DDoS Evaluation Dataset (CICDDoS2019)

Alan Barzilay - 8639515

Caio Lorenzetti Martinelli - 8539899

Origem dos dados

Esse dataset foi gerado conforme a descrição deste paper que buscava trazer uma nova forma de classificar ataques DDoS e gerar um dataset moderno e descritivo desse tipo de ataque, toda a motivação partiu da inexistência de um dataset satisfatório.

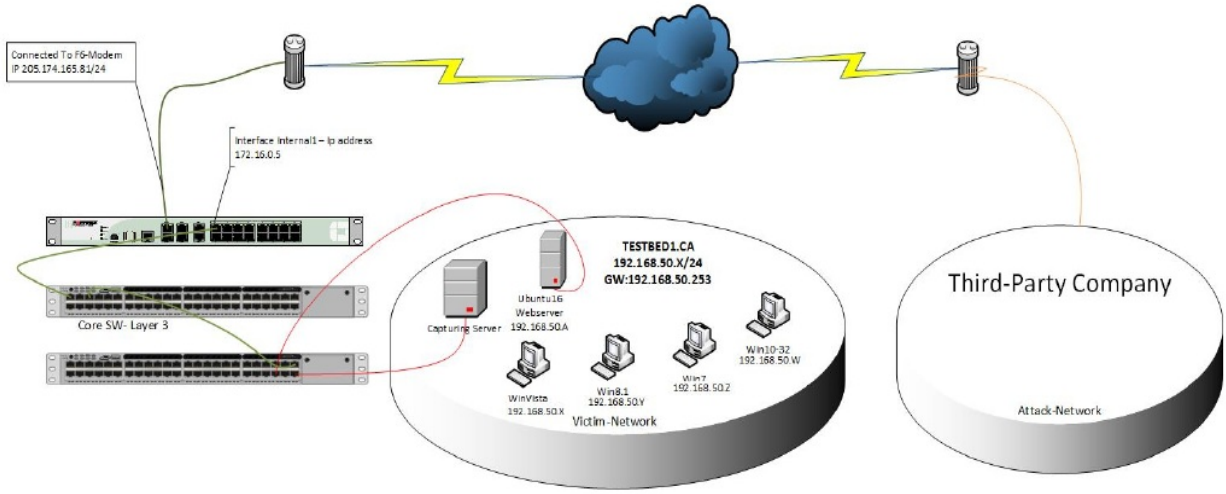


Figure 2: Testbed Architecture

Resumindo o trabalho realizado, eles simularam um trafego padrão, que eles chamaram de benigno, a partir de um perfil gerado em cima de dados de usuários reais. Uma vez criado esse stream benigno de pacotes, eles criaram uma rede capaz de gerar diversos tipos de ataques DDoS distintos e capturaram todos os pacotes que chegavam na rede vítima. O *Testbed* está esquematizado na figura acima. A ideia é que agora eles possuem um dataset próximo o suficiente da realidade para ser utilizado em tarefas mais complicadas como detecção de ataques e a classificação deles. Esse dataset de treino possui 11 tipos de ataques, no paper e na descrição dos dados eles citam 12 tipos de ataques mas nunca explicam de maneira compreensível porque o dataset possui um cenário a menos. A justificativa encontrada para a não inclusão do ataque WebDDoS foi: "*The traffic volume for WebDDoS was so low and PortScan just has been executed in the testing day and will be unknown for evaluating the proposed model*". Os pacotes capturados foram armazenados no formato PCAP e em seguida foram tratados por uma ferramenta chamada CICFlowMeter. Essa ferramenta pega os PCAPS e gera o "flow" dos pacotes para poder extrair features dos flows. Aqui pode ser encontrado uma lista completa das features geradas por essa ferramenta, o nosso dataset possui apenas um recorte de 80 destas features. A escolha das features não foi justificada no trabalho original.

Com o dataset de treino em mãos, os autores geraram um novo conjunto de ataques com apenas 7 categorias desta vez, 6 delas presentes no conjunto de treino e uma categoria de ataque nova (referido como PortScan no paper e Portmap nos dados). Os autores realizaram uma pré-analise onde decidiram quais seriam as features mais relevantes através do uso de uma random forest e por fim eles realizaram testes com 4 modelos diferentes (ID3, Random Forest, Naive Bayes, Multinomial Logistic Regression) para tentar detectar ataques DDoS no conjunto de teste. Nem o código nem os parâmetros utilizados foram compartilhados no trabalho original.

DataSet

	CSV	PCAP.zip
Treino	22Gb	20.9Gb
Teste	8.7Gb	2.0Gb

Os conjuntos em formato CSV possuem um arquivo CSV por categoria de ataque realizado.

