Identificació d'aplicacions amb NetFlow

Pere Barlet
CCABA (UPC)

La classificació de tràfic, i en particular la identificació d'aplicacions de xarxa, és important per a les tasques de gestió i administració de la xarxa. Aquest, però, és un problema complex que requereix l'ús de tècniques d'identificació molt sofisticades, donada la naturalesa canviant del tràfic internet i de les seves aplicacions. En conseqüència, aquest problema ha atret l'interès tant dels operadors de xarxa com de la comunitat científica.

Tradicionalment, els números de port s'han utilitzat per identificar el tràfic d'internet (p. e. els *well-known ports* registrats per la IANA). No obstant això, avui dia és àmpliament conegut que aquest mètode no és vàlid per classificar el tràfic de xarxa actual, a causa de la imprecisió dels seus resultats. L'alternativa més comuna als números de port és la inspecció del contingut dels paquets. Aquest mètode, conegut també com a Deep Packet Inspection (DPI), consisteix a cercar patrons coneguts o signatures en el camp de dades dels paquets. Tot i que aquesta solució potencialment pot ser molt precisa, el seu alt consum de recursos fa inviable el seu ús en les xarxes actuals d'alta velocitat. A més, algunes aplicacions, com ara el P2P, han començat a implementar tècniques sofisticades d'ofuscació del protocol per tal de camuflar el seu tràfic, a part d'utilitzar ports no estàndard o ports d'altres aplicacions, per evadir la detecció o travessar tallafocs.

Per solucionar aquests problemes, la comunitat científica ha proposat diverses tècniques d'aprenentatge automàtic (AA) per a la identificació d'aplicacions. En termes generals, la majoria d'aquestes tècniques es basen en analitzar, en una fase fora de línia, la relació entre un conjunt predefinit d'atributs del tràfic, com ara els números de port, la mida del flux, el temps entre arribades de paquets, etc. i cada aplicació. Aquest conjunt d'atributs és utilitzat per construir un classificador (p.e. arbre de decisió) que posteriorment és utilitzat per identificar el tràfic en línia.

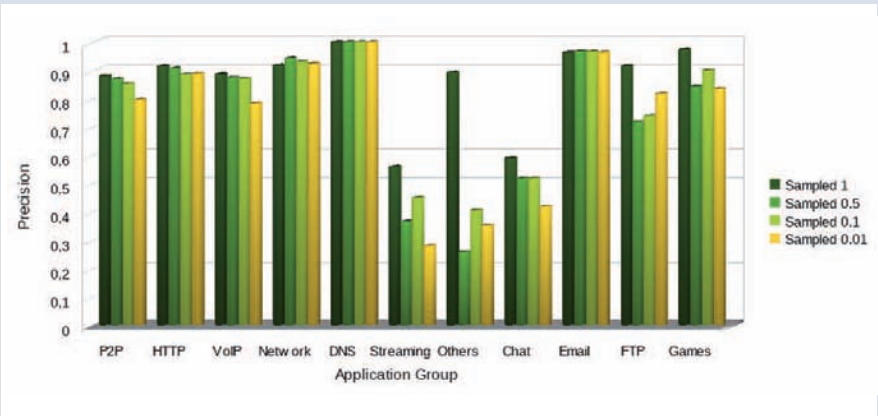
Tot i que existeix un ampli ventall de propostes d'AA per a la identificació d'aplicacions, hi ha alguns aspectes importants que encara romanen sense investigar. Com a resultat, la majoria de tècniques d'AA han tingut un èxit molt limitat entre els operadors de xarxa. Per exemple, la majoria de sistemes de monitoratge de xarxa encara utilitzen els números de ports o tècniques de DPI per identificar el tràfic.

Entre aquests problemes oberts s'inclouen els següents: la majoria de tècniques d'AA només poden operar en traces a nivell de paquet, la qual cosa requereix el desplegament de maquinari addicional de monitoratge molt costós; l'impacte del mostratge de tràfic (*sampling*) en aquestes tècniques d'AA encara és desconegut, tot i que el mostratge, com ara el Sampled NetFlow, és una pràctica habitual entre els operadors; la majoria de les propostes d'AA necessiten d'una fase d'entrenament molt costosa.

En aquest treball s'han investigat aquests tres problemes oberts. En primer lloc, s'ha estudiat el problema de la identificació d'aplicacions utilitzant NetFlow, en comptes de traces a nivell de paquet. Per a aquest efecte, s'ha adaptat el mètode d'AA proposat [1]. En segon lloc, s'ha analitzat l'impacte del mostratge de tràfic en la precisió d'aquesta tècnica d'identificació, que és molt important per les baixes taxes de mostratge utilitzades pels operadors de xarxa (p.e. 1/1000) per fer front a atacs o anomalies de xarxa. En últim lloc, s'ha proposat un mètode automàtic d'entrenament que no requereix la inspecció manual del conjunt d'entrenament i redueix significativament l'impacte del mostratge de tràfic en la precisió dels mètodes d'identificació d'aplicacions basats en AA.

La figura mostra la precisió d'aquest nou mètode de detecció d'aplicacions a la xarxa de la UPC utilitzant NetFlow amb diferents taxes de mostratge (p={*50%, 10%, 1%*}). La precisió mitjana per a p=*1%* va estar al voltant del 85% mentre que, amb la tècnica original proposada a [1] i el mètode basat en els números de port, la precisió va ser inferior al 50% i al 15%, respectivament. Tanmateix, la precisió per algunes aplicacions (p.e. *Streaming*, *Others* i *Chat*) va ser especialment baixa, perquè el conjunt d'entrenament incloïa poc tràfic d'aquestes aplicacions. ■

Precisió per grup d'aplicació del mètode d'identificació d'aplicacions basat en AA



[1] P. Barlet-Ros, E. Codina, i J. Solé-Pareta. "Identificación de aplicaciones de red mediante técnicas de aprendizaje automático". *Boletín de RedIRIS*, 82-83, abril 2008.