Cihaz tanimlama veriseti bulmak

Cihaz tanimlamayla calismalarinda en buyuk sorunlardan birisi yeterince veriseti bulunmamasidir. Pek cok ag calismasinda kullanilan similasyonlarin, cihaz tanimlama calismalarinda kullanilamamasi, verisetinin ancak gercek veriyle olusturulabilmesi zorluklarin en onde gelenidir. duzgun bir Cihaz tanimlama verisetini olusturmak icin pek cok turde IoT cihaza gereksinim duyulur ki bu cihazlarin temini ciddi bir maddi yuk demektir. Ayrica normal veri toplanmasi, uzun sure, is gucu, ve ozellesmis mekan gerektirir. Baseline olan IoTDevID’de cihaz tanimlama calismalari icin uretilmis 2 veriseti olan Altoo university ve UNSW verisetlerini kullanmistik. Bu calismada ilk verisetini metodumuzu gelistirmek, ikinci verisetini ise sonuclarimizin saglamasini almak icin kullanmistik. 2022 yilinda, yeni bir cihaz tanimlama veriseti olan CIC-IoT-22 halka acildi. Bu veriseti oldukca fazla turden ve addette cihaz icermesi, cihazlarin farkli kosullar altindaki durumlarini icermesi ve benign veriye ek olarak saldiri verisi de bulundurmasi bakimindan oldukca ilgi cekicidir. Bu minvalde verisetinin analizini yapacagiz.

Bu veriseti 6 farkli durumda veri toplanmistir. Bu durumlar soyle özetlenebilir.

Power durumunda her bir cihaz, diğer cihazlardan izole edilerek reboot edilerek bu cihazla ilgili ag paketleri toplanir.

Interactions durumunda cihazla tuslar, uygulamalar veya ses komutlariyla etkilesime girilerek bu sirada urettigi ag paketleri toplanir.

Scenarios bu cihazlarin eve giris, evden cikis, gece ve gunduz eve izinsiz giris veya kullanici hatasi gibi durumlardaki ag verileri toplanir.

Saldiri verisinde Flood attacks and RTSP Brute Force attacks uygulanarak veriler toplanir.

IDLE verisinde, cihazlardin calistigi ancak aktif olarak kullanilmadigi aksam saatlerinde 8 saatlik dilimin 30 gun boyunca kaydedilmesinden olusur.

Active cihazlarin 30 gun boyunca gun icerisinde kullanilma verilerini icermektedir. Bu veriler insanlarin labratuara girip cihazlari kullanmasiyla olusturulmustur.

Verisetiyle ilgili bazi onemli noktalar.

Bu calismada, bizim icin en onemli kisim IDLE ve Active bolumleridir. Bu kisimda neredeyse tum cihazlardan yeterince veri toplanmistir. Kagitta bu surecte 60 cihaz kullanildigi belirtilse de, hem kendi deneylerimizde hem de veriseti icersinde verilen bilgilere gore bu kisimlain 40 cihaz icermetedir. Bu 40 cihaz yalnizca lan WIFI cihazlardan olusur, zigbee ve z-wave cihazlarini icermezler. zigbee ve z-wave cihazlari power interaction ve hede hodo olmak uzere digger cihazlardan izole edilmis sekilde verileri vardir ancak bu veriler hem oldukca kisitli miktardadir, hem de normal kullanim verisi icermezler.

Veriseti toplamda 60 cihaz icersede, bizim odaklanacag

Ozellik cikartma:

Ozellik cikartma islemi icin Python ve Scapy ve WS kullanilmistir. Ozellik cikamada yalnizca tekil paket bazli ozellikler kullanilmistir. Sistem toplamda 123 ozellik cikartsa (ozelliklere ve aciklamalarina tablo hededen ulasilablir)da bu ozelliklerin cok azi, yalnızca IoTDevID calismasinda secilen alt ozellikler deneyde kullanilmistir.

Etiketleme:

Etiketleme islemini veriseti icerisinde bulunan cihaz isimleri-MAC adresleri listesi kullanilarak yapilmistir. Cikarilan her parmakizinde kaynak MAC adresi kismi, verilen isimle degistirilmek suretiyle etiketleme yapilmis, Bu listede verilmeyen MAC adresleri (hub, swich veya verinin toplandigi bilgisayara ait oldugu kanaatindeyiz 5 adet MAC adresi gormezden gelinmistir.

Veriseti icerisinde kaynak MAC adresleri,

Ozellik cikarmada kullandigimiz verisetlerinin herbiri birgunde kayit altina alinmiş, ve kaydedildiği tarih ile isimlendirilmiş verilerdir. 24.11.2021 tarihinde kaydedilmiş bir veri Active ise A211124, Idle ise I211124 olarak isimlendirilmistir. Bu baglamda 30 IDLE, 24 tane ise aktif oturum kaydedilmistir. On calisma olarak, tum bu oturumlari birbirleriyle kiyaslayarak basarimlari test etmeyi amacladik. Oturumlarin birbiriyle kiyaslanabilmesi icin benzer cihazlari ivermeleri gerekir. Ancak ne yazik ki, her oturumda her cihazdan veri toplanmamis, bazi oturumlarda bazi cihazlar hic veri uretmemislerdir. Her oturumda hangi cihazin ne kadar veri urettigi tablo 12de verilmistir. Bu yuzden, yalnizca ayni cihazlari iceren oturumlari birbirleriyle kiyasladik. Bu kiyaslama islemi icin bir oturum kimligi olusturduk. Ornegin sessin1 A, ve C cihazini icerirken B cihazini icermesin bu durumda kimlik numarasi 101 olur. Sessin1 ayni kimlik numarasina sahip digger oturumlarla sorunsuz sekilde kiyaslanablir. Bu baglamda Her oturum icin toplamda 40 cihaz uzerinden 40 basamakli bir ID olusturduk.