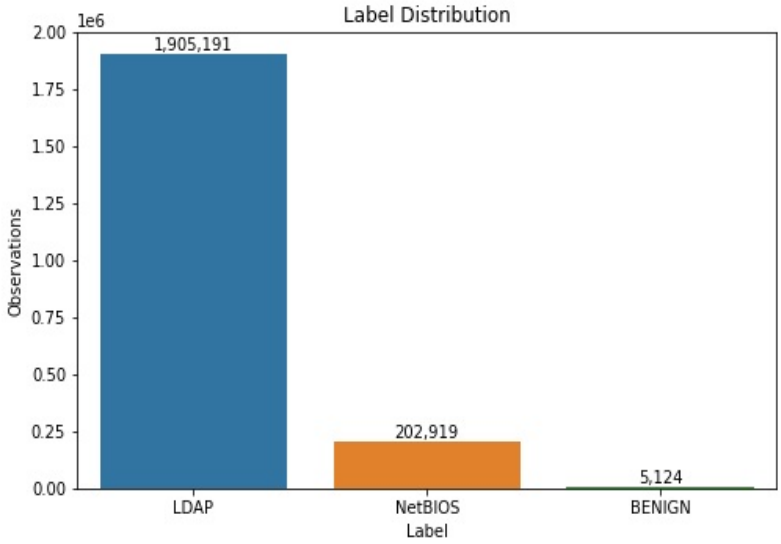
Os conjuntos em formato PCAP.zip são compostos por uma coleção de arquivos PCAP de 200Mb cada e organizados de maneira sequencial representando a totalidade dos pacotes transmitidos. Os dados PCAP não comprimidos estão na ordem centenas de Gb e não cabem no disco da maquina disponível.

Estudo de caso, o ataque LDAP

Para ilustrar algumas características do nosso dataset, decidimos explorar aqui apenas um ataque. Como os diferentes ataques apresentam um comportamento semelhante, acreditamos que nesse primeiro momento esse exemplo é suficiente para se ter uma introdução ao dataset. Essas analises foram replicadas para outros ataques e podem ser conferidas neste repositório do github junto com os notebooks que as geraram.

Histograma das Labels

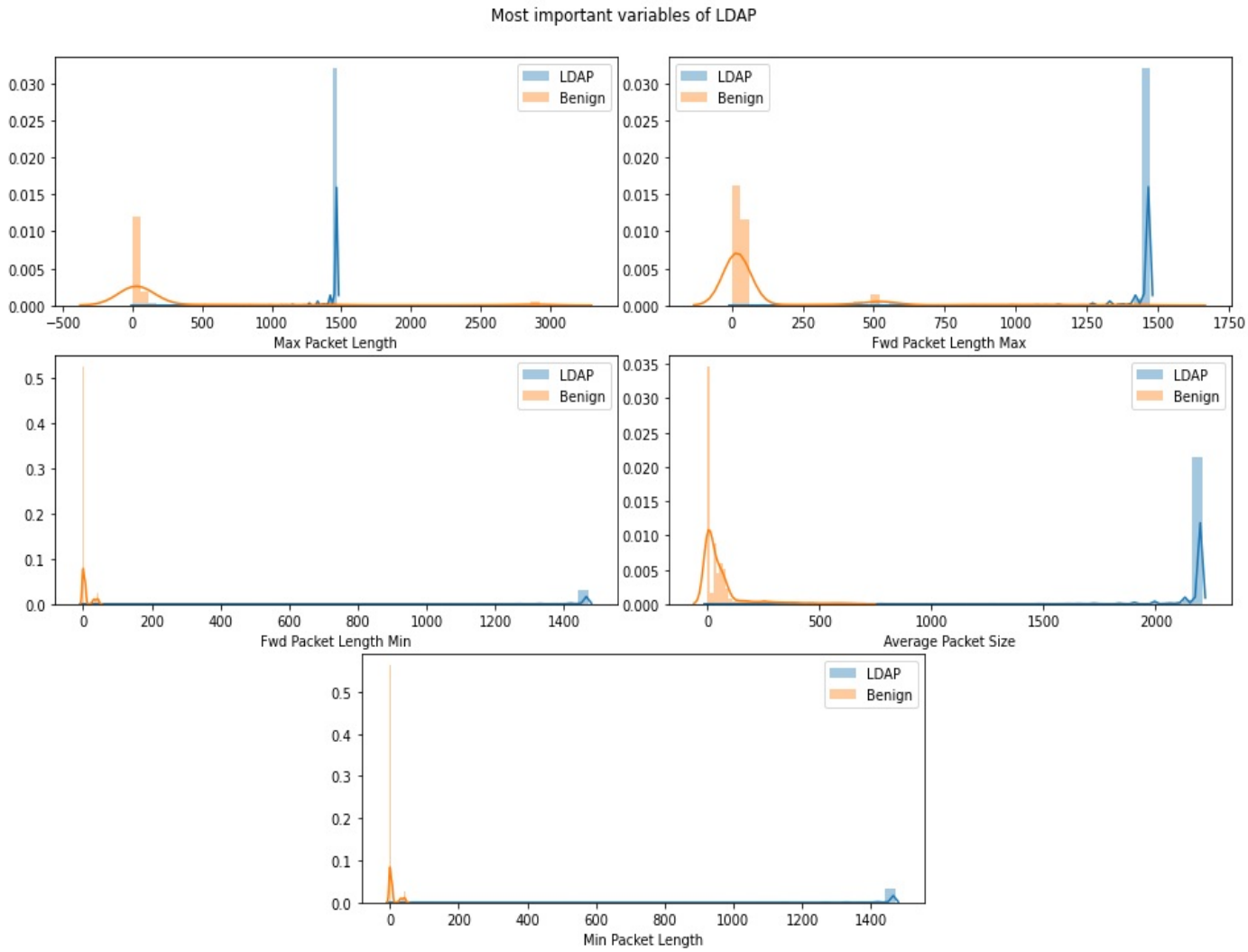
Cada fluxo capturado no dataset possui uma label que o designa como benigno ou como pertencente a um determinado ataque, cada arquivo CSV na teoria possui apenas fluxos de dois tipos, os benignos ou os de um determinado ataque. Aqui plotamos um histograma para as labels do CSV que descreve o ataque LDAP e podemos observar um problema recorrente nesse dataset: a contaminação dos arquivos por fluxos de ataques diferentes. Por não possuir uma janela grande o suficiente de normalidade, o dataset acaba não respeitando as fronteiras entre os ataques. Também podemos notar o grande desbalanceamento de classes. É de se esperar que haja uma certa desproporção entre as classes pela própria natureza dos ataques DDoS, mas essa desproporção é agravada pela total



