد. در این پژوهش امضای هر برنامک مبتنی بر ترافیک http تولیدشده توسط آن درست می‌شود. در مرحله‌ی اول این پژوهش، تمامی ترافیک تولیدشده توسط برنامک جمع‌آوری شده و در مرحله‌ی بعدی ترافیک http جداسازی و تحلیل روی این دسته ادامه پیدا می‌کند. ترافیک حاصل از برنامک‌های اندرویدی در این پژوهش به دو دسته‌ی کلی تقسیم می‌شود:

- ترافیک مرجع<sup>۴۲</sup>: ترافیک تولیدشده توسط برنامک توسعه‌یافته
- ترافیک کتابخانه‌ای: ترافیک تولیدشده توسط کدهای کتابخانه‌ای

جهت جداسازی ترافیک مرجع و ترافیک کدهای کتابخانه‌ای از الگوریتم‌های تطبیق جریان ترافیک<sup>۴۳</sup>

<sup>۴۱</sup> Network Traffic

<sup>۴۲</sup> Source

<sup>۴۳</sup> Traffic Stream Matching

http و الگوریتم تطبیق هانگرین<sup>۴۴</sup> استفاده شده است. در قسمت شباهت سنجی از درخت جست و جوی VPT به جهت افزایش سرعت به دلیل متوازن بودن درخت، استفاده شده است. پژوهش مورد نظر ایده ای نو در زمینه ی تشخیص برنامه های باز بسته بندی شده است اما مشکل اصلی این پژوهش آن است که در مقابل ترافیک رمز نشده مقاوم نیست و عملاً در برنامه هایی که تمامی ترافیک تولیدی توسط آن ها رمز گذاری شده است ناکارآمد می شود.

در پژوهشی که توسط شهری [۵۴] ارائه شده است تفکیک ترافیک به وسیله ی یک طبقه بند انجام می شود. پس از تفکیک ترافیک متغیرهای هر بسته شامل Request, Value, Get, Host مشخص می شود. ترافیک های از نوع Request در این قسمت به دو نوع اجباری و یا غیر اجباری تقسیم می شود. ترافیک غیر اجباری شامل ترافیک مصرفی برنامه برای ارتباط با واسطه های دسترسی عنوان شده است. جداسازی ترافیک های شبکه از آن جهت اهمیت دارد که بسیاری از کتابخانه های رایگان و یا حتی بدافزارهای موجود، یک نقطه ی دسترسی<sup>۴۵</sup> توسط واسطه های برنامه نویسی برای دسترسی به کتابخانه ها ایجاد کرده اند که برنامه های اندرویدی به وفور از این نفاط دسترسی استفاده می کنند. از طرفی تشخیص و جداسازی ترافیک اجباری و غیر اجباری موجب کاهش خطای منفی غلط می شود. برای شباهت سنجی ترافیک اجباری شامل ترافیک اصلی برنامه، از یک الگوریتم جریان درخواست<sup>۴۶</sup> مبتنی بر فاصله ی اقلیدوسی<sup>۴۷</sup> استفاده شده است. پیشنهاد جداسازی ترافیک کتابخانه ای و اصلی به وسیله ی نقطه های دسترسی واسطه های برنامه نویسی ایده ای نو در این پژوهش است که هر چند کامل نمی تواند ترافیک شبکه را جداسازی کند اما در صورت تکامل می تواند دقیق تر عمل کند.

در پژوهش دیگری که توسط آقای هه [۵۵] ارائه شده، از یک طبقه بند به جهت تشخیص برنامه های باز بسته بندی شده استفاده شده است. در پژوهش اخیر، ابتدا تمامی ترافیک کاربران متصل به یک شبکه به یک کارگزار سطح فرستاده می شود و این کارگزار ترافیک شبکه را برچسب زنی<sup>۴۸</sup> کرده و پس از استخراج ویژگی هایی نظیر محتوای<sup>۴۹</sup> اطلاعات هر بسته، آن ها را برای محاسبه ی امضا، سمت یک کارگزار مرکزی که به صورت ابری<sup>۵۰</sup> خدمات ارائه می کند، می فرستد. پس از ارسال ویژگی های مستخرج به سمت کارگزار ابری، فرایند شباهت سنجی آغاز می گردد. به جهت حفظ کامل حریم خصوصی کاربران، تحلیل ترافیک تنها بر روی ترافیک رمز گذاری نشده ی (http) انجام می شود. در سمت کارگزار، با استفاده از تحلیل ترافیک شبکه ی هر کاربر، جریان اطلاعات در ترافیک برچسب زده شده مشخص می شود و سپس با حذف قسمتی

<sup>۴۴</sup> Hungarian Matching

<sup>۴۵</sup> EndPoint

<sup>۴۶</sup> Request Flow

<sup>۴۷</sup> Euclidian Method

<sup>۴۸</sup> Labeling

<sup>۴۹</sup> Packet Content

<sup>۵۰</sup> Cloud



از ترافیک در هر جریان به جهت کاهش اندازه، مانند پاسخ درخواست‌ها<sup>۵۱</sup> و قسمتی از سربرگ<sup>۵۲</sup> ترافیک، طبقه‌بندی صورت می‌گیرد. برای طبقه‌بندی از الگوریتم پرسرعت پژوهش [۵۶] استفاده شده است. مالک و همکاران [۵۷] در پژوهش دیگری تحت عنوان CREDROID از یک روش بر درخواست نام دامنه<sup>۵۳</sup> برای تشخیص برنامه‌های بازبسته‌بندی شده استفاده کرده است. در این پژوهش ابتدا ترافیک کاربر برای تحلیل به یک کارگزار مورد اعتماد فرستاده شده و فرایند بررسی عمیق ترافیک مورد نظر آغاز می‌شود. در ادامه و در سمت کارگزار مورد اعتماد، ترافیک کاربر ارزیابی شده و درخواست‌های نام دامنه جداسازی و برای هر کاربر برچسب‌گذاری می‌شود. در قسمت شباهت‌سنجی ترافیک درخواستی، از یک روش مبتنی بر فاصله‌ی اقلیدوسی جهت بررسی متن درخواست‌های دامنه استفاده و در نهایت برنامه‌های بازبسته‌بندی شده شناسایی می‌شوند.

از آنجایی که بازبسته‌بندی برنامه‌های اندرویدی روشی محبوب برای حمله‌کنندگان به جهت تزریق و گسترش بدافزارهای اندرویدی است، تمرکز پژوهش [۵۸] و همکاران بر روی برنامه‌های اندرویدی بازبسته‌بندی شده و حاوی ترافیک مشکوک به بدافزار، با توجه رفتار مدل‌شده بر اساس درخواست‌های دامنه می‌باشد. این روش، مبتنی بر ترافیک حاصل از درخواست نام دامنه توسط برنامه‌های بازبسته‌بندی شده و سپس ارتباط آن با آدرس به دست آمده از درخواست بوده است. سپس ترافیک فرستاده‌شده از سمت کاربر برای آدرس‌هایی که به دست آمده‌اند بازبینی شده و جمع‌آوری می‌شود و در ادامه با استفاده از روشی مبتنی بر تطابق رشته‌های<sup>۵۴</sup> درخواست، شباهت‌سنجی صورت گرفته و برنامه‌های بازبسته‌بندی شده مشخص می‌شوند. یکی از ایرادات این پژوهش آن است که مبتنی بر رفتار بدافزارهای بیشتر شناخته‌شده انجام شده است و در صورتی که بدافزاری رفتار مشابه با بدافزارهای محبوب نداشته باشد شناسایی نمی‌شود. همچنین تحلیل ترافیک شبکه‌ی کاربر در ترافیک خام و رمزگذاری‌نشده صورت می‌گیرد و در صورتی که ارتباطات بدافزار بازبسته‌بندی شده حاوی ترافیک رمزگذاری‌شده باشد، آن‌گاه روش پیشنهادی در تشخیص آن‌ها ناتوان خواهد بود.

استفاده از روش‌های ترکیب با تحلیل ترافیک شبکه نظیر بررسی دسترسی‌های کاربران، در پژوهش شارما و همکاران [۵۹] استفاده شده است. تشخیص برنامه‌های بازبسته‌بندی شده در این پژوهش مبتنی بر طبقه‌بندی برنامه‌ها با استفاده از تحلیل ترافیک شبکه و بررسی مجوزهای دسترسی مورد نیاز برنامه می‌باشد. پس از دیس‌اسمبل برنامه، مجوزهای دسترسی از فایل فراداده‌ی<sup>۵۵</sup> برنامه مورد نظر استخراج شده و تشکیل یک بردار ویژگی دودویی را می‌دهند. این پژوهش از دو سطح برنامه‌های مورد نظر را بررسی می‌کند، در سطح اول اگر بردار ویژگی برنامه با برداری از برنامه‌های بازبسته‌بندی شده تطابق داشته باشد،

---

Response<sup>۵۱</sup>

Header<sup>۵۲</sup>

Domain Name Requests<sup>۵۳</sup>

String Matching<sup>۵۴</sup>

Manifest File<sup>۵۵</sup>



آنگاه برنامه‌مورد نظر به عنوان یک برنامه‌مشکوک به سطح بعد فرستاده می‌شود. در سطح دوم، ترافیک رمزگذاری‌نشده‌ی برنامه تفکیک شده و تحلیل روی ترافیک TCP ادامه می‌یابد، سپس با استفاده از یک طبقه‌بند درخت تصمیم<sup>۵۶</sup>، برنامه‌های بازبسته‌بندی شده با استفاده از تحلیل ترافیک کاربران، مشخص می‌شوند.

اگرچه پژوهش‌های صورت‌گرفته در این دسته معمولاً سرعت خوبی از نظر مقایسه‌ی مدل استخراج شده دارند، اما باید توجه داشت که محدودیت‌های موجود در این قسمت مانند به خطر افتادن حریم خصوصی کاربران، از طریق بررسی ترافیک رمزنگاری شده، یکی از ویژگی‌های منفی پژوهش‌های مبتنی بر ترافیک برنامه می‌باشد. از طرفی، بسیاری از برنامه‌های اندرویدی، به صورت برون‌خط فعالیت می‌کنند که موجب می‌شود روش‌های این دسته اساساً از بررسی آن‌ها ناکام بمانند.

### روش‌های مبتنی بر تحلیل منابع

عمده‌ی روش‌های موجود در این دسته، فرض استفاده از منابع برنامه‌های بازبسته‌بندی شده را در تعریف بازبسته‌بندی اعمال کرده‌اند. در این دسته از تعریف بازبسته‌بندی، متغلب با تغییر کدهای برنامه و استفاده‌ی مستقیم از منابع آن، سعی در جعل برنامه اصلی دارد. در این حالت رابط کاربری برنامه مشابه با رابط کاربری برنامه اصلی است اما کارکرد منطقی آن دچار تغییرات زیادی می‌شود بنابراین شناسایی این دسته از برنامه‌ها با استفاده از ویژگی‌های مبتنی بر منابع آن انجام می‌شود. به عنوان اولین پژوهش از این دسته، شائو و همکاران [۱]، برای شکل امضا هر برنامه، از دو دسته ویژگی آماری<sup>۵۷</sup> و ساختاری<sup>۵۸</sup> استفاده می‌کنند. در قسمت ویژگی‌های آماری، ۱۵ ویژگی شاخص از منابع برنامه‌های اندرویدی نظیر تعداد فعالیت‌ها<sup>۵۹</sup>، تعداد مجوزهای دسترسی، تعداد فیلترهای هدفمند<sup>۶۰</sup> و میانگین تعداد فایل‌های png و xml استخراج می‌شود. علاوه بر این ۵ ویژگی، ۱۰ ویژگی دیگر آماری شامل میانگین فراخوانی ب ۱۰ منبع پر بازدید در برنامه‌های اندرویدی، مطابق با جدول ۳-۱، استخراج می‌شود.

<sup>۵۶</sup> Decision Tree

<sup>۵۷</sup> Statistical

<sup>۵۸</sup> Structural

<sup>۵۹</sup> Activities

<sup>۶۰</sup> Intent Filters

جدول ۳-۱: ۱۰ منبع پربازدید در پژوهش [۱]

#	نوع منبع
1	id
2	drawable
3	string
4	color
5	style
6	dimen
7	layout
8	xml
9	integer
10	array

پس از استخراج ۱۵ ویژگی آماری، دو ویژگی ساختاری دیگر مبتنی بر لایه‌ی فعالیت و کنترل‌گر رویدادهای<sup>۶۱</sup> برنامه‌ساخته می‌شود. برای ساخت این دو ویژگی از یک درخت انتزاعی مبتنی بر فعالیت‌های برنامه‌ساخته شده است. در قسمت شباهت‌سنجی، پس از نرمال‌سازی ویژگی‌های استخراج شده در مراحل قبل، از دو روش طبقه‌بندی مبتنی بر نزدیک‌ترین همسایه و طبقه‌بندی طیفی<sup>۶۲</sup> برای تشخیص برنامه‌های تقلبی استفاده شده است.

در پژوهش دیگری که توسط لین و همکاران [۶۰] انجام شده است، تشخیص بازبسته‌بندی با استفاده از چکیده‌ی تصاویر موجود در فایل‌های apk موجود در پوشه‌ی منابع، صورت می‌گیرد. در این پژوهش ابتدا چکیده‌ی تمامی تصاویر استخراج شده و شباهت‌سنجی میان آن‌ها به صورت تجمعی صورت می‌گیرد. در نهایت برنامه‌هایی که امتیاز تشابه آن‌ها از یک حد آستانه بیشتر باشد به عنوان برنامه‌های بازبسته‌بندی شده تشخیص داده می‌شوند. یکی از مفروضات پژوهش، به جهت کاهش فضای مقایسه، عدم لو رفتن کلید خصوصی توسعه‌دهنده‌ی برنامه می‌باشد. علاوه بر این، استفاده از روش‌های چکیده‌سازی معمول در این پژوهش، منجر به مقاومت پایین آن در مقابل مبهم‌نگاری می‌شود.

سان و همکاران [۶۱] با استفاده از تحلیل گراف فعالیت‌های برنامه‌ساخته شده، مجموعه‌ای از تصاویر ضبط شده<sup>۶۳</sup> هر برنامه را تحت هر فعالیت استخراج کرده و هر دید<sup>۶۴</sup> موجود در تصاویر را با استفاده از یک مستطیل

<sup>۶۱</sup> Event Handler

<sup>۶۲</sup> Spectral Classification

<sup>۶۳</sup> ScreenShot

<sup>۶۴</sup> View

نشان‌دهی و در نهایت به ازای هر تصویر، یک گروه از ناحیه دیدها ساخته می‌شود. در نهایت با استفاده از ادغام نواحی و گراف فعالیت‌های برنامه، گرافی تحت عنوان گراف گروه-ناحیه ساخته می‌شود. برای مقایسه و شباهت‌سنجی، با استفاده از پیمایش گراف‌ها، گره‌های گراف به عنوان نواحی برنامه با یکدیگر مقایسه می‌شود. دو ناحیه در صورتی با یکدیگر یکسان هستند که تعدادی از مجموعه‌های مدل‌شده به شکل مستطیل در این نواحی با یکدیگر هم‌پوشانی داشته باشند.

هو و همکاران [۶۲] با بررسی برنامه‌های بازبسته‌بندی شده متوجه شدند که ویژگی‌های ساختاری واسط‌های کاربری<sup>۶۵</sup> همانند طول و اندازه‌ی اشکال و دکمه‌ها کم‌تر دچار تغییر می‌شود و بیشتر ویژگی‌های محتوایی مانند رنگ پس‌زمینه تغییر می‌کند. با توجه به این یافته، آن‌ها سه ویژگی مهم مبتنی بر ساختار واسط‌های کاربری را استخراج و تحت عنوان امضای برنامه استفاده می‌کنند و در ادامه، این سه ویژگی را به یک فعالیت نسبت می‌دهند. در نهایت با کنار هم قرار گرفتن سه تایی‌های هر فعالیت و مقایسه‌ی آن‌ها با یکدیگر برنامه‌های تقلبی مشخص می‌شوند.

هدف آقای لین در پژوهش [۶۳] کاهش نرخ منفی غلط و مثبت غلط بوده‌است. در این پژوهش از ساختار سلسله‌مراتبی<sup>۶۶</sup> فایل‌های موجود در یک apk استفاده شده‌است. ساختار سلسله‌مراتبی مبتنی بر فایل‌های برنامه استخراج می‌شود و گره‌های این درخت برچسب‌گذاری می‌شود. برچسب‌گذاری درخت در این پژوهش، مبتنی بر نوع فایل مورد نظر آن گره انجام می‌شود چرا که برچسب‌گذاری بر اساس نام فایل‌ها، مقاومت پایینی در مقابل مبهم‌نگاری دارد و متقلب می‌تواند به راحتی روش را دور بزند. علاوه بر این مشکل دیگری که در این پژوهش وجود دارد، این است که در صورتی که چندین فایل از یک نوع وجود داشته باشد، گره‌های مشابه در درخت سلسله‌مراتبی به وجود می‌آید و در نهایت نرخ مثبت غلط به شدت افزایش پیدا می‌کند. برای حل این مشکل، نویسنده از یک ساختار چند مرحله‌ای برای تشخیص برنامه‌های بازبسته‌بندی شده استفاده می‌کند. در این ساختار ابتدا برچسب‌گذاری به نحوی انجام می‌شود که نرخ مثبت غلط زیادی در نتایج وجود داشته باشد و در ادامه از روشی با نرخ منفی غلط بالا استفاده می‌کند تا دقت پژوهش را افزایش دهد. شباهت‌سنجی درخت‌ها با استفاده از فاصله ویرایشی میان دو درخت انجام می‌شود. روش مورد نظر مقاومت پایینی در مقابل مبهم‌نگاری‌های ساده به خصوص افزونگی فایل‌ها، خواهد داشت چرا که متقلب می‌تواند به راحتی با ایجاد تعداد زیادی فایل بلااستفاده، سیستم تشخیص را دور بزند.

