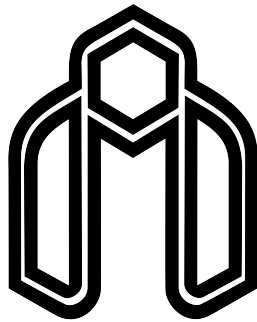


الحمد لله
الرحمن الرحيم



دانشگاه صنعتی شاهرود

پردیس بین المللی خوارزمی

گروه هوش مصنوعی

رساله دکتری مهندسی هوش مصنوعی

استخراج ویژگی‌های آماری ترافیک شبکه جهت شناسایی
برنامه‌های کاربردی در مسیر یاب‌ها

نگارنده:

محمد رضا گندمی

استاد راهنما:

دکتر حمید حسن‌پور

استاد مشاور:

دکتر نعمت‌اله شیری (دانشگاه کنکردیا کانادا)

تیر ۱۳۹۷

شماره: ۱۳۹۷/۴/۲۶

تاریخ: ۱۳۹۷/۴/۲۶

ویرایش:

باسمه تعالی



مدیریت تحصیلات تکمیلی

فرم شماره ۱۲: صورت جلسه نهایی دفاع از رساله دکتری (Ph.D)

(ویژه دانشجویان ورودی های ۹۴ و ما قبل)

بدینوسیله گواهی می شود آقای محمد رضا گندمی دانشجوی دکتری رشته علوم کامپیوتر به شماره دانشجویی ۹۳۰۱۲۵۵ ورودی مهر ماه سال ۹۳ در تاریخ ۹۷/۰۴/۱۰ از رساله نظری / عملی خود با عنوان: استخراج ویژگی های آماری ترافیک شبکه جهت شناسایی برنامه های کاربردی در مسیر یاب ها دفاع و با اخذ نمره ۱۸،۸۸ به درجه بسیار خوب نائل گردید.

☒ الف) درجه عالی: نمره ۱۹-۲۰	☒ ب) درجه بسیار خوب: نمره ۱۸/۹۹ - ۱۷
☐ ج) درجه خوب: نمره ۱۵ - ۱۶/۹۹	☐ د) غیر قابل قبول و نیاز به دفاع مجدد دارد
☐ ه) رساله نیاز به اصلاحات دارد	

ردیف	هیئت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱	حمید حسن پور	استاد/ اساتید راهنما	استاد	
۲	نعمت اله شیرینی	مشاور/ مشاورین	دانشیار	
۳	محمد قاسم زاده	استاد مدعو داخلی / خارجی	دانشیار	
۴	هدی مشایخی	استاد مدعو داخلی / خارجی	استادیار	
۵	امید رضا معروضی	استاد مدعو داخلی / خارجی	استادیار	
۶	علی عباس نژاد	سرپرست (نماینده) تحصیلات تکمیلی دانشکده	استادیار	

مدیر محترم تحصیلات تکمیلی دانشگاه:

ضمن تأیید مراتب فوق مقرر فرمائید اقدامات لازم در خصوص انجام مراحل دانش آموختگی آقای/خانم محمد رضا گندمی بعمل آید.

نام و نام خانوادگی: رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:



تقدیم به پدر و مادر عزیز و همسر مهربانم که در سختی ها و دشواری های زندگی، همواره یآوری دلسوز و فداکار برایم بوده اند.

سپاس‌گزاری...

از استاد گرامی جناب آقای دکتر حمید حسن پور بسیار سپاسگزارم چرا که بدون راهنمایی های ایشان موفقیت در این راه دشوار طاقت فرسا می نمود.