ausência de um trafego regular.

Histograma das features mais relevantes

De acordo com a analise realizada pelo paper original, as features aqui representadas são as mais relevantes para se descrever e detectar um ataque do tipo LDAP. Podemos notar como o perfil do histograma para cada uma dessas features é bem característico e difere entre os pacotes benignos e os do ataque.



Críticas ao dataset e seu paper

- A ferramenta utilizada para processar os dados, CICFlowMeter, foi desenvolvida pelos próprios autores do paper no passado e não parece ter sido utilizada por muitos outros pesquisadores. Ela é aberta num repositório do github porem não possui documentação relevante e sofre de diversas más praticas de programação.
- É comum não existirem intervalos de normalidade entre ataques. As divisões no tempo descritas no paper não existem e isso pode ser constatado na analise que realizamos. Existe uma contaminação entre os arquivos onde pacotes de um determinado ataque podem ser encontrados no arquivo representante do ataque imediatamente anterior ou posterior a ele. Isso tambem agrava o grande desbalanceamento entre as classes.
- A geração dos CSVs a partir dos PCAPS não é reproduzível, não compartilharam código ou parâmetros utilizados.
- A criação dos modelos preditivos e suas métricas não é reproduzível, não compartilharam código ou parâmetros utilizados.
- Não está claro como os modelos foram testados, a existência de uma classe nova no conjunto de treino nos leva a imaginar que os modelos tem como objetivo detectar se há um ataque DDoS ocorrendo ou não sem se importar com classificá-lo. Porém eles se utilizam de um modelo de regressão logística multinomial, o que nos leva a acreditar que os modelos buscavam detectar e classificar o tipo de ataque. Se esse for o caso, não conseguimos compreender a motivação por incluir uma nova classe no conjunto de teste que não estava presente no conjunto de treino para um classificador.
- O paper possui diversos erros de digitação e por vezes é pouco claro.
- Todos os modelos utilizados foram testados a posteriori e não são capazes de detectar um ataque em tempo real.

Ideias possíveis de serem exploradas em nosso paper

- Gerar um modelo “online” capaz de receber o stream de dados e detectar ataques em tempo real. Uma possivel dificuldade que poderemos enfrentar com essa ideia seria na simulação online, como organizar a entrega dos PCAPS de maneira a recriar as suas capturas e gerar um cenario realista para o treino e teste.

- Explorar o desbalanceamento das classes, a proporção de pacotes benignos para pacotes malignos é extremamente baixa.
- Explorar os dados não processados e gerar novas features a partir dos PCAPS que poderiam se mostrar mais relevantes do que as utilizadas nesse dataset.
- Explorar algoritmos mais modernos do que os testados nesse paper.