گوآ و همکاران [۶۴] با استفاده از طرح‌بندی<sup>۶۷</sup> و ایجاد درخت مبتنی بر آن برنامه‌های بازبسته‌بندی شده را تشخیص می‌دهند. ابتدا درختی از تمامی مولفه‌های از نوع منبع و با استفاده از فایل‌های xml به نام درخت طرح‌بندی کامل<sup>۶۸</sup>، استخراج می‌شود. درخت مذکور حاوی تمامی مولفه‌های طرح‌بندی یک برنامه

User Interface<sup>۶۵</sup>  
File Structure<sup>۶۶</sup>  
Layout<sup>۶۷</sup>  
Total tree layout<sup>۶۸</sup>

اندرویدی است که تنها یک ریشه دارد. در ادامه برای مقایسه و شباهت‌سنجی درخت طرح‌بندی کامل، از یک روش مبتنی بر چکیده‌سازی CTPH استفاده شده‌است.

استفاده از طرح‌بندی در پژوهش‌های مبتنی بر منابع محبوبیت زیادی دارد، برای مثال، در روش [۶۵]، ابتدا طرح‌بندی‌های مربوط به کتابخانه‌های اندرویدی حذف شده، سپس هر طرح‌بندی به یک درخت نگاشت می‌شود. به جهت مقایسه، طرح‌بندی‌های برنامه‌ها با یکدیگر ادغام می‌شوند و درخت نهایی در قالب یک فایل XML کدگذاری شده و در ادامه با استفاده از چکیده‌سازی فازی میان فایل‌های XML برنامه‌های بازبسته‌بندی شده مشخص می‌شوند. در نمونه‌ی دیگری، لیو و همکاران [۶۶] در یک ساختار دو مرحله‌ای مبتنی بر طرح‌بندی برنامه‌های بازبسته‌بندی شده را تشخیص می‌دهند. مرحله‌ی اول در این پژوهش شامل تشخیص برنامه‌های مشکوک به صورت درشت‌دانه<sup>۶۹</sup> است که با استفاده از تصاویر موجود در پوشه‌ی منابع انجام می‌گیرد. در ادامه و به صورت ریزدانه، درخت طرح‌بندی تشکیل شده و تشکیل امضای هر برنامه‌ها را می‌دهد. برای مقایسه درخت‌ها، از روش‌های مرسوم مانند فاصله ویرایشی، استفاده شده‌است. به عنوان یکی از آخرین پژوهش‌های موجود در این دسته به جهت بهبود روش‌های [۶۷، ۶۸] ما و همکاران [۶۸] با معرفی گراف انتزاعی طرح‌بندی<sup>۷۰</sup> و گراف طرح‌بندی انتقالی<sup>۷۱</sup>، برنامه‌های بازبسته‌بندی شده را شناسایی می‌کنند. در این پژوهش در ابتدا طرح‌بندی‌های هر برنامه تحت عنوان یک گراف جهت‌دار مدل می‌شود که هر گره گراف مذکور شامل طرح‌بندی‌های مشابه می‌باشد. در مقایسه با [۶۷]، روش ما و همکاران، از سرعت بیشتری در کدگذاری طرح‌بندی و تبدیل آن‌ها به گراف انتزاعی برخوردار است. علاوه بر این در قسمت شباهت‌سنجی و مقایسه، از روشی بهبودیافته، مبتنی بر یافتن تطبیف گراف بیشینه<sup>۷۲</sup> استفاده شده که ارزیابی عملکرد پژوهش نسبت به دو پژوهش مورد مقایسه، افزایش سرعت و دقت را به همراه داشته است.

### ۳-۳-۲ مبتنی بر تحلیل پویا

به صورت کلی پژوهش‌های صورت‌گرفته در این قسمت را می‌توان به دو بخش روش‌های مبتنی بر جعبه‌شن<sup>۷۳</sup> و یا روش‌های خودکار تقسیم‌بندی کرد. در روش‌های مبتنی بر جعبه‌شن، پژوهش‌کنندگان با ایجاد محیطی شبیه‌سازی شده و تعامل حداکثری با برنامه‌اندویدی مدنظر، سعی در شبیه‌سازی رفتار برنامه‌ها و جمع‌آوری مجموعه‌ای از ویژگی‌های موردنیاز به جهت مقایسه با یکدیگر دارند. این دسته از روش‌های پویا دو ایراد اساسی دارند، نخست آن‌که ایجاد یک محیط شبیه‌سازی شده و اجرای هر برنامه‌ها به طوری که تمامی

Coars Grain<sup>۶۹</sup>  
Abstract Layout Graph<sup>۷۰</sup>  
Abstract Transition Graph<sup>۷۱</sup>  
Maximum Graph Matching<sup>۷۲</sup>  
SandBoxing<sup>۷۳</sup>

رفتارهای آن را شبیه‌سازی کند، نیازمند زمان زیادی است که ممکن است روش را ناکارآمد کرده و بررسی برنامه‌های موجود را محدود نماید. علاوه بر این ایجاد محیطی که بتواند به صورتی کامل رفتار برنامه را شبیه‌سازی کند نیازمند آن است که به نوعی تمامی خدمات هر برنامه منحصرأ بررسی شود که این عملی تقریباً ناممکن است مگر آن‌که شبیه‌سازی منحصر یک برنامه خاص پیاده‌سازی شود. ایراد دوم آن است که برنامه‌های بازبسته‌بندی شده که دارای بدافزار هستند، ممکن است پیاده‌سازی آن‌ها قادر به شناسایی محیط شبیه‌سازی شده باشد و در نتیجه، از اجرای قسمت‌هایی از برنامه خودداری کرده و سیستم تشخیص را به خطا بیندازد. برخلاف روش‌های مبتنی بر جعبه‌شن، روش‌های مبتنی بر تحلیل خودکار بیشتر سعی در استفاده از داده‌های کاربران واقعی برنامه‌های اندرویدی دارند. بیشتر روش‌ها در این دسته از معماری‌های ابری و کارخواه-کارگزار استفاده کرده و با استفاده از ارسال گزارشات<sup>۷۴</sup> از سمت کاربران شباهت‌سنجی را انجام می‌دهند. چندین نمونه از این روش‌ها را مانند [۴۹، ۴۰] در قسمت‌های پیشین بررسی کردیم. روش‌های مبتنی بر تحلیل خودکار هیچ شبیه‌سازی در مورد برنامه انجام نمی‌دهند و رفتار برنامه را با توجه به رفتار کاربران مورد تحلیل قرار می‌دهند. تمرکز ما در این قسمت روی پژوهش‌هایی است که بیشتر تمرکز آن‌ها بر پیاده‌سازی روشی مبتنی بر جعبه‌شن و شبیه‌سازی رفتار برنامه‌ها بوده است.

در پژوهش والریو و همکاران [۶۹] با استفاده از اجرای برنامه‌های اندرویدی در چندین جعبه‌شن مشهور و جمع‌آوری اطلاعات برای هر برنامه اندرویدی، یک پروفایل-استفاده<sup>۷۵</sup> ساخته می‌شود. این پروفایل نشان‌دهنده‌ی استفاده‌ی برنامه از ۸ خدمت مبتنی بر منابع سیستمی شامل لیست تماس‌ها، موقعیت مکانی، پیامک<sup>۷۶</sup>، شبکه‌ی ارتباطی بی‌سیم<sup>۷۷</sup>، برنامه‌ها، باتری و تماس‌ها می‌باشد. روند کلی جمع‌آوری امضای برنامه به این شکل است که برنامه‌های موجود در مخزن توسط یک مولفه در چندین جعبه‌شن مشهور و در دسترس اجرا شده و ویژگی‌های ذکر شده استخراج می‌شود و در نهایت تشکیل یک بردار ویژگی می‌دهند. در مرحله‌ی انتهایی، بردارهای ویژگی برنامه‌های اندرویدی با یکدیگر مقایسه شده و بسته‌های بازبسته‌بندی شده مشخص خواهند شد.

از آن‌جا که بررسی پردازش برنامه‌های اندرویدی به ازای ورودی‌های یکسان می‌تواند نتایج مشابه در برنامه‌های اندرویدی داشته باشد، ژانگ و همکاران [۷۰] روشی را مبتنی بر زمان مورد نیاز برای اجرای ورودی‌های یکسان توسط پردازنده<sup>۷۸</sup>، پیاده‌سازی کرده‌اند. در این پژوهش، ابتدا تعدادی ورودی به هر برنامه اندرویدی موجود در مخزن برای اجرا داده می‌شود و پس از اتمام پردازش، بازه‌ی زمانی اجرای پردازش در پردازنده ضبط می‌شود و تحت عنوان یک دوتایی <پردازش، زمان> نمایش داده می‌شود. ورودی‌های یکسان برای تمامی برنامه‌های مورد بررسی اجرا می‌شود و در نهایت لیستی از دوتایی‌های موجود برای هر

Logs<sup>۷۴</sup>  
Usage Profile<sup>۷۵</sup>  
SMS<sup>۷۶</sup>  
Wifi<sup>۷۷</sup>  
CPU<sup>۷۸</sup>

برنامک، تشکل امضای آن را می دهد.

در روش دیگری که توسط نگویان و همکاران [۷۱] اجرا شده، برنامک های اندرویدی با یک روش مبتنی بر جعبه شن و استخراج ویژگی هایی از واسط کاربری برنامک، بسته های بازبسته بندی شده را شناسایی می کنند. در این پژوهش با استفاده از یک جعبه شن که وظیفه ی آن ضبط تصاویر محیط کاربر برنامک می باشد، تصاویری از برنامک های اندرویدی مورد بررسی ذخیره می شود. از آنجایی که روش های چکیده سازی مرسوم، چکیده ای کاملاً متفاوت حتی برای کوچکترین تغییرات تولید می کنند، در این پژوهش از روشی مبتنی بر چکیده سازی ادراکی<sup>۷۹</sup> استفاده شده است. در این دسته از روش های چکیده سازی، تغییرات کوچک به همان اندازه باعث تغییرات کوچک در چکیده ی تولید شده می شود و فایل های مشابه اما با تغییرات اندک می توانند پس از مقایسه ی چکیده سازی ادراکی، شناسایی شوند. در مرحله ی شباهت سنجی، تصاویر ضبط شده از هر برنامک، با استفاده از روش های میانگین چکیده سازی<sup>۸۰</sup>، با تصاویر برنامه های موجود در مخزن مورد مقایسه قرار می گیرند و در نهایت برنامک های بازبسته بندی شده به این طریق تشخیص داده می شوند.

در پژوهش دیگری که مبتنی بر واسط های کاربری انجام شده است، سو و همکاران [۷۲] با استفاده از استخراج فعالیت های برنامک موجود در فایل فراده ی آن ها، فعالیت های مورد نظر را در یک محیط مبتنی بر جعبه شن اجرا و تصاویر حاصل از اجرای هر فعالیت را ضبط می کنند. در قسمت شباهت سنجی، اطلاعاتی از درون هر کدام از تصاویر ضبط شده از برنامک حین اجرای آن، استخراج می شود و تشکیل بردار ویژگی هر برنامک را می دهند. در ادامه با استفاده از روش های مبتنی بر چکیده سازی محلی<sup>۸۱</sup>، شباهت بردارهای ویژگی با یکدیگر سنجیده و در صورتی که میزان شباهت از یک حد آستانه بیشتر باشد، جفت مورد بررسی، بازبسته بندی شده یکدیگر شناسایی می شوند. برنامک برای مقاومت بیشتر در مقابل مبهم نگاری از لیست فعالیت های برنامک استفاده کرده و نقاط ورود<sup>۸۲</sup> را از روی این لیست فعالیت مشخص می کند. این پژوهش، فعالیت های موجود در برنامک های اندرویدی را مستقل از یکدیگر در نظر گرفته است و به همین دلیل نقاط ورود متنوعی مبتنی بر هر فعالیت تعیین می شود، در صورتی که در اکثر برنامک های اندرویدی، فعالیت های موجود کاملاً به هم وابسته هستند و گاهی یک فعالیت نتیجه ی یک فعالیت دیگر است بنابراین فرض پژوهش در این قسمت به نظر اشتباه می باشد.

به طور کلی، تشخیص برنامک های اندرویدی مبتنی بر روش های پویا شامل استخراج ویژگی های متنوع ساختاری و رفتاری برنامک در حین اجرای آن در محیط شبیه سازی شده است. برای مثال، در پژوهش [۷۳] با استفاده از یک برنامک تزریق شده به هسته<sup>۸۳</sup>ی سیستم عامل، تمامی فراخوانی های سیستمی ناشی

Perception Hashing-Phash<sup>۷۹</sup>  
Average Hashing<sup>۸۰</sup>  
Locality Sensitive Hash<sup>۸۱</sup>  
Entry Point<sup>۸۲</sup>  
Kernel<sup>۸۳</sup>

از اجرای برنامه در محیط جعبه‌شن، استخراج می‌شود. استخراج فراخوانی سیستمی از طریق هسته‌ی سیستم عامل، منجر به مقاومت بالای این روش مقابل مبهم‌نگاری می‌شود. کیم و همکاران [۷۴]، برای تشخیص برنامه‌های بازبسته‌بندی شده، از بردارهای ویژگی مبتنی بر فراخوانی‌های واسطه‌های برنامه‌نویسی<sup>۸۴</sup> استفاده کرده‌اند و در نهایت با مقایسه‌ی برنامه‌ها، مسئله را حل می‌کنند. گوان و همکاران [۷۵]، با استفاده از یک روش مبتنی بر معناشناسی<sup>۸۵</sup> در برنامه‌های اندرویدی، بسته‌های بازبسته‌بندی شده شناسایی می‌کنند. آن‌ها با بررسی مشکلات تشخیص در تحلیل‌های ایستا، در ابتدا گراف جریان ارتباطی مبان کلاس‌های برنامه را استخراج کرده و در ادامه با استفاده از وزن‌دهی گراف، مبتنی بر ورودی و خروجی کلاس، کلاس‌های اصلی برنامه را از کلاس‌های فرعی جدا می‌کنند. سپس متدهای هر کلاس با استفاده از گراف جریان میان متدهای کلاسی، به صورت نمادین اجرا می‌شوند و با بررسی تمامی جایگشت‌های ممکن به عنوان ورودی هر متد و خروجی آن‌ها، متدهای یکسان و شبیه به دست می‌آیند. تشخیص بازبسته‌بندی با استفاده از یک حد آستانه، بر اساس تعداد متدهای یکسان مشخص می‌شود. یو و همکاران [۶۷] با استفاده از ساخت گراف فعالیت مبتنی بر واسطه کاربری هر برنامه و در نهایت مقایسه‌ی گرافی، برنامه‌های بازبسته‌بندی شده را تشخیص داده‌اند. برای مقایسه‌ی گرافی، پس از وزن‌دهی گراف بر اساس تعداد تکرار هر فعالیت، از یک الگوریتم تطبیق بیشینه‌ی گراف، استفاده شده است.

### ۳-۳-۳ سایر روش‌ها

وانگ و همکاران [۷۶] از روشی مبتنی بر طبقه‌بندی برای شناسایی کتابخانه‌های اندرویدی استفاده کرده‌اند. با توجه به این که کتابخانه‌های اندرویدی در برنامه‌های مختلف تکرار می‌شوند بنابراین جمع‌آوری دنباله‌ی فراخوانی واسطه‌های برنامه‌نویسی در کتابخانه‌های مشابه، منجر به ساخته‌شدن دنباله‌های مشابه می‌شود. بنابراین پس از جمع‌آوری دنباله‌ی فراخوانی‌ها از برنامه‌های متعدد، طبقه‌بندی روی دنباله‌ی ورودی با استفاده از داده‌ی آموزش انجام می‌شود و در نهایت کدهای کتابخانه‌ای با استفاده از این روش جدا می‌شوند. در قسمت دوم از یک راه حل دو مرحله‌ای برای تشخیص برنامه‌های اندرویدی بازبسته‌بندی شده استفاده شده است. پس از حذف کدهای کتابخانه‌ای، دنباله‌های فراخوانی واسطه‌های اندرویدی در هر برنامه محاسبه و دنباله‌های موجود با یکدیگر مقایسه می‌شوند. در این مرحله، برنامه‌های مشکوک به بازبسته‌بندی تشخیص داده می‌شوند و در مرحله‌ی بعدی بررسی دقیق‌تر صورت می‌گیرد. در مرحله‌ی دوم، متغیرهای موجود در برنامه در روشی مبتنی بر توکن<sup>۸۶</sup>، به صورت برداری نشان داده می‌شوند. به جهت شباهت‌سنجی و مقایسه‌ی بردارهای موجود، تعداد بردارهای یکسان، که مبتنی بر شناسه‌ی هر متغیر و نام آن است، با

API Calls<sup>۸۴</sup>  
Semantic<sup>۸۵</sup>  
Token<sup>۸۶</sup>

یکدیگر مقایسه می‌شود و در صورتی که تعداد توکن‌های مشابه، در یک جفت برنامه‌ک از یک حد آستانه بیشتر باشد، جفت مورد بررسی به عنوان بازبسته‌بندی در نظر گرفته می‌شوند. روش پژوهش، از آن‌جا که در دومرحله برنامه‌ک‌های بازبسته‌بندی شده را شناسایی می‌کند از سرعت خوبی برخوردار است، اما با استفاده از روش‌های مبتنی بر افزودن متغیرهای بیهوده، می‌توان روش یادشده را دور زد.