محمد رضا گندمی

تیر ۱۳۹۷

تعهد نامه

اینجانب محمد رضا گندمی دانشجوی دکتری رشته کامپیوتر گرایش هوش مصنوعی دانشکده پردیس خوارزمی دانشگاه شاهرود، نویسنده پایان نامه با عنوان **استخراج ویژگی‌های آماری ترافیک شبکه جهت شناسایی برنامه‌های کاربردی در مسیر یاب‌ها**، تحت راهنمایی آقای دکتر حسن پور متعهد می‌شوم:

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام “ دانشگاه صنعتی شاهرود “ یا “ Shahrood University of Technology “ به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آوردن نتایج اصلی پایان نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

محمد رضا گندمی

تیر ۱۳۹۷

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

شناسایی ترافیک شبکه یکی از نیازهای اساسی مدیران جهت حفظ امنیت، کنترل و برای بهبود کیفیت سرویس دهی در شبکه می باشد. از میان روش های موجود جهت شناسایی ترافیک برنامه های کاربردی، روش های مبتنی بر ویژگی های آماری بسته های ترافیک شبکه مورد توجه بسیاری از محققین قرار گرفته است. در این روش ها، عملیات شناسایی ترافیک برنامه های کاربردی بدون نیاز به بررسی محتوای بسته های ترافیک و با هدف حفظ محرمانگی اطلاعات مبادله شده انجام می شود. بررسی های انجام شده نشان می دهد که روش های موجود برای شناسایی ترافیک شبکه در مواجهه با مسائلی همچون: به روز شدن برنامه کاربردی، رخداد اتلاف بسته در شبکه و افزایش نرخ تبادل اطلاعات کارایی خود را از دست می دهند. هدف اصلی این پژوهش ارائه ویژگی های آماری مقاوم و موثر در برابر ویرایش های جدید یک برنامه کاربردی و رخداد اتلاف بسته در داده های ترافیک شبکه می باشد.

در این تحقیق، به منظور تحلیل رفتار برنامه های کاربردی، نرم افزارهای فایرفاکس، کروم، اینترنت اکسپلورر، اسکایپ، تلگرام، داندلود منیجر و توییتر، براساس ویژگی های آماری، شامل اندازه بسته ها، مدت زمان مابین ارسال و دریافت بسته ها، مدت زمان جریان ها و نرخ ارسال و دریافت بسته ها در شبکه مورد بررسی قرار گرفتند. این ویژگی های آماری، کارایی برنامه های کاربردی در ایجاد، ارسال و دریافت جریانی از بسته ها جهت تعامل بین سرور و کاربران را به خوبی منعکس می کنند. جهت ارزیابی ویژگی های ارائه شده، اطلاعات متقابل بین هر ویژگی و نوع ترافیک برنامه کاربردی محاسبه شده است. اطلاعات متقابل ارزش و توانمندی هر ویژگی در شناسایی برنامه های کاربردی را نشان می دهد. با استفاده از ویژگی های استخراج شده دقت شناسایی به طور میانگین به ۹۷/۵٪ می رسد.

مساله اصلی و چالش برانگیز در خصوص شناسایی ترافیک شبکه با استفاده از ویژگی های آماری، وابستگی این ویژگی ها به عوامل و رخداد های موجود در ساختار شبکه ی بین فرستنده و گیرنده، و حجم ترافیک شبکه می باشد. یکی از اصلی ترین این رخداد ها اتلاف بسته ها در مسیر یاب ها و در طول مسیر بین فرستنده و گیرنده می باشد. در هنگام وقوع اتلاف بسته، دنباله ای از بسته ها از دست می روند، و در نتیجه ویژگی های آماری مورد استفاده جهت شناسایی برنامه کاربردی نیز تغییر می کنند. به همین منظور پایگاه داده ای از هفت برنامه کاربردی نامبرده شده، با نرخ های مختلف اتلاف بسته فراهم شد تا تاثیر اتلاف بسته بر روی ویژگی ها و دقت شناسایی بررسی شود. در ادامه کار با تحلیل فراوانی توزیع داده ها در چهار ویژگی اندازه بسته ها، فاصله زمانی مابین بسته های ارسالی و دریافتی، مدت زمان جریان بسته ها و نرخ ارسال و دریافت بسته ها، ویژگی هایی مقاوم به رخداد اتلاف بسته جهت شناسایی برنامه های کاربردی استخراج گردید. سپس با اعمال این ویژگی ها به یک شبکه عصبی، عمل شناسایی ترافیک شبکه انجام شده است. نتایج حاصل از بررسی روش ارائه شده نشان می دهد که با استفاده از ویژگی های استخراج شده، متوسط دقت شناسایی برنامه های کاربردی هنگام رخداد اتلاف بسته در شبکه به

