PRIRODOSLOVNO-MATEMATICKI FAKULTET, SVEUCILISTE U ZAGREBU
MATEMATICKI ODSJEK

Detekcija pakiranih datoteka

Jurica Miletic, Roko Kokan, Ante Sosa

April 30, 2018

1 UVOD

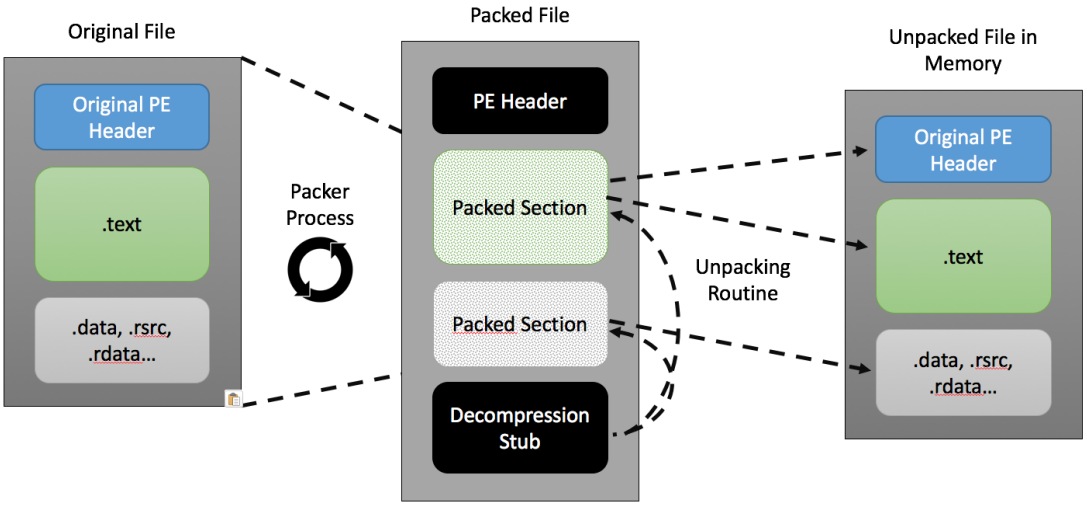
Pakiranje je metoda izmjene izvršnih datoteka bez mijenjanja njihove izvorne funkcionalnosti, ali na način da se datoteka zaštiti od reverznog inženjeringa, da se smanji veličina originalne izvršne datoteke, ili da se prikrije zlonamjeran izvršni kod. Pakiranje podrazumijeva izmjenu sadržaja datoteke te dodavanje instrukcija koje će prilikom izvršavanja taj sadržaj obnoviti.

Packeri modificiraju originalnu izvršnu datoteku na razne načine:

- Kompresijom podataka
- Enkripcijom podataka
- Prikrivanjem (obfuscate)
- Dodavanjem detekcije izvršavanja unutar debuggera ili virtualnog računala
- Modificiranjem raznih dijelova formata izvršne datoteke

Na Slici 1.1 je prikazan proces pakiranja.

U području računalne sigurnosti posebno su učestali packeri za Windows Portable Executable, tj. PE datoteke.



Slika 1.1:

1.1 Opis problema

PE datoteke su povijesno najčešći nositelji malicioznog koda u obliku virusa, ransomwarea, trojanskih konja, itd., te se packeri koriste da bi se taj maliciozni kod prikrio. Klasična statička analiza (bez pokretanja datoteka) koju provode antivirusi bazira se na potpisima. Oni nastaju tako da se prikupe primjeri nekog malwarea te se pronađe niz byteova specifičan za taj malware, koji se zatim traži prilikom skeniranja datoteka antivirusom.

Primjenom packera mijenja se sadržaj datoteke, zbog čega potpisi mogu prestati biti prisutni. Na taj se način iz jedne maliciozne datoteke može napraviti više različitih inačica. Packeri koji imaju nemalicioznu primjenu vrlo su rijetki u odnosu na packere koji se koriste za malware.

1.2 Skup podataka

S obzirom da rjesavamo problem sa natjecanja Mozgalo, skup podataka za ovaj problem nam je omogucio zlatni partner Mozgala, tvrtka ReversingLabs. Oni su razvili ReversingLabs TitaniumCore^{TM} platformu koja prepoznaje PE packere koriste TitaniumCore potpise koje pišu stručnjaci za reverzno inženjerstvo i analizu sigurnosnih prijetnji.

Skup podataka bazirat će se na skupu raznovrsnih packera i dodatnim nepakiranim datotekama. Svaki primjer se sastoji od dva dijela: originalne datoteke i TitaniumCore izvještaja za tu datoteku.

2 Cilj i hipoteze istraživanja problema

Pristup pisanja potpisa za pojedine *packere* je vremenski zahtjevan te zahtjeva vec izdvojene pakiranih datoteka.
Cilj istrazivanja problema je, uz dani skup podataka, napraviti sustav za detekciju pakiranih datoteka.

Pojedine znacajke unutar TitaniumCore izvjestaja nam daju odredene naznake da bi datoteka mogla biti pakirana, poput imena odjeljaka (*Section name*) i entropije dane datoteke.
Primjerice, nekad imena razlicitih odjeljaka nose ime pojedinih packera (UPX), te je entropija pakirane datoteke generalno puno veca nego u obicnim datotekama.
No, problem je sto dosta stvari koje se prikazuju u TitaniumCore izvjestaju mogu biti rucno mijenjane u danoj datoteci pa ne mozemo uzeti pojedinu znacajku zdravo za gotovo.
Zbog svega navedenoga, hipoteze naseg istrazivanja su:

- Pojedine znacajke u TitaniumCore reportu su usko vezane uz to je li neka datoteka pakirana ili ne
- Kombinacijom tih znacajki nadziranim ucenjem, mozemo postici visok rezultat
- Prepoznavanje malware kampanja
- Lakši odabir packera za koje se isplati razvijati unpackere

Metoda automatske detekcije pakiranih datoteka omogucila bi:

- Izdvajanje zanimljivih malware datoteka za detaljniju analizu
- Ranu detekciju nikad prije viđenog malwarea
- Prepoznavanje malware kampanja
- Lakši odabir packera za koje se isplati razvijati unpackere

3 Pregled dosadašnjih istraživanja

Analiza entropije se koristi za uvid u sadržaj PE datoteka, primarno za detekciju kompresije i kriptografije tipično vezanih uz *packere*. Dio pristupa se bazira na analizi svojstava PE zaglavlja i jednostavnim klasifikatorima. Također, znatno poboljšanje točnosti detekcije malwarea dobiveno je koristeći stablo odluke za detekciju pakiranja.

Radovi koji koriste ovaj pristup kvalitetan su izvor analize značajnosti komponenata PE formata, ali su isto tako skloni pretreniranju zbog ograničenosti podataka s kojima rade.