نیشا و همکاران [۷۷]، با استفاده از ویژگی‌های مبتنی بر مجوزهای دسترسی و ایجاد بردار دودویی مجوزهای هر برنامه‌ک، به وسیله‌ی روش‌های یادگیری ماشین برنامه‌ک‌های بازبسته‌بندی شده را تشخیص می‌دهند. یکی از مشکلات روش موجود آن است که برای ایجاد مجموعه‌ی داده آزمون، از یک روش مبتنی بر کپی مجدد برنامه‌ک‌های اصلی استفاده شده‌است که باعث می‌شود روش موجود در عین داشتن دقت بالا در مورد مجموعه‌ی آزمون پژوهش، به صورت کلی مقاومت بالایی در تشخیص بازبسته‌بندی نداشته باشد. مشکل دیگر این روش، آن است که عملاً تعریف تشخیص بازبسته‌بندی در این روش‌ها، حل مسئله‌ی تصمیم در مورد بازبسته‌بندی است و شامل تشخیص جفت تقلبی نمی‌شود.

### ۳-۴ پیش‌گیری از بازبسته‌بندی

دیدگاه پژوهش‌های پیشین در زمینه‌ی تشخیص برنامه‌ک‌های بازبسته‌بندی شده، بررسی برنامه‌ک‌های مشکوک پس از بازبسته‌بندی توسط مهاجم بوده‌است اما عمده‌ی پژوهش‌های موجود به جهت پیش‌گیری از بازبسته‌بندی نیازمند ایجاد تغییراتی در برنامه‌ک پیش از انتشار آن به صورت عمومی می‌باشد. این روش‌ها، از بازبسته‌بندی برنامه‌ک توسط متقلب‌ان جلوگیری کرده و یا در صورتی که بازبسته‌بندی صورت بگیرد توسعه‌دهندگان متوجه آن خواهند شد. مزیت بزرگ روش‌های پیش‌گیری از بازبسته‌بندی آن است که از قطعیت بیشتری برخوردار هستند، یعنی در صورتی که بازبسته‌بندی صورت بگیرد اکثراً به صورت قطعی توانایی تشخیص و جلوگیری از انتشار برنامه‌ک‌ها را دارند. به صورت کلی پژوهش‌های اخیر در این حوزه را می‌توان به دو دسته تقسیم کرد که در ادامه به توضیح مختصری از هرکدام و شرح پژوهش‌های هر قسمت می‌پردازیم:

### ۳-۴-۱ روش‌های مبتنی بر نشان‌گذاری

در این دسته از روش‌ها که نیازمند ایجاد تغییراتی در کد برنامه‌ک توسعه‌داده‌شده است، توسعه‌دهندگان قسمتی از امضای خود را تحت عنوان حق‌تکثیر<sup>۸۷</sup> که با نام نشان شناسایی می‌شود، در برنامه‌ک پنهان می‌سازند و در صورتی که برنامه‌ک جدیدی بازبسته‌بندی شود با توجه به تغییر این حق‌تکثیر قادر به تشخیص

<sup>۸۷</sup> Copy Right



و جلوگیری از بازبسته‌بندی خواهند بود. به صورت کلی روش‌های نشان‌گذاری<sup>۷۸</sup> به دو نوع ایستا و پویا تقسیم می‌شوند. در روش‌های ایستا، پیش از ساخته‌شدن برنامه و انتشار آن، حق‌تکثیر ایجاد و در ساختار داده‌ای و یا کدهای برنامه تزریق می‌شود اما در روش‌های پویا، به جهت غلبه بر محدودیت روش‌های ایستا و بررسی دقیق‌تر، حق‌تکثیر توسعه‌دهندگان در حین اجرای برنامه توسط کدهای مورد توسعه ایجاد می‌شود و در ساختار آن قرار می‌گیرد. در روش‌های مبتنی بر نشان‌گذاری از یک روند مشخص برای شناسایی و تایید نشان تزریق‌شده به برنامه‌ها استفاده می‌شود که عدم بازبسته‌بندی را پیش از اجرا و یا در حین اجرا، بسته به ایستا و یا پویا بودن روش، مشخص می‌کند. از طرفی به دلیل عدم قطعیت تشخیص در عمده‌ی روش‌های تشخیص، روش‌های نشان‌گذاری و روش‌های تشخیص برنامه‌های بازبسته‌بندی شده می‌توانند به صورت مکمل یکدیگر عمل کنند. در این حالت ابتدا برنامه‌های مشکوک با استفاده از روش‌های بیان‌شده در حوزه‌ی تشخیص، شناسایی می‌شوند و در ادامه به جهت اطمینان از صحت برنامه مورد نظر، توسط روش‌های مبتنی بر نشان‌گذاری، حق‌تکثیر توسعه‌دهندگان بررسی می‌شود و در صورتی که برنامه مدنظر بازبسته‌بندی شده باشد، از فروشگاه حذف خواهد شد. استفاده از هر دوی این روش‌های سرعت و دقت تشخیص برنامه‌های اندرویدی بازبسته‌بندی شده را افزایش می‌دهد.

یکی از اولین پژوهش‌های این حوزه توسط وو و همکاران [۷۸] انجام شده‌است. در این پژوهش ابتدا رشته‌ای از اعداد و حروف برای تولید نشان مشخص می‌شود. سپس رقم مشخص‌شده در فرایندی مشخص به یک گراف جایگشت نگاشت شده و در ادامه گراف مورد نظر با استفاده از یک قطعه‌کد نشان‌داده می‌شود. برای جلوگیری از مبهم‌نگاری نشان در فرایند بازبسته‌بندی، کد ایجادشده برای گراف نشان‌گذاری، در قسمت‌های مختلف برنامه هدف تقسیم می‌شود. برای تایید برنامه اصلی و جلوگیری از بازبسته‌بندی از یک برنامه ثانویه که حکم فراداده‌ی برنامه اصلی را دارد استفاده می‌شود. برای ساخت برنامه ثانویه، از نمونه‌ی آزمونی که تمامی کد را پوشش دهد استفاده می‌شود هر ورودی آزمون در این نمونه، یک گرداننده دارد که هر تکه از کد نشان‌گذاری شده در قسمت قبل درون آن قرار می‌گیرد. در واقع هدف از تولید برنامه ثانویه، نگه‌داشت مکان‌هایی از برنامه اصلی است که کد نشان‌گذاری در آن قسمت‌ها قرار گرفته‌است چون همانطور که اشاره کردیم جهت جلوگیری از مبهم‌نگاری نشان‌گذاری در حین بازبسته‌بندی، کدهای نشان‌گذاری در قسمت‌های مختلف برنامه اصلی تقسیم می‌شوند. بنابراین در صورتی که برنامه ثانویه، در دسترس متقلب قرار گیرد، می‌تواند تمامی نشان‌گذاری‌های موجود در کد را شناسایی و آن‌ها را حذف کند بنابراین توسعه‌دهندگان نمی‌توانند از بازبسته‌بندی جلوگیری کنند.

ژانگ و همکاران [۷۹]، از یک روش نشان‌گذاری مبتنی بر تصاویر نشان‌گذاری شده استفاده می‌کنند. نویسندگان، با توجه به این که در پژوهش [۷۸] هیچ وابستگی داده‌ای میان کدهای نشان‌گذاری شده وجود

ندارد، روشی را پیشنهاد داده‌اند تا مهاجم نتواند با تحلیل برنامه‌کد و یافتن کدهای نشان‌گذاری شده، در آن‌ها مبهم‌نگاری ایجاد کرده و سیستم تشخیص را دور بزند. در این روش، با استفاده از کاراکترهای اسکی<sup>۸۹</sup>، تصویر ورودی را کدگذاری کرده و طی یک دنباله‌ی تصادفی از رخدادها، کد تولیدشده‌ی اسکی در مسیر رخدادهای تولیدشده تقسیم می‌شود. به جهت این‌که از تغییر نشان‌ها توسط مهاجم جلوگیری شود، تعدادی از کاراکترهای تصویر ورودی را با اعداد ثابت موجود در کد برنامه‌کد جایگزین می‌سازند که موجب می‌شود تصویر ورودی تغییر چندانی نکند و در عین حال وابستگی داده‌ای میان کدهای برنامه‌کد و تصویر ورودی ایجاد شود. یکی از مشکلات دو روش اخیر، عدم امکان بررسی اصالت برنامه‌کد توسط کاربران می‌باشد. برای رفع این مشکل رن و همکاران [۸۰] از کدگذاری بلوکی مبتنی بر ساختارهای خودرمزگشایی<sup>۹۰</sup> استفاده کرده‌اند. در این پژوهش، کد نشان‌گذاری شده پس از تولید، در بلوک‌های شرطی کد برنامه‌کد مورد توسعه قرار گرفته و رمزگذاری صورت می‌گیرد و در حین اجرا رمزگشایی بلوک‌های اجرایی منجر به تایید اصالت کد نشان‌گذاری شده می‌شود.

یکی از مشکلات روش‌های مطرح‌شده در این زمینه آن‌است که جلوگیری از بازبسته‌بندی همواره باید توسط توسعه‌دهندگان صورت گیرد. به همین جهت ایجاد روندی خودکار جهت جلوگیری از اجرای برنامه‌کدهای بازبسته‌بندی شده توسط خود برنامه‌کد ایده‌ای بود که در پژوهش [۸۱] توسط سان و همکاران دنبال شده‌است. در این پژوهش ابتدا تمامی ساختارهای شرطی<sup>۹۱</sup> برنامه‌کد استخراج می‌شود و با استفاده از تحلیل و تغییر شروط به صورتی که منطق برنامه‌کد تغییر نکند، کد نشان‌گذاری شده در ساختار برنامه‌کد جایگذاری می‌شود. در نهایت با استفاده از یک جریان مبتنی بر بارگذاری پویا، در صورتی که کد نشان‌گذاری شده تغییری کرده‌باشد، از اجرای برنامه‌کد جلوگیری خواهد شد.

لوآ و همکاران [۸۲] در سال ۲۰۱۶ روشی را مبتنی بر کلید عمومی توسعه‌دهنده‌ی برنامه‌کد اندرویدی، برای جلوگیری از بازبسته‌بندی ارائه کرده‌اند. نویسندگان در طی روشی به نام شبکه‌ی پنهانی مرتبط<sup>۹۲</sup>، به عنوان ورودی، کد برنامه‌کد اندرویدی به همراه کلید عمومی کاربر توسعه‌دهنده‌ی برنامه‌کد را دریافت می‌کند و با استفاده مقایسه‌ی کلید عمومی کدسخت شده در داخل برنامه‌کد و زیررشته‌ای از محاسبات مبهم‌نگاری شده در برنامه‌کد هنگام اجرای آن، اصالت برنامه‌کد را تشخیص می‌دهند. برای به دست آوردن کلید عمومی در حین اجرای برنامه‌کد، از ویژگی بازتاب در زبان جاوا استفاده شده‌است. یکی از مشکلات این پژوهش آن‌است که به دلیل استفاده از ویژگی بازتاب برای دستیابی به کلید عمومی، حمله‌کننده می‌تواند قسمتی که متد فراخوانی می‌شود را بازنویسی کرده و از طریق ورودی‌های قبلی، متد را بازفراخوانی کند. در این صورت حمله‌کننده به راحتی می‌تواند به کلید عمومی دسترسی پیدا کرده و در ادامه با کمی تحلیل ایستا، قسمتی از برنامه‌کد را که

Ascii<sup>۸۹</sup>

Self Decryption<sup>۹۰</sup>

Conditional Statement<sup>۹۱</sup>

Stochastic Stealthy Network<sup>۹۲</sup>

کلید عمومی در آن کدسخت شده است را بیابد. به جهت مقابله با این نوع از حملات، زنگ در پژوهش [۸۳] راه حلی مبتنی بر بمب‌های منطقی<sup>۹۳</sup> ارائه کرده‌اند. بمب‌های منطقی، قطعه کدهای کوچکی هستند که تنها در حالاتی خاص فعال می‌شوند. در طی این روش، کدهای نشان‌گذاری در ساختار شرط‌های منطقی در برنامهک تزریق می‌شود به صورتی که برخی از این شروط در طی اجرای برنامهک در سیستم عامل کاربران معمولی اجرا می‌شود. در صورتی که مهاجم از روش‌های مبتنی بر جعبه سیاه استفاده ننماید، در آن صورت محیط اجرای وی محدود به تعدادی سیستم عامل و ویژگی‌های مبتنی بر آن خواهد بود، این در حالی است که به صورت کلی کاربران گستردگی بیشتری در تنوع این ویژگی‌ها دارند. بنابراین می‌توان شروط منطقی این بمب‌ها را به گونه‌ای نوشت، که کاربر مهاجم شانس کمی برای دسترسی به نشان‌گذاری داشته باشد. با استفاده از این روش، پژوهش [۸۲] توسعه یافت و تا حدودی مشکلات آن برطرف گردید، اما پژوهش جدید نیز دچار مشکلاتی شد که در ادامه به آن خواهیم پرداخت. در صورتی که کاربر بتواند با تحلیل پویا و فعال‌سازی شروط و در نهایت دسترسی به بلوک بمب‌های منطقی، پیش از اجرای کدهای رمزگشایی شده آن‌ها را تغییر دهد، آنگاه می‌تواند از تشخیص اصالت برنامهک جلوگیری نماید. تانر و همکاران [۸۴] برای حل این مشکل از کدهای بومی استفاده کرده‌اند. از آنجایی که تغییر کدهای بومی در حین بازبسته‌بندی برای مهاجم دشوار و پرهزینه است، نویسندگان پس از دستیابی به کدهای بومی و کدهای اصلی برنامهک، مجموعه‌ای از بایت‌کدهای دالویک رمزنگاری شده را در قسمت‌های مختلف برنامهک اندرویدی اصلی تزریق کرده و متدهای بررسی صحت برنامهک را در کدهای بومی پیاده‌سازی می‌کنند. رمزنگاری بایت‌کدها در کنار فراخوانی توابع بررسی صحت در کدهای بومی، موجب شده است که روش موجود، در مقابل تحلیل ایستا و پویا توسط مهاجم امنیت داشته باشد.

### ۳-۴-۲ روش‌های مبتنی بر مبهم‌نگاری

همانطور که در فصل ۲ اشاره شد، مبهم‌نگاری یکی از روش‌های پیش‌رو برای جلوگیری از بازبسته‌بندی توسط مهاجمان است. توسعه‌دهندگان از مبهم‌نگاری استفاده می‌کنند تا تغییر دادن برنامهک‌های اندرویدی خود را سخت و هزینه‌بر کنند. به صورت کلی مبهم‌نگاری در وهله اول موجب می‌شود تا ساختار منطقی اصلی و به طور مشخص‌تر، بدنه‌ای اصلی منطقی برنامهک از دید مهاجمان مخفی بماند. در مرحله بعدی در صورتی که مهاجم بتواند بدنه‌ی منطقی اصلی برنامهک اندرویدی قربانی را شناسایی کند، تغییر دادن آن و سپس پیاده‌سازی منطق مهاجم به صورت یک بدافزار، کاری سخت، زمان‌بر و هزینه‌بردار خواهد بود، به طوری که ممکن است مهاجم از انجام این عمل امتناع ورزد. همانطور که در فصل پیش دیدیم، برخی از ابزارهای موجود، مبهم‌نگاری در برنامهک‌های اندرویدی را انجام می‌دهند. اما پژوهش‌های موجود در این

<sup>۹۳</sup> Logic Bomb

قسمت نیز، به صورت پیش‌رو، روش‌هایی را توسعه داده‌اند که می‌توانند نقش پیش‌گیری در بازبسته‌بندی برنامه‌های اندرویدی را داشته باشند.

یکی از روش‌های مبهم‌نگاری در برنامه‌های اندرویدی، تغییر نام شناسه‌ها، کلاس‌ها و متدهای برنامه می‌باشد. این روش موجب می‌شود تا تحلیل و جست‌وجوی رشته‌ها در برنامه دشوار شده و مهاجم نتواند با تحلیل برنامه، آن را تغییر دهد. از آنجایی که تغییر نام در بیشتر مواقع تاثیری بر منطق برنامه ندارد، تغییر نام شناسه‌های برنامه، روشی است که ورمک و همکاران [۸۵] از آن استفاده کرده‌اند. پروگارد ابزار دیگری است که با تغییر نام کلاس‌ها، متدها و شناسه‌های موجود در برنامه مبهم‌نگاری را انجام می‌دهد. تغییر نام در پروگارد با استفاده از کاراکترهای محدود اسکی صورت می‌گیرد، به همین جهت دکس‌گارد معرفی شد تا نسخه‌ی پیشرفته‌تری از پروگارد را ارائه دهد که تغییر نام در آن‌ها با دامنه‌ی بیشتر از کاراکترهای اسکی انجام شود [۸۶].

برنامه‌های مبتنی بر زبان جاوا، با استفاده از پیشوندهای مشخص<sup>۹۴</sup> گروه‌بندی می‌شوند و در یک مسیر جاری قرار می‌گیرند که موجب می‌شود کلاس‌های موجود در یک گروه به صورت مستقیم به کلاس‌های هم‌گروهی خود دسترسی داشته باشند. تغییر نام بسته‌های اندرویدی، یکی از روش‌های معمول به جهت جلوگیری در بازبسته‌بندی است چرا که موجب می‌شود تحلیل برنامه برای مهاجم سخت شود. وانگ و همکاران [۸۷] از این روش برای افزایش مقاومت مقابل مبهم‌نگاری استفاده کرده‌اند. در روشی که توسط نویسندگان پیاده‌سازی شده است، تعدادی کلاس کمکی برای مخفی کردن کلاس‌های اصلی برنامه به بسته‌های نرم‌افزاری اضافه شده است. پژوهش [۸۶] نیز در تولید ابزار پروگارد از تغییر نام بسته‌ها استفاده کرده است.