۹۱/۵٪ می‌رسد که نزدیک به دقت شناسایی در حالت عدم رخداد اتلاف بسته می‌باشد.

کلمات کلیدی: استخراج ویژگی، ترافیک شبکه، شناسایی ترافیک برنامه‌های کاربردی، ویژگی آماری، یادگیری ماشین

مقالات مستخرج از رساله

- M. Gandomi and H. Hassanpour, "Behavioral Analysis of Traffic Flow for an Effective Network Traffic Identification". International Journal of Engineering-Transactions B: Applications, vol.30, no.11, p.1740, 2017.
- M. Gandomi and H. Hassanpour, "Network Traffic Identification in Network Layer Using Behavioral Analysis of Traffic Flow". The CSI Journal On Computer Science And Engineering. (Submitted on May 2018).

- م. گندمی، ح. حسن پور، "استخراج ویژگی جهت شناسایی ترافیک شبکه با در نظر گرفتن اثرات اتلاف بسته ها"، مجله پردازش علائم و داده ها. (پذیرفته شده در اردیبهشت ۱۳۹۷)
- م. گندمی، ح. حسن پور، "شناسایی ترافیک شبکه در مسیریاب ها با حفظ حریم خصوصی کاربران"، سومین کنفرانس پردازش سیگنال و سیستم های هوشمند، آذر ماه ۱۳۹۶، شاهرود.
- م. گندمی، ح. حسن پور، "استفاده از تکنیکهای داده کاوی جهت دسته بندی داده های مهم و حساس"، مجله مطالعات اطلاعاتی. (پذیرفته شده در شهریور ۱۳۹۶)

فهرست مطالب

ف	فهرست اشکال
ش	فهرست جداول
ث	فهرست اختصارات
۱	فصل اول ترافیک شبکه و شناسایی آن
۲	۱.۱ مقدمه
۳	۲.۱ ساختار شبکه و تبادل داده در آن
۴	۳.۱ ساختار بسته‌های ترافیک شبکه
۶	۴.۱ روش‌های شناسایی ترافیک شبکه
۹	۵.۱ چالش‌های پیش‌رو
۱۴	۶.۱ تعریف مساله
۱۵	۷.۱ ضرورت توسعه روش شناسایی ترافیک شبکه
۱۷	۸.۱ فرض‌های مسئله
۱۷	۹.۱ مدل روش ارائه‌شده
۲۱	۱۰.۱ پایگاه داده‌های استفاده شده در رساله
۲۳	۱۱.۱ ساختار رساله
۲۵	۱۲.۱ جمع‌بندی
۲۷	فصل دوم مروری بر روش‌های شناسایی ترافیک شبکه بر پایه ویژگی‌های آماری
۲۸	۱.۲ مقدمه

۲۸		رهیافت روش‌های شناسایی ترافیک بر پایه ویژگی‌های آماری	۲.۲
۳۰		کارهای انجام شده در زمینه شناسایی بر پایه ویژگی‌های آماری	۳.۲
۴۲		جمع‌بندی	۴.۲

فصل سوم تحلیل رفتاری ترافیک شبکه برای شناسایی برنامه‌های کاربردی

۴۴		مقدمه	۱.۳
۴۴		تحلیل رفتاری	۲.۳
۵۲		اطلاعات متقابل	۳.۳
۵۸		اتلاف بسته و تاثیر آن بر شناسایی ترافیک	۴.۳
۶۷		جمع‌بندی	۵.۳