Ayni IDye sahip cihazlari egitim ve test verisi olarak kullanmak suretiyle elde etttigimiz sonuclar sekil 1-2-de verilmistir. Bu sonuclarin sunumunda F1 skor kullandik. Bunun kabaca iki nedeni vardir. Birincisi f1 skor dogrulugun aksine, dengesiz verisetlerinde guvenilir sonuclar verir. Ikincisi yalnizca overall sonuclar vermez, sonuclari sinif bazli inceleme imkani da verir. Bu baglamda sonuclar incelendiginde ikili oturum kiyaslamalarinda f1 skorun 40-88 arasinda degistigi gorulur. Burada dikkat cekmek istedigimiz bir baska nokta bu islemin yaklasik 40 sinifli bir coklu siniflandirma islemi olmasidir. Bu baglamda 40 bile, sans basarisindan oldukca uzak bir sonuctur.

Sekiller incelendiginde sekil 1de spesifik tarihlere rastgeelen sonuclarin (211108, 211109, 211206, 211208, 211223, 211225, 211228) basarisiz oldugu digger taraftan sekil 2 incelendiginde ozellikle birbirini takip eden belli tarih araliklarindaki verilerin daha basarili oldugu gorulmektedir.

Bu sonuclar overall basriyi yansitmaktadir. Tekil sonuclara odaklanmanin daha fazla velgi verecegi kanaatindeyiz. Toplamda 600den fazla oturum kiyaslamasi yaptik. Bu kiyaslamada ilk oturum egitim ikincisi test olarak kullanildi. Eger bu oturumlari aktif ve Idle olarak ayirirsak 4 farkli olasilik mumkundur AA,AI,IA ve II. Bu baglamda oturum kiyaslamalarinin dagilimi sekilde verilmistir. Her oturum icin cihaz temelli sonuclari inceleyerek, sorunlu cihazlara odaklanmak istiyoruz. Bu baglamda oturumlarin herhangibirinde basarisiz olan, sinif bazli 0.50 f1 skordan asagi alan bir cihaz, bu davranisini 12den fazla tekrarlarsa (12 tum oturum kiyaslamasinin %2sine karsilik gelir), listemize eklenir. Grafik 12 de bu liste verilmistir. Liste cihazi, kackez hata verdigini ve bu hatalarin oturum kiyaslama turlerine gore dagilimini gostermektedir.

Genel manada sonuclar incelendiginde,

I211124

verilerdeir.

Fig.~\ref{fig:bar} incelendiginde, bazi ufak istisnalar haricinde, pasta grafikteki genel dagilimin ayni sekilde korundugunu gormekteyiz. Bu bize idle ile active arasinda anlamli bir fark olmadigini gostermektedir. Diger taraftan, basarimi dusuk olan bazi cihazlara odaklandigimizda bunlarin neden listeye eklendigini rahatlikla anlayabiliriz. En fazla basarizliga ugrayan cihazlarin Amazon Alexa Echo Dot, Gosund Plug, Gosund Socket, Teckin Plug, Yutron Plug gibi deney setinde birden fazla ornegi olan cihazlar oldugu gorulur. Bu cihazlar ayni cihazin (ayni marka ve model) farkli ornekleri olduklari icin tek tek etiket altinda toplanmalari gerekir (ornegin Teckin Plug 1 ve Teckin Plug 2 -> Teckin Plug ). Bunlarin disindaki cihazlarin pek cogunun basari seviyesinin orneklem cesitliligini arttirarak yukseltilebilecegini dusunuyoruz. Bu baglamda birden fazla oturumdan orneklerm alarak, orneklem cesitliligini arttirmayi, bunun bir sonucu olarak da model basarisini yukseltmeyi hedefledik.

Bu baglamda tum oturumlari agitim ve test olmak uzere oturumlar birbirinden izole olacak sekilde ayirdik. egitim ve test olarak belirledigimiz oturumlari birlestirerek oldukca buyuk bir veriseti elde ettik. Bu verisetini olusturuken idle ve active verilerini kendi cilerinde izole ettik. sadece idle verilerinden olusan Idle train ve test, sadece active verilerinden olusan active train ve test olmak uzere 2 cift veriseti elde ettik. Ancak active oturumlarinin icermedigi bir cihaz olan (cihazinin verilerini ) idle oturumlarindan aldik. Bir baska degisiklik de LG TV ile ilgiliydi. Bu cihaza ait veri 54 oturumun yalnizca uc tanesinde vardir. Ancak bu cihazin verisi o kadar dengesidir ki yalnizca 3 oturumdan toplanan cihaz verisi toplam paket sayisinin yaklasik yuzde dokuzunu olusturur. Bu cihazin hem yeterince oturumda verisi olmamasi hem de, asiri paket sayisiyla veriseti dagilimini bozmasi nedeniyle verisetinden cikardik.

Oturumlardaki cesitliligi yansitan ancak cok da buyu olmayan bir veriseti elde etmek amaciyla, olusturdugumuz bu 4 verisetini toplam paket sayilarini \%10 olacak kadar kuculttuk. Bu islem sirasinda rastgele orneklerm kullandigimiz icin, cihazlardan elde edilen paket oranlari sabit kaldi boylece, verisetinin dogal dagilina zarar vermedik.

olusturdugumuz 4 farkli durumun F1 skor cinsinden kiyaslamasi grafikte verilmistir.