حذف کدهای بومی‌ای که در هنگام اجرای برنامه نیازمند آن‌ها نیستیم، روش دیگری از مبهم‌نگاری برنامه‌های اندرویدی است که در پژوهش [۸۸] از آن استفاده شده است. حذف کدهای بومی موجب می‌شود که خوانایی برنامه در حین دی‌کامپایل توسط مهاجمان کاهش یابد. علاوه بر این موضوع، به جهت خوانایی پایین کدهای بومی نسبت به دالویک بایت‌کد که یک زبان میانی محسوب می‌شود، در پژوهش [۸۸]، نویسندگان قسمتی از برنامه را که شامل الگوریتم‌های رمزنگاری می‌شود، به صورت کدهای بومی پیاده‌سازی کرده که موجب می‌شود خوانایی این الگوریتم‌ها و در نتیجه دسترسی به الگوریتم‌های رمزگشایی سخت شود.

تغییر گراف کنترل جریان، روشی دیگری است که برای مبهم‌نگاری از آن استفاده می‌شود. تحلیل گراف کنترل جریان توسط مهاجمان، یکی از روش‌های محبوب بدافزارنویسان برای وارد ساختن بدافزار و بازبسته‌بندی است. گراف کنترل جریان نشان‌دهنده‌ی روال فراخوانی میان متدهای یک برنامه را نشان می‌دهد. باسی و همکاران [۸۹] با استفاده از افزودن یال به گراف جریان، آن را تغییر می‌دهند. افزودن یال با استفاده از اضافه

<sup>۹۴</sup> NameSpace

کردن یک گره جدید در میان هر فراخوانی و ایجاد فراخوانی واسط انجام می‌شود. پردا و همکاران [۹۰] از همین روش برای تغییر گراف جریان استفاده کرده‌اند با این تفاوت که متد میانی اضافه‌شده، به صورت ایستا و با همان ورودی‌های متد اصلی تعریف می‌شود تا شناسایی متدهای اصلی فراخواننده دشوار شود. افزودن کدهای بیهوده، روشی دیگری است که در پژوهش‌های [۹۱، ۹۲، ۹۳]، استفاده شده‌است. در پژوهش [۹۱]، لی و همکاران، با استفاده از افزودن کدهای nop، پرش‌های قطعی<sup>۹۵</sup> و ثبات‌های داده<sup>۹۶</sup> پردازش‌های بیهوده‌ای در برنامهک اضافه می‌کنند که منطق برنامهک را تغییر نمی‌دهد. راستوگی در پژوهش [۹۳]، حالتی از دستورات شرطی را معرفی کرد که همواره یک نتیجه می‌دهد اما گراف جریان را به دو شاخه تقسیم می‌کند. در این حالت یکی از حلقه‌های این دستور، شامل کد اصلی خواهد بود و حلقه‌ی دیگر شامل پرش بدون شرط به ابتدای کد اصلی. ژانگ [۹۲] در پژوهش خود از این روش برای مبهم‌نگاری برنامهک‌های اندرویدی استفاده کرده‌است. جابه‌جایی دستورات، روش دیگری است که در پژوهش‌های [۸۶، ۹۱، ۹۰] استفاده شده‌است. باسی و همکاران [۹۱] با استفاده از جابه‌جایی دستورات به صورت کاملاً تصادفی، با دانه<sup>۹۷</sup>‌های ناهم‌گون و در نهایت استفاده از دستوران پرش، خطوط کدهای اصلی را جابه‌جا کرده و به این شکل مهاجم را در تحلیل کدهای اصلی برنامهک، ناکام می‌گذارد. پردا و همکاران [۹۰] با تحلیل دالویک بایت‌کدها، بلوک‌های مستقل اجرایی در آن‌ها را استخراج کرده و ترتیب آن‌ها را عوض می‌کنند. برای تضمین اجرای منطق اصلی برنامهک، از شرط‌های منطقی قطعی استفاده شده‌است.

استفاده از رمزگذاری در حین کامپایل برنامهک‌های اندرویدی و دسترسی به اطلاعات در حین اجرا، یکی دیگر از روش‌های مبهم‌نگاری است که در سال‌های اخیر محبوبیت زیادی پیدا کرده‌است. روش‌های رمزگذاری را به صورت کلی می‌توان به دو دسته‌ی رمزگذاری داده و رمزگذاری منابع برنامهک تقسیم کرد. در روش‌های مبتنی بر رمزگذاری داده، قسمت‌های حساسی از کد برنامهک، مانند کلید دسترسی به واسط‌های برنامه‌نویسی و یا گذرواژه‌های مهم مانند آنچه برای درخواست از پایگاه‌داده مورد استفاده قرار می‌گیرد، رمزگذاری می‌شود. باسی و همکاران [۸۹] با استفاده از این ایده و استفاده از رمزگذاری مبتنی بر سیستم رمز سزار<sup>۹۸</sup>، گذرواژه‌های حساس و حیاتی برنامهک را رمزنگاری و در هنگام اجرا رمزگشایی می‌کند. منابع برنامهک‌های اندرویدی، نیازمند کامپایل توسط ماشین مجازی جاوا نیستند و توسط کدهای برنامهک، فراخوانی می‌شوند. پردا و همکاران [۹۰] با استفاده از چکیده‌سازی ۱۲۸ بیتی مبتنی بر MD5، فایل‌های منابع را چکیده و آن‌ها را در مسیر جدیدی مقداردهی می‌کند. از مشکلات این پژوهش آن‌است که نویسندگان به درستی، نحوه‌ی کارکرد کلاسی را که وظیفه‌ی بارگیری پویای فایل‌های منابع رمز شده از بر عهده دارد، مشخص نکردند. علاوه بر رمزگذاری منابع و اطلاعات حساس، رمزگذاری کلاس‌های برنامهک نیز روش دیگری

Unconditional Jump<sup>۹۵</sup>

Data Register<sup>۹۶</sup>

Seed<sup>۹۷</sup>

Caesar Cipher<sup>۹۸</sup>

است نویسندگان پژوهش [۹۴] از آن استفاده کرده‌اند. در این پژوهش، هر یک از کلاس‌های اندرویدی پس از رمزگذاری، در آرایه‌هایی از جنس بایت ذخیره می‌شود و کلاسی مجزا در هنگام اجرا، با استفاده از قابلیت بازتاب در برنامه‌های جاوا، کلاس رمزگذاری شده را بازیابی و اجرا می‌کند.

به صورت کلی می‌توان تاثیر پژوهش‌های مبهم‌نگاری بر برنامه‌های اندرویدی را از سه دیدگاه بررسی کرد. در حالت اول مبهم‌نگاری‌هایی نظیر تغییرنام و یا اجرای کلاس‌های رمزگذاری شده، منجر به افزایش امنیت و جلوگیری از بازبسته‌بندی می‌شود. در حالت دوم، رمزگذاری داده‌ها و یا استفاده از کدهای بومی در هنگام اجرا، مانع از اجرای تحلیل بدافزار روی برنامه مورد نظر می‌شود. در حالت پایانی، روش‌های مبهم‌نگاری نظیر چکیده‌سازی، تغییر نام و یا حذف کدهای مرده و بیهوده، منجر به افزایش کارایی و بهینه‌شدن اجرای برنامه‌های اندرویدی می‌شود. از آنجایی که مهاجمان نیز از روش‌های مبهم‌نگاری استفاده می‌کنند، بنابراین آشنایی با پژوهش‌های اخیر موجود در این حوزه می‌تواند ما را در ارائه‌ی روشی با دقت بیشتر و سرعت بالاتر یاری دهد.

### ۳-۵ مقایسه‌ی روش‌ها

با توجه به فصلی که گذشت می‌توان گفت که اکثر پژوهش‌های موجود در زمینه‌ی تشخیص برنامه‌های بازبسته‌بندی شده، نیازمند تقویت ایده‌ی اولیه هستند و هنوز پژوهش کاملی در این زمینه که علاوه بر دقت بالا، سرعت تشخیص مناسبی نیز داشته باشد، توسعه داده نشده‌است. اما به صورت کلی، در مقایسه‌ی روش‌های پویا و ایستا می‌توان گفت که عمده‌ی روش‌های موجود با تمرکز بر تحلیل ایستا سعی در افزایش کارایی روش‌های پیشین دارند. علت این موضوع آن است که تحلیل ایستا پاسخ‌گوی طیف وسیعی از برنامه‌های اندرویدی خصوصاً برنامه‌های رایگان است در حالی که پژوهش‌های موجود در زمینه‌ی تحلیل پویا، اکثراً از مجموعه‌ی داده‌ی آزمون کوچک و محدودی، خصوصاً در زمینه‌ی برنامه‌های تجاری، استفاده کرده‌اند. علت دیگر این موضوع آن است که تهدید اصلی کاربران در زمینه‌ی برنامه‌های اندرویدی بازبسته‌بندی شده، از سمت برنامه‌های رایگان رخ می‌دهد و برنامه‌های غیررایگان و تجاری، معمولاً پیش از توسعه به طور کاملی از روش‌های جلوگیری از بازبسته‌بندی استفاده می‌کنند. پژوهش پیاده‌سازی شده در این پایان‌نامه، مبتنی بر تحلیل ایستا با استفاده از گراف و استخراج ویژگی‌های مقاوم در مقابل مبهم‌نگاری است، بنابراین در ادامه‌ی این پژوهش، مقایسه‌ی روش‌های مبتنی بر تحلیل ایستا را انجام خواهیم داد. اگر به طور کلی دو معیار دقت و سرعت اجرا را در دسته روش‌های مبتنی بر تحلیل ایستا در نظر بگیریم، روش‌های مبتنی بر گراف از دقت بالایی در تشخیص برخوردار هستند چرا که تغییر گراف‌های حاصل از فراخوانی و یا اجرای برنامه سخت و هزینه‌بر است اما به صورت کلی، از آنجایی که الگوریتم‌های مقایسه‌ی گرافی، به خصوص الگوریتم‌های تشخیص گراف هم‌ریخت، سرعت بسیار پایینی در مقایسه دارند، معمولاً

روش‌های این دسته به قدری کند هستند که عملاً استفاده از آن‌ها در یک محیط صنعتی ناممکن خواهد بود. از طرفی روش‌های مبتنی بر تحلیل ترافیک و تحلیل منابع، خصوصاً در روش‌هایی که از طبقه‌بندی استفاده می‌شود، از سرعت بالایی در تشخیص برخوردار هستند اما به جهت این‌که ایجاد مبهم‌نگاری در منابع برنامه‌آسان و کم‌هزینه‌است، بنابراین بازبسته‌بندی ممکن است مبتنی بر منابع برنامه‌آسان نیز اتفاق بیافتد. علاوه بر این، هدف بسیاری از برنامه‌های بازبسته‌بندی، فریب کاربران ناآگاه با استفاده از ایجاد یک واسط کاربری مشابه با برنامه اصلی می‌باشد، بنابراین بررسی این دسته از ویژگی‌ها می‌تواند ما را در تشخیص برنامه‌های اندرویدی بازبسته‌بندی شده یاری دهد. در جدول زیر سعی داریم که پژوهش‌های مطرح مبتنی بر تحلیل پویا را از نظر سه خاصیت یعنی جست‌وجوی دودویی، سربار زمانی و دقت تشخیص مورد مقایسه قرار دهیم.

دسته‌بندی روش‌ها	پژوهش‌کنندگان	سربار محاسباتی	دقت تشخیص	مقایسه‌ی دودویی
مبتنی بر آپکد و دنباله‌ی دستورات	ژو و همکاران	متوسط	پایین	✓
	ژو و همکاران			✓
	دزنور و همکاران			✓
	جرومه و همکاران			×
	سرنیل و همکاران			×
	لین و همکاران			×
مبتنی بر گراف	سان و همکاران	زیاد	خوب	✓
	هو و همکاران	متوسط	خوب	✓
	ژو و همکاران	زیاد	خوب	✓
	چن و همکاران	کم	خوب	✓
	جنگ و همکاران	زیاد	خوب	✓
	وانگ	زیاد	خوب	✓
	ترکی	زیاد	خوب	✓
	الشهری و همکاران	متوسط	پایین	✓
مبتنی بر تحلیل ترافیک شبکه	هه و همکاران	متوسط	پایین	×
	مالک و همکاران	متوسط	پایین	✓
	پوچر و همکاران	متوسط	پایین	✓
	شارما و همکاران	متوسط	متوسط	×
	شائو و همکاران	کم	خوب	×
مبتنی بر تحلیل منابع	سان و همکاران	زیاد	خوب	✓
	هو و همکاران	متوسط	خوب	✓
	لین و همکاران	متوسط	پایین	✓
	گوا و همکاران	زیاد	متوسط	✓
	لیو و همکاران	متوسط	خوب	✓
	ما و همکاران	متوسط	خوب	✓



## فصل ۴

### راهکار پیشنهادی

تقریباً در تمامی راهکارهای ارائه شده جهت تشخیص برنامه‌های باز بسته بندی شده، نیازمند انجام مقایسه‌ی دودویی ویژگی‌های استخراج شده از برنامه‌ها در راستای تشخیص جفت تقلبی هستیم. به همین جهت اصولاً تشخیص جفت برنامه باز بسته بندی شده از میان مخزن برنامه‌ها روشی زمان‌بر و پرهزینه است. از طرفی همانطور که بررسی شد، برای تشخیص برنامه باز بسته بندی شده، نیازمند ویژگی‌هایی به صورت ایستا و یا پویا از برنامه‌ها هستیم که رفتار منطقی و غیرمنطقی آن‌ها را نشان دهد. از بررسی کارهای پیشین و مطالعه در زمینه‌ی تحلیل برنامه‌های اندرویدی، می‌توان دریافت که پارامترهای دقت و سرعت در تشخیص برنامه‌های باز بسته بندی شده، رابطه‌ی مستقیمی با ویژگی‌های منتخب به جهت مقایسه‌ی دودویی برنامه‌ها دارد. پیچیدگی ویژگی‌های منتخب منجر به کاهش سرعت و مقاومت آن‌ها در مقابل روش‌های مبهم‌نگاری، منجر به افزایش دقت در تشخیص جفت تقلبی خواهد شد. در این فصل ابتدا نمای کلی از پژوهش را توضیح می‌دهیم و مولفه‌های اصلی پژوهش را به صورت مختصر بررسی خواهیم کرد. سپس هر کدام از مولفه‌های مطروحه را با جزئیات بیشتری بررسی و معایب و مزایای هر کدام را شرح خواهیم داد.

#### ۴-۱ نمای کلی

از آنجایی که مقایسه‌ی دودویی برنامه و ورودی با تمامی برنامه‌های موجود در مخزن هزینه‌ی محاسباتی زیادی را به پژوهش تحمیل می‌کند، استفاده از یک مرحله پیش پردازش می‌تواند به افزایش سرعت تشخیص کمک کند. همانطور که در شکل — مشخص شده است، در این پژوهش از یک طبقه‌بند نزدیک‌ترین همسایه به عنوان مرحله‌ی پیش پردازش ورودی‌ها استفاده شده است. در این مولفه با استفاده از تعدادی

ویژگی مبتنی بر واسطه‌های کاربری، برنامه‌های موجود در مخزن و برنامه و ورودی طبقه‌بندی و تعدادی از نزدیک‌ترین همسایه‌های برنامه و ورودی به عنوان برنامه‌های مشکوک به جهت بررسی دقیق‌تر به مرحله‌ی بعد خواهند رفت. پژوهش به صورت کلی از دو مولفه‌ی «تشخیص کتابخانه‌های اندرویدی» و «تشخیص برنامه بازبسته‌بندی شده» تشکیل شده‌است. مولفه‌ی «تشخیص کتابخانه‌های اندرویدی» با الهام از پژوهش [۵۱] و همراه با تغییر امضای کلاسی به جهت افزایش سرعت، پیاده‌سازی شده‌است. مخزن کتابخانه‌های اندرویدی و برنامه هدف، به عنوان ورودی به این مولفه داده می‌شوند. به صورت کلی در این مولفه کلاس‌های هر برنامه، با تمامی کلاس‌های موجود در مخزن کتابخانه‌ها مقایسه می‌شود. برای کاهش فضای مقایسه‌ی کتابخانه‌ها با کلاس‌های برنامه هدف، از «ماژول فیلتر کتابخانه‌ها» استفاده شده‌است. این ماژول تعداد کلاس‌های مورد بررسی در هر کتابخانه را با استفاده از دو «فیلتر ساختاری» و «طول امضا» کاهش می‌دهد و در نهایت با استفاده از توابع چکیده‌سازی محلی، مقایسه‌ی کلاس‌های برنامه هدف و مخزن کتابخانه‌های اندرویدی در ماژول «استخراج نگاشت» انجام می‌شود. حل مسئله‌ی نگاشت میان کلاس‌های برنامه و کلاس‌های کتابخانه‌ای، همان حل مسئله‌ی تخصیص است که در آن نگاشتی از کمترین هزینه (مبتنی بر چکیده‌سازی محلی) از یک گراف دوبخشی استخراج می‌شود.

ب مولفه‌ی دیگری که در این پژوهش پیاده‌سازی شده‌است، مولفه‌ی تشخیص برنامه‌های بازبسته‌بندی شده است. این مولفه، در یک پیش‌پردازش با استفاده از مولفه‌ی تشخیص کتابخانه‌ها، نگاشتی از کلاس‌های برنامه و کلاس‌های کتابخانه‌ای مخزن به دست می‌آورد. سپس دو برنامه مورد بررسی و نگاشتی از کلاس‌های هر دو، به عنوان ورودی ماژول طبقه‌بند، وارد می‌شود. برای کاهش فضای مقایسه‌ی دودویی میان برنامه‌های اندرویدی، از یک طبقه‌بند نزدیک‌ترین همسایه مبتنی بر ویژگی‌های واسطه‌های کاربری استفاده شده‌است. این ماژول تعدادی از نزدیک‌ترین همسایه‌های برنامه هدف را به عنوان برنامه‌های مشکوک به ماژول مقایسه‌ی دودویی ارسال می‌کند. در این ماژول ابتدا کدهای کتابخانه‌ای با استفاده از نگاشت حاصل از مرحله‌ی قبلی حذف می‌شود و کلاس‌های باقی‌مانده برای ساخت امضا توسط ماژول ساخت امضای برنامه، تحلیل می‌شوند. در نهایت با استفاده از چکیده‌سازی محلی و تعیین حد‌آستانه، تعیین بازبسته‌بندی صورت می‌گیرد.