فصل چهارم ارزیابی روش‌های ارائه شده

۷۶		مقدمه	۱.۴
۷۶		معیارهای ارزیابی روش‌های شناسایی ترافیک	۲.۴
۷۹		معیارهای ارزیابی روش ارائه شده	۳.۴
۸۰		ارزیابی ویژگی‌های رفتاری	۴.۴
۸۸		ارزیابی ویژگی‌های مرتبه سوم	۵.۴
۹۰		جمع‌بندی	۶.۴

فصل پنجم جمع‌بندی و کارهای آینده

۹۲		مقدمه	۱.۵
۹۲		خلاصه‌ای از رساله	۲.۵
۹۴		پیشنهادهایی برای ادامه کار	۳.۵

پیوست اول مبانی شبکه

۹۸		ساختار شبکه	۱.آ
۹۹		سطوح مختلف ترافیک	۲.آ
۱۰۲		تبادل داده در شبکه	۳.آ

۱۰۸	مراجع
۱۱۳	واژه نامه انگلیسی به فارسی
۱۱۴	واژه نامه فارسی به انگلیسی
۱۱۷	نمایه

فهرست تصاویر

۱.۱	 اجرای برنامه‌های کاربردی مختلف بر روی رایانه کاربر شبکه	۲
۲.۱	 بازگشایی بسته‌ها در مسیر یاب تا لایه شبکه	۴
۳.۱	 پارامترهای D_i و T_i ، d_i ، u_i ، t_i	۵
۴.۱	توزیع اندازه بسته‌های شش برنامه کاربردی در دوره‌ها و ساختارهای مختلف. (آ). (Skype)، (ب. Safari)، (ج. Torrent)، (د. Firefox)، (ه. Mail)، (و. Thunderbird)،	
۵.۱	توزیع و فراوانی اندازه سرآیندها برنامه کاربردی Skype در دوره‌ها و ساختارهای مختلف. (آ.ج. سال ۲۰۰۹)، (ب.د. سال ۲۰۱۶)،	۱۰
۶.۱	 نمایی از قابلیت تنظیم شماره درگاه پویا در Skype Ver 6.11	۱۳
۷.۱	 اجرای برنامه‌های کاربردی مختلف بر روی شبکه	۱۴
۸.۱	 ارسال بسته‌های برنامه‌های کاربردی مختلف بر روی شبکه	۱۵
۹.۱	 نمای فرآیند کلی مدل ارائه شده	۱۸
۱۰.۱	 عملیات پیش‌پردازش بسته‌های ترافیک شبکه	۱۸
۱۱.۱	 عملیات شناسایی نشست‌ها و جریان‌ها در داده‌های ترافیک شبکه	۲۰
۱۲.۱	 عملیات محاسبه و استخراج ویژگی از بسته‌های ترافیک شبکه	۲۱
۱۳.۱	 نمونه سرآیند IP یک بسته	۲۱
۱۴.۱	 نمونه سرآیند TCP یک بسته	۲۲
۱.۲	 الگوریتم‌های یادگیری ماشین از نوع بانظارت	۳۰