## ۴-۲ تشکیل امضای کلاس

در این پژوهش، مولفه‌ی تشخیص کتابخانه‌های اندرویدی و تشخیص بازبسته‌بندی، هر دو از امضای کلاس برای شایه‌سنجی استفاده کرده‌اند. امضای کلاس متشکل از مهم‌ترین ویژگی‌های آن است که علاوه بر مدل‌کردن رفتار کلاس، مقاومت بالایی در مقابل روش‌های مبهم‌نگاری داشته‌باشد. در ابتدای

این قسمت، ابتدا طرح کلی از امضای هر متد را توضیح می‌دهیم، سپس چگونگی ساخته شدن امضای کلاس را با استفاده از امضای متدهای درون آن شرح خواهیم داد. استخراج ویژگی‌های مبتنی بر کلاس در برنامه‌های اندرویدی در این قسمت با استفاده از چارچوب سوت انجام شده است. چارچوب سوت ابزاری مبتنی بر زبان جاوا برای تحلیل ایستای برنامه‌های مبتنی بر این زبان است. از پارامترهای زیر برای ساخت امضای متد استفاده شده است:

- **تغییردهنده‌ی متد:** تغییردهنده‌های هر متد، مشخص‌کننده‌ی چگونگی پیاده‌سازی متدهای جاوا هستند که اکثراً به سختی قابل تغییر می‌باشند چرا که در صورت تغییر آن‌ها بدنه‌ی متد الزاماً باید به صورت کلی تغییر کند، بنابراین استفاده از این ویژگی، مقاومت بالایی مقابل مبهم‌نگاری را به ارمغان می‌آورد. برای ساخت امضای متد از ۴ نوع تغییردهنده‌ی بومی، انتزاعی، ایستا و سازنده استفاده شده است. متدهای بومی، شامل دستورات به زبان بومی جاوا هستند که توسط پردازنده به صورت مستقیم اجرا می‌شود. متدهای ایستا توسط تمامی نمونه‌های کلاس قابل دسترسی هستند و تنها به متغیرهای ایستا در کلاس دسترسی دارند. متدهای سازنده، در واقع وظیفه‌ی مقداردهی اولیه به متغیرهای پویا در هنگام ساخت نمونه‌ی کلاسی در برنامه‌های اندرویدی را بر عهده دارند.

- **نوع داده‌ی خروجی و ورودی متد:** ویژگی دیگری که در امضای متد مورد استفاده قرار گرفته است، نوع داده‌ی متغیرهای خروجی و ورودی به هر متد می‌باشد. به دلیل آن‌که تغییر ورودی و خروجی متدها غالباً بدون تغییر بدنه‌ی اصلی توابع امکان‌پذیر نیست، بنابراین تغییر آن‌ها منجر به هزینه‌ی زیادی برای متقلبان است، چرا که نیازمند تغییر بدنه‌ی متد، به طوری که منطق اصلی متد حفظ شود، می‌باشد.

- **متدهای فراخوانی شده:** در این قسمت از دو دسته‌ی مهم از متدهایی که در طول بدنه‌ی برنامه فراخوانی شده اند استفاده شده است. دسته‌ی اول متدهایی هستند که متدهای کتابخانه‌ای جاوا نیستند و به صورت غیرایستا، توسط توسعه‌دهنده پیاده‌سازی شده‌اند. لازم به ذکر است که به دلیل آن‌که در قسمت تشخیص برنامه‌های اندرویدی بازبسته‌بندی شده، کدهای کتابخانه‌ای حذف شده‌اند، بنابراین تعریف متدهای فراخوانی شده در این قسمت، بدون در نظر گرفتن متدهای کتابخانه‌ای اندرویدی خواهد بود. دسته‌ی دوم متدهای فراخوانی شده، متدهای زبان جاوا هستند که داخل بسته‌ی کتابخانه‌ای جاوا حضور دارند. تغییر این دسته از متدها آسان‌تر از متدهای قسمت قبلی است، اما در هر حال نمی‌توان بدون پرداخت هزینه‌ی محاسباتی آن‌ها را حذف کرد.

- **فراخوانی واسطه‌های برنامه‌نویسی:** واسطه‌های برنامه‌نویسی، هسته‌ی اصلی رفتار هر متد در برنامه‌های اندرویدی هستند که بدافزارنویسان نیز از آن‌ها استفاده می‌کنند. تغییر این دسته از فراخوانی‌ها در

متهای کلاسی، سخت و با پیچیدگی همراه است.

در امضای هر کلاس نیز از ویژگی‌های زیر استفاده شده است:

- **هسته‌ی کلاس:** هسته‌ی کلاس شامل امضای تمامی متهای موجود با استفاده از ویژگی‌های بخش قبل.
- **کلاس‌های داخلی:** از آنجا که تغییر سلسله مراتب کلاس‌های جاوایی، اکثراً ناممکن است بنابراین امضای تمامی کلاس‌های داخلی به همراه متهای پیاده‌سازی شده در این قسمت استفاده می‌شود. تغییر سلسله مراتب کلاس‌های داخلی نیازمند تغییر تمامی فراخوانی‌های موجود از کلاس‌های مورد نظر است بنابراین تغییر سلسله مراتب تقریباً ممکن نیست.
- **سطح کلاس:** همانطور که گفته شد، تغییر سلسله مراتب کلاس‌ها سخت و زمان‌بر است بنابراین شماره‌ی سطح کلاس در ساختار سلسله‌مراتبی ویژگی مقاوم دیگری است که در این قسمت استفاده می‌شود.
- **کلاس پدر:** تغییر ساختار ارث‌بری کلاس‌ها، هزینه‌بر است چرا که تغییر این ساختار نیازمند تغییر فراخوانی‌های متعدد ناشی از ساختن نمونه‌های کلاسی است. بنابراین در امضای هر کلاس، از امضای پدر آن کلاس نیز استفاده شده است.
- **نام کلاس داخلی و بیرونی:** برای بررسی ریزدانه‌تر، علاوه بر سطح کلاس، از نام کلاس‌های درونی و بیرونی نیز استفاده شده است. اکثر مبهم‌نگارها، از دانه‌ی یکسانی برای تولید نام‌های تصادفی استفاده می‌کنند بنابراین ایجاد نام‌های متنوع از یک نام نیازمند تغییر نام اولیه‌ی متد است.
- **طول هسته‌ی کلاس:** طول هسته‌ی کلاس، شامل متهای کلاسی و کلاس‌های درونی، ویژگی دیگری است که در امضای کلاس مورد استفاده قرار گرفته است.

## ۳-۴ ساخت امضا

از آنجایی که امضا برای هر کلاس تولید می‌شوند، بنابراین به صورت کلی دو مرحله برای تولید امضای کلاس استفاده شده است. در ابتدا با استفاده از ویژگی‌های مبتنی بر متد، امضای متهای پویای کلاس استخراج شده و سپس با استفاده از طرحی که در ادامه توضیح داده خواهد شد امضای کلاس ساخته

می‌شود. در ادامه ابتدا نحوه‌ی ساخت امضای متد را توضیح می‌دهیم، سپس به شرح امضای هسته‌ی کلاس و امضای کلاس می‌پردازیم: —————

## ۱-۳-۴ امضای متد

به صورت کلی امضای متد از ۵ قسمت متفاوت تشکیل شده‌است:

< *Modifier, RetType, InputType, JLibMethodCallee, NonStaticAppMethodCallee, ApiCallSootSignature* >

شکل ۱-۴: شمای کلی امضای متد در پژوهش

- *Modifier*: در این بخش رشته‌ای شامل تغییردهنده قرار می‌گیرد. همانطور که در بخش پیشین بررسی کردیم، تغییردهنده‌ی متد، شامل یکی از انواع بومی، انتزاعی، ایستا و یا سازنده‌است. در صورتی که تغییردهنده‌ی متد هیچ‌کدام از موارد ذکر شده نباشد، مقدار *Null* به جای این ویژگی قرار می‌گیرد.
- *RetType*: نوع داده‌ی بازگشتی از متد، به صورت رشته در این قسمت قرار می‌گیرد. در صورتی که نوع داده‌ی متد، از انواع داده‌ای زبان جاوا نباشد، نام کلاس این نوع داده‌ای به صورت رشته جایگزین می‌شود.
- *InputType*: نوع داده ورودی‌های متد، مرتب شده به صورت الفبایی، به صورت رشته در این قسمت قرار می‌گیرند. در صورت نبودن نوع داده در زبان جاوا، همانند *RetType* عمل می‌شود.
- *JLibMethodCallee*: رشته‌ای متشکل از تمامی توابع کتابخانه‌ای جاوا، که در بدنه‌ی تابع صدا زده شده اند در این قسمت قرار می‌گیرد. رشته‌ی حاصل به صورت الفبایی مرتب شده‌است.
- *NonStaticAppMethodCallee*: رشته‌ی مرتب شده به صورت الفبایی شامل نام تمامی توابع غیر کتابخانه‌ای و غیرایستا که توسط تابع فراخوانی شده‌اند. لازم به ذکر است، از آنجایی که فراخوانی توابع ایستا، مقاومت بالایی در مقابل مبهم‌نگاری ندارد، در این قسمت از توابع غیر ایستا استفاده شده‌است. افزودن فراخوانی توابع ایستا، یکی از روش‌های محبوب در مبهم‌نگاری برنامه‌های اندرویدی است.

- *ApiCallSootSignature*: در این قسمت رشته‌ای از سه ویژگی مهم توابع واسطه‌های برنامه‌نویسی، شامل نوع کلاس متد فراخوانی، کلاس نوع بازگشتی متد و نام متد قرار می‌گیرد. تغییر نام واسطه‌های برنامه‌نویسی قطعاً نیازمند تغییر کتابخانه‌های اندرویدی و ایجاد مبهم‌نگاری در آن‌ها است، به همین دلیل در این قسمت نام متد، مقاوم در مقابل مبهم‌نگاری برنامه‌نامک می‌باشد.

*< DeclaringClass, RetType, MethodName >*

شکل ۲-۴: شمای کلی امضای واسطه‌های برنامه‌نویسی

## ۲-۳-۴ امضای کلاس

امضای کلاس، مطابق با شکل زیر، شامل ۶ قسمت است که در ادامه به بررسی هر کدام می‌پردازیم:

*< ClassCoreSig, InnerClassesSig, InheritedClassesSig, ImpInterfacesSig, ClassLevel, InnerOuterClassName, ClassLen >*

شکل ۳-۴: شمای کلی امضای کلاس

- *ClassCoreSig*: رشته‌ی حاصل از الحاق امضای تمامی متدهای کلاس در این قسمت قرار می‌گیرد. ترتیب قرارگرفتن امضای متدهای کلاس، بر اساس خطوطی است که پیاده‌سازی شده‌اند، یعنی اگر متد *A* در کلاس *B* پیش از متد *C* پیاده‌سازی شده‌بود، آنگاه امضای متد *A* پیش از متد *C* قرار خواهد گرفت.

*< Method\_1, Method\_2, Method\_3, Method\_3, ..., Method\_n >*

شکل ۴-۴: شمای کلی هسته‌ی کلاس

- *InnerClassesSig*: رشته‌ی حاصل از الحاق امضای کلاس‌های داخلی کلاس مورد نظر، مرتب‌شده به صورت الفبایی در این قسمت قرار خواهد گرفت.
- *InheritedClassesSig*: در صورتی که کلاس پدر، یکی از کلاس‌های کتابخانه‌ای جاوا باشد، نام کلاس پدر در این قسمت قرار می‌گیرد، اما در غیر این صورت، امضای کلاس پدر، شامل *CoreClassesSig* و *InnerClassesSig* به صورت رشته در این قسمت قرار می‌گیرد.

- *ImpInterfacesSig*: در صورتی که واسط پیاده‌سازی شده، از نوع کلاس‌های کتابخانه‌ای جاوا باشد، نام آن به صورت رشته در این قسمت قرار خواهد گرفت و در غیر این صورت همانند قسمت قبل، امضای واسط‌ها، شامل الحاق *CoreClassesSig* و *InnerClassesSig* تمامی آن‌ها، مرتب شده به صورت الفبایی خواهد بود.
- *ClassLevel*: عمق کلاس در ساختار سلسله‌مراتبی کلاس‌های هر بسته به صورت عددی در این قسمت قرار خواهد گرفت. برای مثال اگر کلاس مورد نظر، ریشه‌ی درخت سلسله‌مراتبی بسته‌ها باشد، عدد ۰ و اگر فقط یک کلاس بیرون داشته باشد، عدد ۱ در این قسمت قرار خواهد گرفت.
- *InnerOuterClassesName*: نام کلاس بیرونی و کلاس‌های درونی مرتب‌شده به صورت الفبایی به صورت رشته در این قسمت قرار خواهد گرفت.
- *ClassLen*: در این قسمت طول رشته‌ی امضای کلاس، به صورت عددی درج می‌شود.
- *NumOfInnerClass*: تعداد کلاس‌های داخلی کلاس هدف، به صورت عددی در این قسمت قرار می‌گیرد.

## ۴-۴ تشخیص برنامه‌های بازبسته‌بندی شده

در این قسمت، به بررسی ماژول‌های مولفه‌ی تشخیص برنامه‌های بازبسته‌بندی مطابق با شکل — می‌پردازیم. همانطور که در شکل — آمده‌است، مولفه‌ی تشخیص کدهای برنامه‌های بازبسته‌بندی شده، شامل سه ماژول پیدا جداسازی کدهای کتابخانه‌ای، یافتن  $K$  نزدیک‌ترین همسایه و مقایسه‌ی دودویی می‌باشد. در ادامه هر کدام از ماژول‌های این مولفه را به دست بررسی می‌کنیم.

### ۴-۴-۱ مولفه‌ی تشخیص کدهای کتابخانه‌ای

این مولفه، دارای سه ماژول اصلی یافتن کلاس‌های کاندید، فیلتر کلاس‌های کاندید مبتنی بر فیلترهای ساختاری و طولی و در نهایت ماژول یافتن نگاشت میان کلاس‌های برنامه و کلاس‌های کتابخانه‌ای است.

آسان‌ترین روش یافتن کلاس‌های کتابخانه‌ای مقایسه‌ی دودویی کلاس‌های موجود در مخزن کتابخانه‌ها با تمامی کلاس‌های برنامه می‌باشد اما این روش بسیار پرهزینه است، در مرحله‌ی اول از تشخیص کدهای کتابخانه‌ای با استفاده از دو فیلتر ساختاری و طولی تعداد کلاس‌های مورد مقایسه با برنامه مقصد را

کاهش می‌دهیم. در قسمت ابتدایی توسط فیلتر طولی، تنها کلاس‌هایی را بررسی خواهیم کرد که طول آن‌ها از کلاس مبدا از یک حد آستانه کوچکتر و یا بیشتر باشد. در قسمت فیلتر ساختاری، با استفاده از بررسی و تحلیل چندین ویژگی روی کلاس‌های کتابخانه‌ای، مجموعه‌ی کلاس‌هایی که در خروجی فیلتر طولی ظاهر شده‌اند را مجدداً کاهش می‌دهیم. ویژگی‌های مورد بررسی در این قسمت شامل تعداد کلاس‌های بیرونی و داخلی، ارث‌بری کلاسی کلاس هدف، ارث‌بری از کلاس‌های کتابخانه‌ای جاوا و یا اندروید، تعداد واسط‌های پیاده‌سازی شده، واسط‌های پیاده‌سازی شده و معادل بودن کلاس‌های بیرونی می‌باشد. کلاس خروجی پس از فیلتر طولی، تنها در صورتی پس از اعمال فیلتر ساختاری نیز حضور خواهد داشت که در تمامی ویژگی‌های ذکر شده با کلاس هدف تطابق داشته‌باشد.

در این ماژول تمامی کلاس‌های برنامه‌ک و یک کلاس هدف از میان کلاس‌های کتابخانه‌ای به عنوان ورودی به ماژول مورد نظر پاس داده می‌شود و در نهایت لیستی از ماژول‌های کلاس‌های برنامه‌ک به عنوان کلاس‌های مشابه در خروجی مشخص می‌شود. شبکه‌کد الگوریتم پیدا کردن کلاس‌های کاندید در شکل ۱ مشخص شده‌است. پس از اجرای ماژول کلاس‌های کاندیدی، با استفاده از ماژول یافتن کتابخانه‌ها، کتابخانه‌هایی که کلاسی مشابه در برنامه‌ک دارند را شناسایی می‌کنیم. به جهت شناسایی کتابخانه‌های موجود در برنامه‌ک، پیش از اجرای مولفه‌ی تشخیص کدهای کتابخانه‌ای، مخزنی شامل کتابخانه‌های اندرویدی مشهور ایجاد می‌کنیم و تمامی کلاس‌های کتابخانه‌ای را در این مخزن قرار می‌دهیم. سپس تعداد  $N$  کلاس با طول بیشینه از هر کتابخانه را انتخاب و به عنوان کلاس‌های هدف قرار می‌دهیم. در ادامه، الگوریتم پیدا کردن کلاس‌های کاندید ۱ را برای هر کدام از کلاس‌های هدف کتابخانه‌ی *lib* اجرا کرده و تعدادی کلاس کاندید از میان کلاس‌های برنامه‌ک انتخاب می‌کنیم. در این مرحله، کلاس‌های کاندید برنامه‌ک را با کلاس *Cl* از کتابخانه‌ی *lib* با استفاده از روش‌های چکیده‌سازی محلی مقایسه کرده و در صورتی که میزان تشابه دو کلاس بیش از یک حد آستانه باشد، کلاس کاندید مورد نظر با کلاس *Cl* مطابقت یافته و برای نگاشت میان کلاس‌های کتابخانه‌ای و کلاس‌های برنامه‌ک، به مرحله‌ی بعد خواهد رفت. طبیعتاً در صورتی که کلاسی در کتابخانه‌ی *lib* حضور داشته‌باشد، که در خروجی ماژول فیلتر کتابخانه‌ها، حضور پیدا کند، بنابراین کتابخانه‌ی مورد نظر در کد برنامه‌ک حضور دارد و برای انجام نگاشت میان کلاس‌های کتابخانه‌ای و کلاس‌های برنامه‌ک، به مجموعه‌ی *Libs* اضافه می‌شود. شبکه‌کد ماژول پیدا کردن کتابخانه‌ها را می‌توان در الگوریتم — مشاهده نمود.

ماژول نگاشت، به جهت ایجاد نگاشتی میان کلاس‌های کتابخانه‌ای (خروجی ماژول فیلتر کتابخانه‌ها) و کلاس‌های برنامه‌ک پیاده‌سازی شده‌است. ورودی ماژول لیست *ClsLib* شامل کلاس‌های کتابخانه‌ای برنامه‌ک، حاصل از اجرای الگوریتم ۲ و لیست *ClsApp* کلاس‌های برنامه‌ک هدف می‌باشد. خروجی ماژول نگاشتی یک به یک و پوشا از کلاس‌های کتابخانه‌ای به کلاس‌های برنامه‌ک می‌باشد و مشخص می‌کند کلام



یک از کلاس‌های برنامه، کلاس‌های کتابخانه‌ای هستند.

$$f : L \rightarrow A \quad L \subseteq Cls_{lib} \quad , A \subseteq Cls_{app} \quad (۱-۴)$$

---

### الگوریتم ۱ پیدا کردن کلاس‌های کاندید

---

**ورودی:** کلاس هدف برای یافتن لیستی از کلاس‌های مشابه با آن  $t_c$ ، لیستی از کلاس‌های برنامه  $S_{app}$

**خروجی:** لیستی از کلاس‌های کاندید برنامه مشابه با کلاس هدف  $t_c$

۱: قرار بده  $Candidate = \emptyset$

۲: قرار بده  $L = length(Sig_{t_c})$  مقدار امضای نهایی کلاس  $t_c$  می‌باشد

۳:  $\Delta$  فیلتر طولی کلاس‌های برنامه

۴: به ازای  $Class \in S_{app}$ :

۵: اگر  $L - T_L < length(Sig_{Class}) < L + T_L$ :

۶:  $Candidate \cup Class$

۷:  $\Delta$  فیلتر ساختاری کلاس‌های برنامه

۸: به ازای  $Class \in Candidate$ :

۹: قرار بده  $Condition = True$

۱۰: قرار بده  $Condition = Condition \wedge (\#outerClass(Class) = \#outerClass(t_c))$

۱۱: قرار بده  $Condition = Condition \wedge (\#InnerClass(Class) = \#InnerClass(t_c))$

۱۲: قرار بده  $Condition = Condition \wedge (DoesInherit(Class) = DoesInherit(t_c))$

۱۳: قرار بده  $Condition = Condition \wedge (SDKInherit(Class) = SDKInherit(t_c))$

۱۴: قرار بده  $Condition = Condition \wedge (\#Interfaces(Class) = \#Interfaces(t_c))$

۱۵: قرار بده  $Condition = Condition \wedge (SDKInterface(Class) = SDKInterface(t_c))$

۱۶: قرار بده  $Condition = Condition \wedge (OuterClass(Class) = OuterClass(t_c))$

۱۷: اگر  $Condition \neq True$ :

۱۸:  $Candidate = Candidate - \{Class\}$

---

در ابتدای این ماژول، کلاس‌های کتابخانه‌ای در سطح  $i$  ام را با استفاده از متد  $FilterClassByLevel$

محاسبه می‌کنیم و آن را با  $S_{lib,i}$  نمایش می‌دهیم. در ادامه به ازای تمامی کلاس‌های موجود در مجموعه‌ی

$S_{lib,i}$ ، کلاس‌های کاندید را با استفاده از ماژول یافتن کلاس‌های کاندید محاسبه و تحت عنوان  $Candidate_{cls}$

ذخیره می‌کنیم. سپس برای محاسبه‌ی کلاس‌های منطبق بر کلاس هدف، تمامی کلاس‌های کاندید حاصل از مرحله‌ی قبل را با استفاده از توابع چکیده‌سازی محلی، با کلاس هدف مقایسه و در صورتی که میزان تشابه دو کلاس از حد آستانه‌ی  $Thr_s$  بیشتر بود آنگاه کلاس مورد نظر را به عنوان کلاس تطابق داده شده به مجموعه‌ی  $Matched_{Class}$  اضافه می‌کنیم. در ادامه سعی می‌شود تا مسئله‌ی پیدا کردن کلاس‌های

---

## الگوریتم ۲ فیلتر کتابخانه‌ها

---

ورودی: لیست کلاس‌های برنامه‌ی  $S_{app}$ ، مخزن کتابخانه‌های  $libs$

خروجی: لیست کتابخانه‌های استفاده‌شده در برنامه

۱: قرار بده  $Matched = \emptyset, Libs = \emptyset$

۲: به ازای  $lib \in libs$ :

۳: قرار بده  $GetMaxLengthClasses(lib, N) = TargetClasses$

۴: به ازای  $Class \in TargetClasses$ :

۵: قرار بده  $Candidates = FindCandidate(Class, S_{app}), Matched_{Class} = \emptyset$

۶: به ازای  $Ca \in Candidates$ :

۷: اگر  $FHashCompare(Sig_{ca}, Sig(Class)) > Thr_F$ :

۸: قرار بده  $Matched_{Class} = Matched_{Class} \cup Ca$

۹: قرار بده  $Matched = Matched \cup Matched_{class}$

۱۰: اگر  $Matched \neq \emptyset$ :

۱۱: قرار بده  $Libs = Libs \cup lib$

۱۲: برگردان  $Libs$

---

کتابخانه‌ای را به مسئله‌ی تخصیص در یک گراف دوبخشی (حاصل از کلاس‌های کتابخانه‌ای و کلاس‌های برنامه) تبدیل کرد. به همین جهت گراف دو بخشی وزن‌دار  $G$  با تابع هزینه‌ی  $C \times V \rightarrow \mathbb{N}$  را با استفاده از کلاس‌های کتابخانه‌ای و کلاس‌های برنامه تشکیل می‌دهیم. گراف  $G$  به صورت زیر تعریف می‌شود.

$$G = (U, V, E) \quad (2-4)$$

$$U = \{cls \mid cls \in S_{lib,i} \wedge M_{cls} \neq \emptyset\} \quad (3-4)$$

$$V = \bigcup_{cls \in U} Matched_{cls} \quad (4-4)$$

$$\{(cls, c) \mid cls \in s_{lib,i}, c \in Matched_{cls}\} \quad (5-4)$$

$$C(cls, c) = 100 - FhashCompare(Sig_c, Sig_{cls}) \quad (6-4)$$

$$FhashCompare : Sig_1 \times Sig_2 \rightarrow R \quad Sig_1, Sig_2 \in ClassesSigs, R \in \mathbb{N} \quad (7-4)$$

### الگوریتم ۳ نگاشت کلاس‌های کتابخانه و برنامه

ورودی: لیست کلاس‌های کتابخانه‌ای برنامه شامل  $Cls_{lib}$ ، لیست کلاس‌های برنامه شامل  $Cls_{apps}$

خروجی: لیستی از دوتایی‌های  $(LibraryClass, AppClass)$

۱: قرار بده  $f = \emptyset, i = \emptyset$

۲: تا وقتی  $Cls_{lib} \neq \emptyset$ :

۳: قرار بده  $S_{lib,i} = FilterClassByLevel(S_{lib}, i)$

۴: قرار بده  $S_{lib} = S_{lib} - S_{lib,i}$

۵: قرار بده  $U = \emptyset, V = \emptyset, E = \emptyset$

۶: به ازای  $cls \in S_{lib,i}$ :

۷: قرار بده  $Candidate_{cls} = FindCandidates(cls, Cls_{app})$

۸: قرار بده  $M_{cls} = \emptyset$

۹: به ازای  $c \in Candidate_{cls}$ :

۱۰: اگر  $FhashCompare(Sig_c, Sig_{cls}) > Thr_s$ :

۱۱: قرار بده  $Matched_{cls} = Matched_{cls} \cup c$

۱۲: قرار بده  $E = E \cup (c, cls)$

۱۳: اگر  $M_{cls} \neq \emptyset$ :

۱۴: قرار بده  $U = U \cup \{cls\}$

۱۵: قرار بده  $V = V \cup M_{cls}$

۱۶: قرار بده  $f_i = AssignmentSolver(U, V, E)$

۱۷: قرار بده  $f = f \cup f_i$

۱۸: قرار بده  $i = i + 1$

۱۹: برگردان  $f$

همانطور که در شکل — نشان داده شده است، گراف  $G$  یک گراف دو بخشی حاصل از گره‌های  $U$  و  $V$  می‌باشد که توسط مجموعه یال‌های  $E$  به یکدیگر متصل شده‌اند. برای نگاشت گره‌های مجموعه‌ی  $V$  به مجموعه‌ی  $U$  از حل مسئله‌ی تخصیص برای سطح  $i$ ، در گراف  $G$  استفاده شده‌است. خروجی ماژول نگاشت کتابخانه‌ها، تابع یک‌به‌یک و پوشای مشروح در — خواهد بود. در نهایت اجتماع توابع یک‌به‌یک و پوشای به‌دست آمده به ازای هر یک از کلاس‌های کتابخانه‌ای، خروجی ماژول نگاشت می‌باشد که همان نگاشت میان کلاس‌های کتابخانه‌ای و کلاس‌های برنامه می‌باشد.

## ۴-۴-۲ مولفه‌ی یافتن نزدیک‌ترین همسایه

جهت افزایش سرعت تشخیص برنامه‌های بازسته‌بندی شده، از یک مرحله‌ی پیش‌پردازش شامل الگوریتم  $K$  نزدیک‌ترین همسایه استفاده شده‌است. به جهت اجرای الگوریتم نزدیک‌ترین همسایه، ابتدا ۵ ویژگی، مبتنی بر منابع برنامه با استفاده از ماژول استخراج ویژگی، استخراج می‌شود و درخت ۵ بعدی حاصل از اجرای الگوریتم ۵ ساخته می‌شود. در نهایت با استفاده از اجرای الگوریتم یافتن نزدیک‌ترین همسایه، برنامه‌های کاندید جهت مقایسه‌ی دودویی را انتخاب می‌کنیم.

### انتخاب ویژگی

- **تعداد فعالیت‌ها:** هر فعالیت یک صفحه‌ی جداگانه را در برنامه‌های اندرویدی به منظور پیاده‌سازی واسط کاربری در اختیار توسعه‌دهندگان قرار می‌دهد. از آنجایی که پیاده‌سازی فعالیت‌های جدید، نیازمند افزودن آن‌ها در حلقه‌ی فراخوانی فعالیت‌هاست، بنابراین، برنامه‌های بازسته‌بندی شده عموماً دارای تعداد فعالیت مشخص و مشابه هستند چرا که بخشی از اهداف آن‌ها در صورتی محقق می‌شود که کاربران به جهت شباهت‌های واسط کاربری دچار خطا شوند. برای استخراج فعالیت‌های هر برنامه، از فایل *AndroidManifest.xml* استفاده شده‌است. همانطور که در الگوریتم ۴ مشاهده می‌شود، فعالیت‌های اندرویدی در فایل *AndroidManifest.xml* با تگ مشخص *Activity* مشخص شده‌اند. نمونه‌ای از تعریف یک فعالیت در تصویر ۴-۵ مشاهده می‌شود.

```

<manifest ... >
  <application ... >
    <activity android:name=".ExampleActivity" />
    ...
  </application ... >
  ...
</manifest >

```

شکل ۴-۵: نمونه‌ای از تعریف یک فعالیت

- **تعداد مجوزهای دسترسی برنامه:** ویژگی دیگری که برای طبقه‌بندی از آن استفاده شده است، تعداد مجوزهای دسترسی هر برنامه است که در طی اجرای برنامه از کاربر درخواست می‌شود. از آنجایی که اجرای بدافزارهای بازبسته‌بندی شده، نیازمند دسترسی‌های گوناگون است بنابراین بدافزارنویسان برای اهداف خود از مجوزهای دسترسی به کرات استفاده می‌کنند. همانند تعداد فعالیت‌های برنامه، برای شمارش تعداد مجوزهای دسترسی، نیازمند شمارش تگ مخصوص *uses – permission* در فایل فراداده‌ی *AndroidManifest.xml* هستیم. نمونه‌ای از تعریف حق دسترسی جهت استفاده از دوربین کاربر را در تصویر ۴-۶ مشاهده می‌کنید.

```

<manifest ...>
  <uses-permission android:name="android.permission.CAMERA"/>
  <application ...>
    ...
  </application>
</manifest>

```

شکل ۴-۶: نمونه‌ای از تعریف یک حق دسترسی

- **تعداد فیلترهای تصمیم:** جابه‌جایی میان صفحات برنامه و یا جابه‌جایی میان صفحات برنامه‌های اندرویدی، با استفاده از فیلترهای تصمیم صورت می‌گیرد. با استفاده از فیلترهای تصمیم برنامه مولفه‌ی درخواستی خود را به سیستم عامل اعلام می‌کند. بنابراین رفتار برنامه‌های اندرویدی، از جهتی مبتنی بر فیلترهای تصمیم هستند چرا که عمده‌ی فعالیت‌های یک برنامه مبتنی بر جابه‌جایی میان صفحات مختلف (فعالیت‌ها) با استفاده از فیلترهای تصمیم می‌باشد. در این قسمت تمامی فیلترهای تصمیم ضمنی و یا صریح، با استفاده از فایل *AndroidManifest.xml* و تگ مخصوص *intent – filter* استخراج می‌شود.

- **میانگین تعداد فایل های png:** برای محاسبه ی میانگین تعداد فایل های تصویری که پسوند *png* دارند، از شمارش تعداد تعداد پوشه های *drawable* که حاوی تصاویر برنامه هستند استفاده شده است. منظور از میانگین در این ویژگی، محاسبه ی تقسیم تعداد تمامی فایل های *png* بر تعداد پوشه های *drawale* می باشد.

- **میانگین تعداد فایل های xml:** همانطور که در الگوریتم ۴ مشاهده می شود، برای محاسبه ی میانگین تعداد فایل های با پسوند *xml* از محاسبه ی تعداد پوشه های موجود در پوشه ی منابع برنامه یعنی */res* استفاده شده است. از آنجایی که المان های موجود در واسط کاربری برنامه توسط فایل های *xml* پیاده سازی می شود، بنابراین حذف و یا اضافه کردن آن ها نیازمند صرف هزینه می باشد. دلیل اسفاده از میانگین در دو ویژگی اخیر، ناشی از عدم تغییر این عدد به ازای اضافه کردن پوشه های جدید است. برای مثال در صورتی که یک برنامه با استفاده از یک زبان دوم بازبسته بندی شود، یک پوشه شامل منابع آن ایجاد می شود. حال در این صورت تعداد کل منابع *xml* اضافه شده است اما میانگین آن ها تغییر نمی کند.

پس از جمع آوری ویژگی های ذکر شده، با فرض آن که  $\vartheta = \{\alpha^i\}_{i=1,2,3,...,N}$  مجموعه ی برنامه های موجود در مخزن باشد، بردار ویژگی هر برنامه را با  $Features_i = (F_1^i, ..., F_5^i)$  که به ترتیب شامل تعداد فعالیت ها، تعداد دسترسی های درخواستی، تعداد فیلترهای تصمیم، میانگین تعداد فایل های *xml* و میانگین تعداد فایل های *png* خواهد بود. پس از ایجاد بردار ویژگی  $Features_i$ ، به منظور همسانی داده های آزمون، نرمال سازی با استفاده از تابع *Normalized* روی داده ها صورت می گیرد. معادله ی تابع *Normalized* به ازای هر ویژگی  $f_j^i$  زیر مشخص می شود.

$$Normalized(f_j^i) = \sqrt{\frac{f_j^i - \min(f_j^{1,...,N})}{\max(f_j^{1,...,N}) - \min(f_j^{1,...,N})}} \quad (۸-۴)$$

ماژول استخراج ویژگی های مبتنی بر منابع به عنوان ورودی برنامه های موجود در مخزن و فایل فراداده ی *AndroidManifest.xml* و به عنوان خروجی لیستی از ویژگی های ۵ تایی ذکر شده برای هر برنامه باز می گرداند.

## ساخت درخت KD

در ادامه برای طبقه بندی داده های هر برنامه از یک طبقه بند مبتنی بر نزدیک ترین همسایه استفاده شده است. به صورت کلی دو روش پیاده سازی برای جست و جوی نزدیک ترین همسایه وجود دارد. در بدیهی ترین

حالت ابتدا فاصله‌ی یک نقطه از تمامی نقاط دیگر را محاسبه، در ادامه با استفاده از مرتب‌سازی مبتنی بر فاصله، تعدادی از نزدیک‌ترین همسایه‌ها را استخراج می‌کنیم. همانطور که مشخص است روش مورد نظر نیازمند محاسبه‌ی فاصله‌ی نقطه‌ی ورودی با تمامی نقاط دیگر است. در روش دیگری که به  $KD - KNN$  شهرت دارد، ابتدا تمامی داده‌های برنامه را در یک داده‌ساختار درختی ذخیره و برای جست‌وجوی  $K$  تا از نزدیک‌ترین همسایه‌های هر گره، از آن استفاده می‌کنیم. در این ساختار درختی برای جست‌وجوی گره‌های نزدیک، تمامی داده‌های آزمون به بلاک‌های نزدیک به هم تقسیم و پردازش گره‌های نزدیک تنها در این بلاک صورت می‌گیرد. در ادامه برای درک بهتر الگوریتم جست‌وجوی درختی  $KNN$  مثالی از یک مجموعه داده‌ی آزمون ۲ بعدی را در الگوریتم ۵ بررسی خواهیم کرد.