۳۲	مدل ارائه شده در روش Loo [۱]		۲.۲
۳۳	مدل ارائه شده در روش Qin [۲]		۳.۲
۳۴	مراحل تشخیص تعداد بسته در روش Peng [۳]		۴.۲
۳۷	مراحل شناسایی ترافیک شبکه در روش Yamansavascular [۴]		۵.۲
۴۵	تحلیل رفتاری برنامه کاربردی		۱.۳
۴۶	شبکه Peer2Peer		۲.۳
۴۷	شبکه مشتری - خدمت گذار		۳.۳
۴۷	شبکه اسکایپ		۴.۳
۴۸	پروتکل TCP در اتصال مداوم و چند اتصالی		۵.۳
۴۹	شبکه تلگرام		۶.۳
۵۶	روابط بین انتروپی ها		۷.۳
۵۷	انتروپی متغیر تصادفی باینری		۸.۳
۵۹	محاسبه تاخیر انتها به انتها		۹.۳
۵۹	نحوه ایجاد تغییرات تاخیر در بسته های دریافتی در مقصد		۱۰.۳
۶۰	عوامل اتلاف بسته در شبکه های کامپیوتری		۱۱.۳
۶۳	معماری سیستم جهت شبیه سازی رخداد اتلاف بسته و ضبط داده ها		۱۲.۳
۶۴	نتایج بدست آمده از شناسایی برنامه های کاربردی مختلف از پایگاه داده با درصد های مختلف اتلاف بسته		۱۳.۳
۶۹	نمودار توزیع جمعیتی چهار ویژگی آماری مربوط به هفت برنامه کاربردی با نرخ های مختلف اتلاف بسته		۱۴.۳
۷۳	فراوانی تابع توزیع جمعیتی چهار ویژگی مربوط به بسته های ترافیک هفت برنامه کاربردی در ده بازه یکسان از ۰ تا ۱		۱۵.۳

۱.۴		توزیع ویژگی آماری اندازه سرآیندها در پایگاه داده UNIBS (ویرایش قدیمی)
۸۱		و پایگاه داده جمع آوری شده (ویرایش جدید) برای برنامه‌های کاربردی
۲.۴		نمودار فراوانی چهار ویژگی آماری استخراج شده برای دو گروه کاربردی
۸۲		Client-Server و P2P
۳.۴		نمودار توزیع تجمعی چهار ویژگی آماری استخراج شده برای دو گروه
۸۳		برنامه کاربردی P2P و Client-Server
۴.۴		میزان اطلاعات متقابل هر ویژگی و برنامه کاربردی تولید کننده آن در پایگاه داده UNIBS (r=Packet Flow duration, fl=Inter-Arrival Time, t=Size, s=Rate)
۸۴		Rate)
۵.۴		میزان اطلاعات متقابل هر ویژگی و برنامه کاربردی تولید کننده آن در پایگاه داده جمع آوری شده (s=Size, t=Inter-Arrival Time, fl=Flow duration, r=Packet Rate)
۸۴		r=Packet Rate)
۶.۴		مقادیر Precision، Recall و F-Measure بدست آمده از آزمایش دو مجموعه داده با استفاده از تکنیک‌های یادگیری ماشین
۸۶		مقایسه میانگین دقت شناسایی برنامه‌های کاربردی با سه تکنیک Chen [۵]، Peng [۳] و Qin [۲]
۸۶		مقایسه دقت شناسایی روش ارائه شده با روش‌های Qin و Chen
۸۹		مقایسه دقت شناسایی روش ارائه شده با روش‌های Qin و Chen
۹۹		سطوح مختلف ترافیک
۱۰۳		ساختار پیام در مدل هفت لایه شبکه‌های رایانه‌ای [۶]
۱۰۴		مسیریاب‌های بازگشایی پیام‌ها در مسیر یاب تا لایه شبکه (شکل 1.24 مرجع [۶])
۱۰۵		سرآیند بسته‌های پروتکل TCP
۱۰۶		شروع ارتباط در پروتکل TCP