پس از اجرای الگوریتم ابتدا نقاط ورودی بر اساس عنصر اول دوتایی‌های ورودی جداسازی و مرتب می‌شوند. در ادامه میانه‌ی داده‌های مرتب‌شده‌ی حاصل از مرحله‌ی قبل، محاسبه و عناصر داده‌ای بیش از مقدار میانه در سمت راست درخت و عناصر کم‌تر در سمت چپ درخت تقسیم می‌شوند. این مرحله، در واقع سازوکار ایجاد بلاک‌های داده‌ای انجام می‌شود. از آنجایی که فاصله‌ی دو نقطه در فضا از یکدیگر با استفاده از معادله‌ی ۴-۹ محاسبه می‌شود، بنابراین جداسازی بلاک‌های داده‌ای در الگوریتم ۵ به صورت یکی در میان، بر اساس عناصر اول و دوم از لیست داده‌ای تقسیم می‌شوند.

$$d(x, y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (9-4)$$

برای ساخت درخت مبتنی بر ویژگی‌های استخراج‌شده از الگوریتم ۴، از بردار ویژگی حاصل استفاده کرده و در نهایت درخت ۵ بعدی حاوی اطلاعات برنامه‌های اندرویدی ساخته می‌شود. هر گره درخت مذکور حاوی یک بردار ویژگی ۵ تایی است و مطابق الگوریتم ۵ تصمیم‌گیری در مورد تقسیم داده‌های موجود در هر سطح با توجه به یکی از عناصر ۵ تایی صورت می‌گیرد. به عنوان مثال، زمانی که در سطح اول و ریشه‌ی درخت هستیم، عناصر اول ۵ تایی‌های مجموعه داده را جدا و آن‌ها را مرتب می‌کنیم. سپس با توجه به محاسبه‌ی میانه مقدار  $Median$  را برای این گروه داده‌ای محاسبه کرده و ۵ تایی مربوط به عنصر  $Median$  در ریشه قرار می‌گیرد. سپس باقی مجموعه‌ی داده‌ای با توجه به بزرگ‌تر بودن و یا کوچک‌تر بودن از عنصر  $axis$  که همان شاخص تصمیم است، تقسیم می‌شوند و این رویه به صورت بازگشتی برای تمامی بلاک‌های داده‌ای و گره‌های درخت تکرار می‌شود تا زمانی که درخت ساخته شود.

ماژول ساخت درخت  $KD$  در این پژوهش لیستی از عناصر ۵ تایی حاوی ویژگی‌های مبتنی بر منابع و به طول  $n$  می‌باشد. خروجی ماژول یک درخت ۵ بعدی حاوی ویژگی‌های مذکور است که می‌توان الگوریتم جست‌وجوی نزدیک‌ترین همسایه‌ی ۶ را از مرتبه‌ی  $\log(n)$  اجرا کرد.

## جست‌وجوی $k$ همسایه‌ی نزدیک

پس از ساخت داده‌ساختار درختی ۵ بعدی در مرحله‌ی قبل، حال هدف یافتن نزدیک‌ترین همسایه‌های هر برنامه‌ک به جهت بررسی دوتایی آن‌ها می‌باشد. به همین جهت در این قسمت از الگوریتم جست‌وجوی نزدیک‌ترین همسایه که در پژوهش [۹۵] معرفی شده‌است، استفاده کرده‌ایم. شبه‌کد این روش را می‌توان در الگوریتم ۶ مشاهده نمود. ورودی الگوریتم شامل ریشه‌ی درخت  $KD$  و هدف آن یافتن نزدیک‌ترین نقطه از میان نقاط درون درخت به نقطه‌ی ورودی  $Tpoint$  می‌باشد. به طور کلی الگوریتم مذکور را می‌توان به دو قسمت تقسیم کرد، در قسمت اول در طی یک روند رو به جلو سعی می‌کنیم با توجه به نقطه‌ی ورودی، مکان درست آن را در برگ درخت بیابیم. به همین منظور از یک ساختار مبتنی بر درخت تصمیم استفاده می‌کنیم. در این ساختار، در صورتی که مقدار شاخص تصمیم  $axis$  بردار ورودی، از همین مقدار در بردار گره بیشتر باشد، آن‌گاه جست‌وجو را در زیردرخت سمت راست ریشه انجام می‌دهیم و در صورتی که مقدار آن کوچکتر باشد آن‌گاه زیر درخت سمت چپ را بررسی می‌کنیم. این روند بازگشتی را تا زمانی که به گره برگ برسیم ادامه می‌دهیم و مسیر طی شده را ذخیره می‌کنیم. در مرحله‌ی بعدی با توجه به نقاط درون مسیر  $Path$  تمامی گره‌های مسیر را به صورت عقب‌گرد بررسی و نزدیک‌ترین بردار ویژگی به گره ورودی را پیدا کرده و آن را بر بازمی‌گردانیم.

متد  $Distance$  در الگوریتم ۶ فاصله‌ی میان دو بردار ۵ تایی را با استفاده از فاصله‌ی اقلیدوسی به صورت زیر محاسبه می‌کند.

$$d(f^1, f^2) = \sqrt{(f_1^1 - f_1^2)^2 + (f_2^1 - f_2^2)^2 + \dots + (f_5^1 - f_5^2)^2} \quad (10-4)$$

پس از ایجاد درخت و پیاده‌سازی الگوریتم نزدیک‌ترین همسایه، تعداد  $K$  فراخوانی از الگوریتم ۶ منجر به دریافت لیستی از  $K$  نزدیک‌ترین همسایه‌های بردار ورودی می‌شود.

## ۴-۳-۴ مقایسه‌ی دودویی و تشخیص برنامه‌های بازبسته‌بندی شده

هدف از پیاده‌سازی مرحله‌ی طبقه‌بندی، کاهش فضای مقایسه‌ی دودویی در این مرحله بوده‌است که منجر به افزایش سرعت مقایسه می‌شود. پس از اجرای الگوریتم یافتن  $K$  تا از نزدیک‌ترین همسایه‌های برنامه‌ک ورودی با استفاده از ویژگی‌های مبتنی بر منبع، در این قسمت مقایسه‌ی دودویی میان  $K$  برنامه‌ک و برنامه‌ک ورودی صورت می‌گیرد.

پس از شناسایی کلاس‌های کتابخانه‌ای در الگوریتم ۳، نگاشتی از کلاس‌های هر برنامه‌ک داخل مخزن و برنامه‌ک هدف با کلاس‌های کتابخانه‌ای به صورت  $F : LibraryClasses \rightarrow AppClasses$  به دست می‌آید. در ادامه امضای هر کدام از کلاس‌های برنامه‌ک را مطابق با شکل ۴-۳ تشکیل می‌دهیم. امضای هر



برنامک از الحاق تمامی امضای کلاس‌های آن ایجاد می‌شود. سپس شروع به حذف کلاس‌های کتابخانه‌ای با توجه به نگاشت  $F$  از هر برنامک می‌کنیم و مجموعه‌ی کلاس‌های برنامک با حذف کتابخانه‌های اندرویدی تشکیل مجموعه‌ی  $S'_{app}$  را می‌دهند که به صورت زیر محاسبه شده‌است.  $AppClasses$  مجموعه‌ی تمامی کلاس‌های کتابخانه‌ای است که در نگاشت  $F$  به دست آمده‌است.

$$Cls'_{app} = Cls_{app} - Appclasses \quad (۱۱-۴)$$

---

#### الگوریتم ۴ استخراج ویژگی‌های مبتنی بر منابع

---

ورودی: برنامک‌های موجود در مخزن  $DsApps$  و فایل فراداده‌ی هرکدام  $Manifest_{app}$

خروجی: ویژگی‌های ۵ تایی از هر برنامک شامل مجموعه‌ی  $Features$

۱: قرار بده  $Features = \emptyset$

۲: به ازای  $app \in DsApps$ :

۳: به ازای  $tag \in Manifest_{app}$ :

۴: اگر  $tag = Activity$ :

۵: قرار بده  $Features_{app,AcNumber} = Features_{app,AcNumber} + ۱$

۶: اگر  $tag = uses - premission$ :

۷: قرار بده  $Features_{app,PNumber} = Features_{app,PNumber} + ۱$

۸: اگر  $tag = intent - filter$ :

۹: قرار بده  $Features_{app,IfNumber} = Features_{app,IfNumber} + ۱$

۱۰: قرار بده  $\#TotalDrawableDir = GetTotalDrNum()$

۱۱: قرار بده  $\#TotalPngFiles = GetTotalPngs("/res")$

۱۲: قرار بده  $Features_{app,AvgPngPerDir} = \frac{\#TotalPngFiles}{\#GetTotalDrNum}$

۱۳: قرار بده  $\#TotalResDir = GetTotalDirNum()$

۱۴: قرار بده  $\#TotalXmlFiles = GetTotalXmles("/res")$

۱۵: قرار بده  $Features_{app,AvgXmlPerDir} = \frac{\#TotalXmlFiles}{\#GetTotalDirNum}$

۱۶: برگردان  $Features$

---

پس از حذف کلاس‌های کتابخانه‌ای از هر امضا، در این مرحله امضای اصلی برنامک حاوی کلاس‌هایی که توسط توسعه‌دهنده پیاده‌سازی شده‌اند را با یکدیگر مقایسه می‌کنیم. مقایسه‌ی امضای برنامک‌ها با استفاده از روش‌های چکیده‌سازی محلی صورت می‌گیرد. در روش‌های چکیده‌سازی معمولی نظیر MD۵

در صورتی که قسمت کوچکی از محتوای متنی موارد مقایسه تغییر کند آنگاه چکیده‌ی خروجی حاصل از این‌گونه روش‌ها نیز کاملاً تغییر می‌کنند. از آنجایی که در علم داده‌کاوی، مقایسه‌ی محتوای دو متن با استفاده از روش‌های چکیده‌سازی معمولی امکان‌پذیر نیست، روش‌هایی مبتنی بر چکیده‌سازی فازی ارائه شده‌است که در آن‌ها بر خلاف روش‌های معمولی، تغییرات محتوای موارد مقایسه، منجر به خروجی‌های کاملاً متفاوت نمی‌شود بلکه در صورتی که دو محتوا مشابه اما با تغییرات اندک باشند، آنگاه دو چکیده به یک سبد یکسان نگاشت خواهند شد.

---

#### الگوریتم ۵ ساخت درخت دوبعدی جست‌وجوی $K$ نزدیک‌ترین همسایه

---

ورودی: لیستی از داده‌های دوتایی  $Points$  به صورت  $(x, y)$  و به طول  $n$

ورودی: عمق سطح  $Depth$

خروجی: درخت دوبعدی شامل داده‌های ورودی

۱: قرار بده  $Dimension = Depth \bmod 2$

۲: اگر  $Dimension = 1$  :

۳: قرار بده  $Median = GetMedian(Sort(x_1, \dots, x_n))$

۴: در غیر این صورت:

۵: قرار بده  $Median = GetMedian(Sort(y_1, \dots, y_n))$

۶: قرار بده  $Node.data = Median$

۷: قرار بده  $Node.axis = Dimension$

۸: قرار بده  $Node.leftChild = kdtree(points \text{ in } points \text{ before } median, depth + 1)$

۹: قرار بده  $Node.rightChild = kdtree(points \text{ in } points \text{ after } median, depth + 1)$

---

---

## الگوریتم ۶ جست و جوی $K$ نزدیک ترین همسایه

---

ورودی: ریشه ی درخت  $Root$

ورودی: نقطه ی آزمون  $TPoint$

خروجی: نزدیک ترین نقطه به ورودی  $NearestPoint$

۱:  $Path = \emptyset$  قرار بده

۲: تا وقتی  $Root \neq leaf$ :

۳:  $Path.add(Root)$

۴: اگر  $Root[data][Root.axis] > TPoint[axis]$

۵:  $Searchtree(rightChild)$

۶: در غیر این صورت:

۷:  $Searchtree(leftChild)$

۸: به ازای  $Point \in Path$ :

۹: قرار بده  $NearestDist = Distance(Point, TPoint)$

۱۰: اگر  $|Point[Data][axis] - TPoint[axis]| > |TPoint[axis] - Root[Data][axis]|$ :

۱۱:  $Travel(Root.nextchild)$

۱۲: قرار بده  $Dist = Distance(ChildPoint, TPoint)$

۱۳: اگر  $Dist > NearestDist$ :

۱۴: قرار بده  $NearestDist = Dist$

۱۵: قرار بده  $NPoint = Point$

۱۶: قرار بده  $NearestPoint = NPoint$

۱۷: برگردان  $NearestPoint$

---

## فصل ۵

### نتیجه‌گیری

در این فصل، ضمن جمع‌بندی نتایج جدید ارائه‌شده در پایان‌نامه یا رساله، مسائل باز باقی‌مانده و همچنین پیشنهادهایی برای ادامه‌ی کار ارائه می‌شوند.

## فصل ۶

### نتیجه‌گیری

- [1] Y. Shao, X. Luo, C. Qian, P. Zhu, and L. Zhang. Towards a scalable resource-driven approach for detecting repackaged android applications. In *Proceedings of the 30th Annual Computer Security Applications Conference, ACSAC '14*, page 56–65, New York, NY, USA, 2014. Association for Computing Machinery.
- [2] M. S. Bhatt, H. Patel, and S. Kariya. A Survey Permission Based Mobile Malware Detection. *Int.J.Computer Technology and Applications*, 6(5):852–856, 2015.
- [3] N. T. Cam, N. H. Khoa, T. T. An, N. P. Bach, and V.-H. Pham. Detect repackaged android applications by using representative graphs. In *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 102–106, 2021.
- [4] Global mobile OS market share 2022 | Statista — statista.com. <https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/#:~:text=Android%20maintained%20its%20position%20as,the%20mobile%20operating%20system%20market>. [Accessed 02-Feb-2023].
- [5] Play Protect | Google Developers — developers.google.com. <https://developers.google.com/android/play-protect>. [Accessed 02-Feb-2023].
- [6] Decompile and modify an Android application | cylab.be — cylab.be. <https://cylab.be/blog/69/decompile-and-modify-an-android-application>. [Accessed 02-Feb-2023].
- [7] A. Dizdar. OWASP Mobile Top 10 Vulnerabilities and How to Prevent Them — brightsec.com. <https://brightsec.com/blog/owasp-mobile-top-10/>. [Accessed 02-Feb-2023].
- [8] D. J. Wu, C. H. Mao, T. E. Wei, H. M. Lee, and K. P. Wu. DroidMat: Android malware detection through manifest and API calls tracing. *Proceedings of the 2012*

*7th Asia Joint Conference on Information Security, AsiaJCIS 2012*, pages 62–69, 2012.

- [9] K. Khanmohammadi, N. Ebrahimi, A. Hamou-Lhadj, and R. Khoury. Empirical study of android repackaged applications. *Empirical Software Engineering*, 24(6):3587–3629, 2019.
- [10] T. Vidas and N. Christin. Sweetening android lemon markets: Measuring and combating malware in application marketplaces. *CODASPY 2013 - Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy*, 2011:197–207, 2013.
- [11] P. Maniriho, A. N. Mahmood, and M. J. M. Chowdhury. A study on malicious software behaviour analysis and detection techniques: Taxonomy, current trends and challenges. *Future Generation Computer Systems*, 130:1–18, 2022.
- [12] Z. Ma, H. Wang, Y. Guo, and X. Chen. Libradar: Fast and accurate detection of third-party libraries in android apps. In *2016 IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C)*, pages 653–656, 2016.
- [13] S. Dong, M. Li, W. Diao, X. Liu, J. Liu, Z. Li, F. Xu, K. Chen, X. F. Wang, and K. Zhang. *Understanding android obfuscation techniques: A large-scale investigation in the wild*, volume 254. Springer International Publishing, 2018.
- [14] V. Rastogi, Y. Chen, and X. Jiang. DroidChameleon: Evaluating Android anti-malware against transformation attacks. *ASIA CCS 2013 - Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, pages 329–334, 2013.
- [15] Trail: The reflection api the javax; tutorials — docs.oracle.com. <https://docs.oracle.com/javase/tutorial/reflect/index.html>. [Accessed 02-Feb-2023].
- [16] X. Zhang, F. Breiting, E. Luechinger, and S. O’Shaughnessy. Android application forensics: A survey of obfuscation, obfuscation detection and deobfuscation techniques and their impact on investigations. *Forensic Science International: Digital Investigation*, 39:301285, 2021.
- [17] ProGuard Manual: Home | Guardsquare — guardsquare.com. <https://www.guardsquare.com/manual/home>. [Accessed 02-Feb-2023].
- [18] Allatori Java Obfuscator — codedemons.net. <http://www.codedemons.net/allatori.html>. [Accessed 02-Feb-2023].

- [19] Y. Wang. Obfuscation-Resilient Code Detection Analyses for Android Apps. 2018.
- [20] L. Ardito, R. Coppola, S. Leonardi, M. Morisio, and U. Buy. Automated Test Selection for Android Apps Based on APK and Activity Classification. *IEEE Access*, 8:187648–187670, 2020.
- [21] L. Li, T. F. Bissyandé, J. Klein, and Y. Le Traon. An investigation into the use of common libraries in android apps. 1:403–414, 2016.
- [22] N. Karankar, P. Shukla, and N. Agrawal. Comparative study of various machine learning classifiers on medical data. pages 267–271, 2017.
- [23] Y. Shao, X. Luo, C. Qian, P. Zhu, and L. Zhang. Towards a scalable resource-driven approach for detecting repackaged android applications. *ACM International Conference Proceeding Series*, 2014-Decem(December):56–65, 2014.
- [24] F. Zhang, H. Huang, S. Zhu, D. Wu, and P. Liu. Viewdroid: Towards obfuscation-resilient mobile application repackaging detection. page 25–36, 2014.
- [25] J. Crussell, C. Gibler, and H. Chen. Attack of the clones: Detecting cloned applications on Android markets. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7459 LNCS:37–54, 2012.
- [26] H. Gonzalez, N. Stakhanova, and A. A. Ghorbani. Droidkin: Lightweight detection of android apps similarity. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, 152(January 2015):436–453, 2015.
- [27] X. Chen, C. Li, D. Wang, S. Wen, J. Zhang, S. Nepal, Y. Xiang, and K. Ren. Android HIV: A Study of Repackaging Malware for Evading Machine-Learning Detection. *IEEE Transactions on Information Forensics and Security*, 15(8):987–1001, 2020.
- [28] A. Salem. Stimulation and Detection of Android Repackaged Malware with Active Learning. 2015.
- [29] T. Nguyen, J. T. McDonald, W. B. Glisson, and T. R. Aniel. Detecting repackaged android applications using perceptual hashing. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2020-January:6641–6650, 2020.
- [30] F. Alswaina and K. Elleithy. Android malware family classification and analysis: Current status and future directions. *Electronics (Switzerland)*, 9(6):1–20, 2020.



- [31] F. Akbar, M. Hussain, R. Mumtaz, Q. Riaz, A. W. A. Wahab, and K.-H. Jung. Permissions-based detection of android malware using machine learning. *Symmetry*, 14(4), 2022.
- [32] X. Chen, C. Li, D. Wang, S. Wen, J. Zhang, S. Nepal, Y. Xiang, and K. Ren. Android HIV: A study of repackaging malware for evading machine-learning detection. *IEEE Transactions on Information Forensics and Security*, 15:987–1001, 2020.
- [33] Q. Zhang, X. Zhang, Z. Yang, and Z. Qin. An efficient method of detecting repackaged android applications. pages 056 (4 .)–056 (4 .), 01 2014.
- [34] W. Zhou, Y. Zhou, X. Jiang, and P. Ning. Detecting repackaged smartphone applications in third-party android marketplaces. page 317–326, 2012.
- [35] A. Desnos. Android: Static analysis using similarity distance. pages 5394–5403, 2012.
- [36] B. B. Rad and M. Masrom. Metamorphic Virus Detection in Portable Executables Using Opcodes Statistical Feature. *International Journal on Advanced Science, Engineering and Information Technology*, 1(4):403, 2011.
- [37] B. B. Rad, M. Masrom, and S. Ibrahim. Opcodes histogram for classifying metamorphic portable executables malware. pages 209–213, 2012.
- [38] Q. Jerome, K. Allix, R. State, and T. Engel. Using opcode-sequences to detect malicious android applications. In *2014 IEEE International Conference on Communications (ICC)*, pages 914–919, 2014.
- [39] C. Liangboonprakong and O. Sornil. Classification of malware families based on n-grams sequential pattern features. pages 777–782, 2013.
- [40] Y. D. Lin, Y. C. Lai, C. H. Chen, and H. C. Tsai. Identifying android malicious repackaged applications by thread-grained system call sequences. *Computers and Security*, 39(PART B):340–350, 2013.
- [41] P. Faruki, V. Laxmi, V. Ganmoor, M. Gaur, and A. Bharmal. Droidolytics: Robust feature signature for repackaged android apps on official and third party android markets. In *2013 2nd International Conference on Advanced Computing, Networking and Security*, pages 247–252, 2013.

- [42] J. Ko, H. Shim, D. Kim, Y.-S. Jeong, S.-j. Cho, M. Park, S. Han, and S. B. Kim. Measuring similarity of android applications via reversing and k-gram birthmarking. page 336–341, 2013.
- [43] S. Kishore, R. Kumar, and S. Rajan. *Towards Accuracy in Similarity Analysis of Android Applications*, volume 11281 LNCS. Springer International Publishing, 2018.
- [44] R. Potharaju, A. Newell, C. Nita-Rotaru, and X. Zhang. Plagiarizing smartphone applications: Attack strategies and defense techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7159 LNCS:106–120, 2012.
- [45] J. Crussell, C. Gibler, and H. Chen. Scalable Semantics-Based Detection of Similar Android Applications. *Esortics*, pages 182–199, 2013.
- [46] W. Hu, J. Tao, X. Ma, W. Zhou, S. Zhao, and T. Han. Migdroid: Detecting app-repackaging android malware via method invocation graph. In *2014 23rd International Conference on Computer Communication and Networks (ICCCN)*, pages 1–7, 2014.
- [47] W. Zhou, Y. Zhou, M. Grace, X. Jiang, and S. Zou. Fast, scalable detection of ”piggybacked” mobile applications. *CODASPY 2013 - Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy*, pages 185–195, 2013.
- [48] K. Chen, P. Liu, and Y. Zhang. Achieving accuracy and scalability simultaneously in detecting application clones on Android markets. *Proceedings - International Conference on Software Engineering*, (1):175–186, 2014.
- [49] L. Adhianto, S. Banerjee, M. Fagan, M. Krentel, G. Marin, J. Mellor-Crummey, and N. R. Tallent. HPCTOOLKIT: Tools for performance analysis of optimized parallel programs. *Concurrency and Computation: Practice and Experience*, 22(6):685–701, 2010.
- [50] J. Zheng, K. Gong, S. Wang, Y. Wang, and M. Lei. Repackaged apps detection based on similarity evaluation. pages 1–5, 2016.
- [51] M. Torki. Detecting repackaged android applications. 2018.
- [52] Y. Wang. Obfuscation-resilient code detection analyses for android apps,” phd thesis, the ohio state university. 2018.

- [53] X. Wu, D. Zhang, X. Su, and W. Li. Detect repackaged android application based on http traffic similarity. *Sec. and Commun. Netw.*, 8(13):2257–2266, sep 2015.
- [54] M. Alshehri. APP-NTS: a network traffic similarity-based framework for repacked Android apps detection. *Journal of Ambient Intelligence and Humanized Computing*, 13(3):1537–1546, 2022.
- [55] G. He, B. Xu, L. Zhang, and H. Zhu. On-device detection of repackaged android malware via traffic clustering. *Security and Communication Networks*, 2020:8630748, May 2020.
- [56] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.
- [57] J. Malik and R. Kaushal. Credroid: Android malware detection by network traffic analysis. In *Proceedings of the 1st ACM Workshop on Privacy-Aware Mobile Computing*, PAMCO ’16, page 28–36, New York, NY, USA, 2016. Association for Computing Machinery.
- [58] D. Iland and A. Pucher. Detecting android malware on network level. 2011.
- [59] S. Kandukuru and R. M. Sharma. Android malicious application detection using permission vector and network traffic analysis. pages 1126–1132, 2017.
- [60] M. Lin, D. Zhang, X. Su, and T. Yu. Effective and scalable repackaged application detection based on user interface. In *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, 2017.
- [61] S. Yue, Q. Sun, J. Ma, X. Tao, C. Xu, and J. Lu. Regiondroid: A tool for detecting android application repackaging based on runtime ui region features. In *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 323–333, 2018.
- [62] Y. Hu, G. Xu, B. Zhang, K. Lai, G. Xu, and M. Zhang. Robust app clone detection based on similarity of ui structure. *IEEE Access*, 8:77142–77155, 2020.
- [63] S. Li. Juxtapp and dstruct: Detection of similarity among android applications. 2012.
- [64] J. Guo, D. Liu, R. Zhao, and Z. Li. Wltdroid: Repackaging detection approach for android applications. In G. Wang, X. Lin, J. Hendler, W. Song, Z. Xu, and G. Liu, editors, *Web Information Systems and Applications*, pages 579–591, Cham, 2020. Springer International Publishing.

- [65] F. Lyu, Y. Lin, J. Yang, and J. Zhou. Suidroid: An efficient hardening-resilient approach to android app clone detection. In *2016 IEEE Trustcom/BigDataSE/ISPA*, pages 511–518, 2016.
- [66] X. Liu, Z. Yu, Z. Song, L. Chen, and Z. Qin. Mdsdroid: A multi-level detection system for android repackaged applications. In *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, pages 1128–1133, 2021.
- [67] S. Yue, W. Feng, J. Ma, Y. Jiang, X. Tao, C. Xu, and J. Lu. Repdroid: An automated tool for android application repackaging detection. In *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC)*, pages 132–142, 2017.
- [68] J. Ma, Q.-W. Sun, C. Xu, and X.-P. Tao. Griddroid—an effective and efficient approach for android repackaging detection based on runtime graphical user interface. *Journal of Computer Science and Technology*, 37(1):147–181, Feb 2022.
- [69] V. Costamagna, C. Zheng, and H. Huang. Identifying and evading android sandbox through usage-profile based fingerprints. page 17–23, 2018.
- [70] Y. Zhang, K. Huang, Y. Liu, K. Chen, L. Huang, and Y. Lian. Timing-based clone detection on android markets. 12 2015.
- [71] T. Nguyen, J. T. McDonald, W. B. Glisson, and T. R. Aniel. Detecting repackaged android applications using perceptual hashing. 2020.
- [72] C. Soh, H. B. Kuan Tan, Y. L. Arnatovich, and L. Wang. Detecting clones in android applications through analyzing user interfaces. In *2015 IEEE 23rd International Conference on Program Comprehension*, pages 163–173, 2015.
- [73] T. Bläsing, L. Batyuk, A.-D. Schmidt, S. A. Camtepe, and S. Albayrak. An android application sandbox system for suspicious software detection. In *2010 5th International Conference on Malicious and Unwanted Software*, pages 55–62, 2010.
- [74] D. Kim, A. Gokhale, V. Ganapathy, and A. Srivastava. Detecting plagiarized mobile apps using API birthmarks. *Automated Software Engineering*, 23(4):591–618, 2016.
- [75] Q. Guan, H. Huang, W. Luo, and S. Zhu. Semantics-based repackaging detection for mobile apps. In J. Caballero, E. Bodden, and E. Athanasopoulos, editors, *Engineering Secure Software and Systems*, pages 89–105, Cham, 2016. Springer International Publishing.

- [76] H. Wang, Y. Guo, Z. Ma, and X. Chen. Wukong: A scalable and accurate two-phase approach to android app clone detection. In *Proceedings of the 2015 International Symposium on Software Testing and Analysis, ISSTA 2015*, page 71–82, New York, NY, USA, 2015. Association for Computing Machinery.
- [77] O. S. J. Nisha and S. M. S. Bhanu. Detection of repackaged android applications based on apps permissions. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*, pages 1–8, 2018.
- [78] W. Zhou, X. Zhang, and X. Jiang. Appink: Watermarking android apps for repackaging deterrence. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security, ASIA CCS '13*, page 1–12, New York, NY, USA, 2013. Association for Computing Machinery.
- [79] Y. Zhang and K. Chen. Appmark: A picture-based watermark for android apps. In *2014 Eighth International Conference on Software Security and Reliability (SERE)*, pages 58–67, 2014.
- [80] C. Ren, K. Chen, and P. Liu. Droidmarking: Resilient software watermarking for impeding android application repackaging. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering, ASE '14*, page 635–646, New York, NY, USA, 2014. Association for Computing Machinery.
- [81] X. Sun, J. Han, H. Dai, and Q. Li. An active android application repacking detection approach. In *2018 10th International Conference on Communication Software and Networks (ICCSN)*, pages 493–496, 2018.
- [82] L. Luo, Y. Fu, D. Wu, S. Zhu, and P. Liu. Repackage-proofing android apps. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 550–561, 2016.
- [83] Q. Zeng, L. Luo, Z. Qian, X. Du, and Z. Li. Resilient decentralized android application repackaging detection using logic bombs. In *Proceedings of the 2018 International Symposium on Code Generation and Optimization, CGO 2018*, page 50–61, New York, NY, USA, 2018. Association for Computing Machinery.
- [84] S. Tanner, I. Vogels, and R. Wattenhofer. Protecting android apps from repackaging using native code. In A. Benzekri, M. Barbeau, G. Gong, R. Laborde, and J. Garcia-Alfaro, editors, *Foundations and Practice of Security*, pages 189–204, Cham, 2020. Springer International Publishing.

- [85] D. Wermke, N. Huaman, Y. Acar, B. Reaves, P. Traynor, and S. Fahl. A large scale investigation of obfuscation use in google play. In *Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC '18*, page 222–235, New York, NY, USA, 2018. Association for Computing Machinery.
- [86] Virus Bulletin :: Obfuscation in Android malware, and how to fight back — virusbulletin.com. <https://www.virusbulletin.com/virusbulletin/2014/07/obfuscation-android-malware-and-how-fight-back>. [Accessed 12-Feb-2023].
- [87] Y. Wang and A. Rountev. Who changed you? obfuscator identification for android. In *2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*, pages 154–164, 2017.
- [88] S. Alam, Z. Qu, R. Riley, Y. Chen, and V. Rastogi. Droidnative: Automating and optimizing detection of android native code malware variants. *Computers and Security*, 65:230–246, 2017.
- [89] A. Bacci, A. Bartoli, F. Martinelli, E. Medvet, and F. Mercaldo. Detection of obfuscation techniques in android applications. In *Proceedings of the 13th International Conference on Availability, Reliability and Security, ARES 2018*, New York, NY, USA, 2018. Association for Computing Machinery.
- [90] M. D. Preda and F. Maggi. Testing android malware detectors against code obfuscation: a systematization of knowledge and unified methodology. *Journal of Computer Virology and Hacking Techniques*, 13:209–232, 2017.
- [91] Z. Li, J. Sun, Q. Yan, W. Srisa-an, and Y. Tsutano. Obfusifier: Obfuscation-resistant android malware detection system. In S. Chen, K.-K. R. Choo, X. Fu, W. Lou, and A. Mohaisen, editors, *Security and Privacy in Communication Networks*, pages 214–234, Cham, 2019. Springer International Publishing.
- [92] L. Zhang, H. Meng, and V. L. L. Thing. Progressive control flow obfuscation for android applications. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 1075–1079, 2018.
- [93] V. Rastogi, Y. Chen, and X. Jiang. Droidchameleon: Evaluating android anti-malware against transformation attacks. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security, ASIA CCS '13*, page 329–334, New York, NY, USA, 2013. Association for Computing Machinery.

- [94] D. Maiorca, D. Ariu, I. Corona, M. Aresu, and G. Giacinto. Stealth attacks: An extended insight into the obfuscation effects on android malware. *Computers & Security*, 51:16–31, 2015.
- [95] W. Hou, D. Li, C. Xu, H. Zhang, and T. Li. An advanced k nearest neighbor classification algorithm based on kd-tree. In *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*, pages 902–905, 2018.

# واژه‌نامه

## الف

ابتکاری heuristic.....  
 ابعاد بالا high dimensions.....  
 اریب bias.....  
 آستانه threshold.....  
 اصل لانه‌ی کبوتری pigeonhole principle.....  
 ان‌پی-سخت NP-Hard.....  
 انتقال transition.....

## ت

تجربی experimental.....  
 تراکم density.....  
 تقریب approximation.....  
 تقسیم‌بندی partition.....  
 توری mesh.....  
 توزیع‌شده distributed.....

## ب

برخط online.....  
 برنامه‌ریزی خطی linear programming.....  
 بهینه optimum.....  
 بیشینه maximum.....

## ج

جدایپذیر separable.....  
 جعبه سیاه black box.....  
 جویبار داده data stream.....

## پ

پرت outlier.....  
 پرسمان query.....  
 پوشش cover.....  
 پیچیدگی complexity.....

## ح

حدی extreme.....  
 حریصانه greedy.....

## خ

خوشه cluster.....  
 خطی linear.....



## د

داده ..... data  
داده‌کاوی ..... data mining  
داده‌ی پرت ..... outlier data  
دوبرابر سازی ..... doubling  
دودویی ..... binary

## ف

فاصله ..... distance  
فضا ..... space

## ق

قطعی ..... deterministic

## ر

رأس ..... vertex  
رسمی ..... formal

## ک

کارا ..... efficient  
کاندیدا ..... candidate  
کمینه ..... minimum

## ز

زیرخطی ..... sublinear

## م

مجموعه ..... set  
مجموعه هسته ..... coreset  
سطح ..... planar  
موازی سازی ..... parallelization  
میان گیر ..... buffer

## س

سرشکن ..... amortized  
سلسله مراتبی ..... hierarchichal

## ش

شبه کد ..... pseudocode  
شیء ..... object

## ن

نابه جایی ..... inversion  
ناوردا ..... invariant  
نقطه‌ی مرکزی ..... center point  
نیم فضا ..... half space

## ص

صدق پذیری ..... satisfiability

## ه

هزینه‌ی آشوب ..... price of anarchy (POA)

## غ

غلبه ..... dominate

## ی

یال ..... edge

پیوست آ

مطالب تکمیلی

پیوست‌های خود را در صورت وجود می‌توانید در این قسمت قرار دهید.

## **Abstract**

We present a standard template for typesetting theses in Persian. The template is based on the `XYPersian` package for the `LATEX` typesetting system. This write-up shows a sample usage of this template.

**Keywords:** Thesis, Typesetting, Template, `XYPersian`



Sharif University of Technology  
Department of Computer Engineering

M.Sc. Thesis

# **Performance Improvement of Android Repackaged Applications**

By:

**Mojtaba Moazen**

Supervisor:

**Dr. Amini**

february 2023