فهرست جداول

۷		شماره درگاه‌های پروتکل‌های مختلف	۱.۱
۸		امضای پروتکل‌های مختلف	۲.۱
۲۲		پایگاه داده UNIBS	۳.۱
۲۳		پایگاه داده جمع‌آوری شده	۴.۱
۳۲		ویژگی‌های آماری مورد استفاده در روش Muehlstein [۷]	۱.۲
۳۳		ویژگی‌های آماری مورد استفاده در روش Loo [۱]	۲.۲
۳۵		نتایج الگوریتم‌های مورد استفاده در روش علی‌اکبریان [۸]	۳.۲
۳۶		ویژگی‌های آماری مورد استفاده در روش alshammari [۹]	۴.۲
۳۷		ویژگی‌های آماری مورد استفاده در روش Ertam [۱۰]	۵.۲
۳۸		کارایی روش‌های ارائه شده	۶.۲
۵۱		خصوصیات رفتاری برنامه‌های کاربردی Client-Server و P2P	۱.۳
۶۴		پایگاه داده جمع‌آوری شده با رخدادهای مختلف اتلاف بسته	۲.۳
۸۰		معیارهای ارزیابی روش ارائه شده	۱.۴
۸۷		مقایسه نتایج مابین روش ارائه شده و کارهای اخیر	۲.۴
۸۸		مشخصات شبکه عصبی مورد استفاده	۳.۴
۸۹		نتایج شناسایی هفت برنامه کاربردی با استفاده از ویژگی‌های مرتبه سوم	۴.۴
۹۰		ماتریس درهم‌ریختگی شناسایی هفت برنامه کاربردی به کمک روش ارائه شده	۵.۴

فهرست اختصارات

A

ABA	Application Behavior Analysis
ACK	Acknowledgement
API	Application Programming Interface
ARPA	Advanced Research Projects Agency
ARPANET	Advanced Research Projects Agency Network
ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode

B

BGP	Border Gateway Protocol
-----------	-------------------------

C

CDF	Cumulative Distribution Function
-----------	----------------------------------

D

DOS	Denial Of Service
DPI	Deep Packet Inspection

F

FTP	File Transfer Protocol
-----------	------------------------

I

IANA	Internet Assigned Numbers Authority
ICA	Independent Component Analysis
ID	Identification
i.i.d	Independent identically distributed
IP	Internet Protocol
ISP	Internet Service Provider

N

NAT	Network Address Translation
-----------	-----------------------------

P

P2P	Peer-to-Peer
PCA	Principal Component Analysis

Q

QoE	Quality of experience
QoS	Quality of Service

R

ROC	Receiver Operating Characteristics
-----------	------------------------------------

S

SPI	Stochastic Packet Inspection
SSL	Secure Sockets Layer

T

TCP	Transmission Control Protocol
-----------	-------------------------------

U

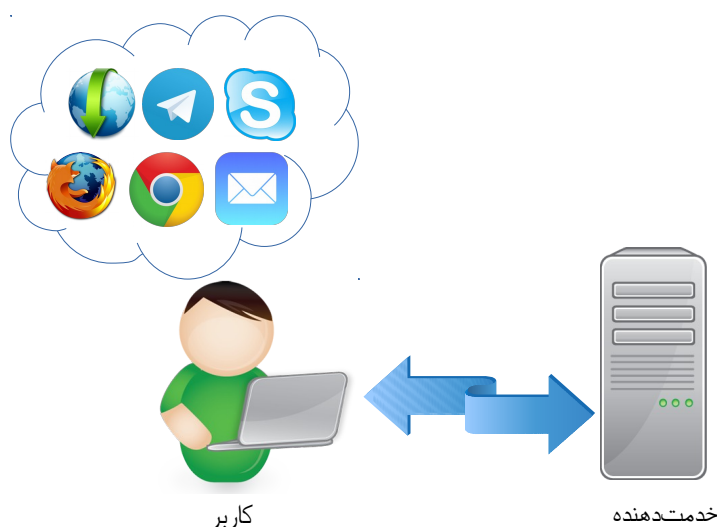
UDP..... User Datagram Protocol

فصل اول

ترافیک شبکه و شناسایی آن

۱.۱ مقدمه

کاربران شبکه جهانی اینترنت به طور پیوسته رو به افزایش است و هر روز شاهد ورود برنامه‌های کاربردی جدید و متنوع به ترافیک اینترنت هستیم. بازی‌های برخط، کاربردهای به اشتراک گذاری فایل، ویدئو کنفرانس‌ها، برنامه‌های چت، کار، تجارت و مکالمات بر روی شبکه IP¹ به مرور سهم بیشتری از ظرفیت خطوط ارتباطی شبکه اینترنت را به خود اختصاص می‌دهند (شکل ۱.۱). همچنین انواع ترافیک‌های مخرب و تهدیدات امنیتی در فضای سایبری رو به افزایش است که امنیت کاربران را به خطر می‌اندازد. از این رو مدیریت و امنیت اینترنت مشکلی جدی برای تمامی کشورها می‌باشد. از طرفی دیگر، تقاضای



شکل ۱.۱: اجرای برنامه‌های کاربردی مختلف بر روی رایانه کاربر شبکه

رو به رشد برای استفاده از پهنای باند بیشتر، از یک سو و محدودیت ظرفیت فیزیکی خطوط ارتباطی شبکه از سوی دیگر، سرویس دهندگان اینترنت را بر آن می‌دارد تا در پی یافتن راهکارهایی جهت بهبود کیفیت بهره برداری کاربران از منابع شبکه باشند. یکی از این راهکارها، تخصیص پهنای باند مشخص به هر یک از برنامه‌ها و سرویس‌های موجود در شبکه و اولویت‌دهی به آن‌ها در استفاده از ظرفیت خطوط ارتباطی و منابع پردازشی تجهیزات شبکه است. گام نخست در نیل به این هدف، شناسایی و طبقه‌بندی ترافیک شبکه است. شناسایی ترافیک، تکنیک ابتدایی می‌باشد که توسط ارائه دهندگان سرویس‌های اینترنتی، جهت مدیریت منابع شبکه و ضمانت امنیت شبکه مورد استفاده قرار می‌گیرد. در حقیقت جهت تحقق مدیریت شبکه و کنترل امنیت، مدیران شبکه بایستی هویت و وضعیت‌های برنامه‌های کاربردی را با مکانیزم‌های شناسایی ترافیک معلوم نمایند. همچنین با توجه به افزایش بدافزارها و تلاش آن‌ها برای

¹Internet Protocol

پنهان سازی ترافیک خود به منظور گریز از سیستم‌های تشخیص نفوذ و دورزدن دیوارهای آتش، شناسایی ترافیک به عنوان یک گام اولیه در تأمین امنیت شبکه در مقابل تهدیدات سایبری دارای اهمیت است [۱۱، ۱۲].

۲.۱ ساختار شبکه و تبادل داده در آن

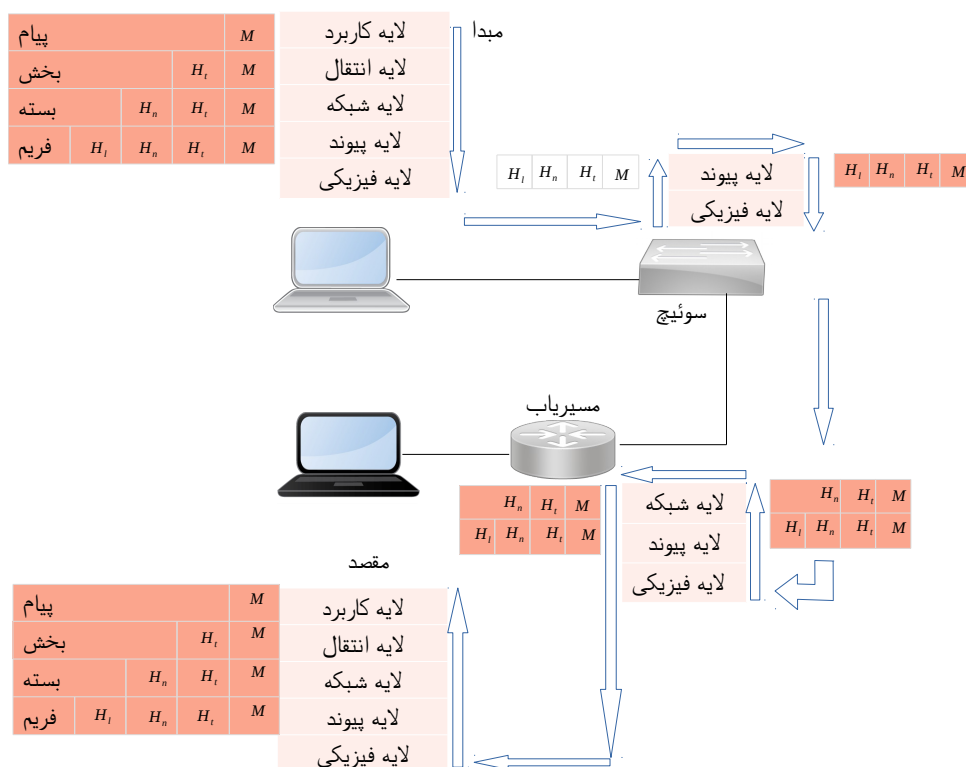
شبکه‌های کامپیوتری مجموعه‌ای از کامپیوترهای مستقل و متصل به یکدیگرند، که با یکدیگر ارتباط داشته و تبادل داده می‌کنند. همه شبکه‌ها از اجزای سخت‌افزاری پایه‌ای تشکیل شده‌اند تا گره‌های شبکه را به یکدیگر متصل کنند، مانند کارت‌های شبکه، تکرارکننده^۲، هاب^۳، پل^۴، سوئیچ^۵ و مسیریاب^۶‌ها. مسیریاب‌ها از قراردادهای مسیریابی مانند کوتاه‌ترین مسیر استفاده می‌کنند تا با یکدیگر ارتباط برقرار کرده و بهترین مسیر بین هر دو ایستگاه را پی‌ریزی کنند. ترافیک شبکه یا داده‌های ارسالی بر روی بستر شبکه، دارای سطوح مختلفی همچون جلسه^۷، جریان^۸، توده^۹ و بسته^{۱۰} می‌باشد. در این رساله ترافیک شبکه در سطح بسته‌ها مورد ارزیابی قرار می‌گیرد. (جهت دریافت اطلاعات بیشتر در مورد مفاهیم اولیه شبکه به پیوست [آ](#) مراجعه شود).

فرض کنید کاربر یک برنامه کاربردی را بر روی رایانه خود اجرا می‌کند. بر طبق مدل هفت لایه‌ای شبکه‌های رایانه‌ای [۶]، هر لایه بسته‌های لایه بالاتر را دریافت می‌کند، و پس از اعمال برخی پردازش‌ها بر روی آن، مقداری داده به عنوان سرآیند به آن اضافه کرده، و آن را به سوی لایه پایین‌تر رهسپار می‌کند. در نهایت در لایه فیزیکی^{۱۱} بسته نهایی به صورت امواجی به سوی گیرنده رهسپار می‌گردد. ذکر این نکته حائز اهمیت است که تحلیل‌گر ترافیک می‌تواند به شرط عدم رمزنگاری، بسته‌ها را تا لایه کاربرد بازگشایی کند و از محتوای آن آگاه شود. وظیفه اصلی مسیریاب تعیین مسیر بسته‌ها با استفاده از جدول ارسال می‌باشد. در شبکه نیز تعیین مسیر توسط آدرس IP صورت می‌گیرد. در نتیجه مسیر یاب‌های شبکه، کار خود را در لایه شبکه انجام داده و بسته‌ها تا لایه شبکه بازگشایی می‌شوند (شکل ۲.۱). این بدان معناست که، بررسی محتویات بسته تا لایه شبکه سربار چندانی نخواهد داشت. اما اگر تحلیل‌گر ترافیک، لایه‌های بالاتر را نیز بازگشایی کند، بایستی به جهت بازگشایی هر لایه و خواندن سرآیند و محتوای بسته‌ها، متحمل سربار محاسباتی و زمانی زیادی شود.

²Repeater
³Hub
⁴Bridge
⁵Switch
⁶Router

⁷Session
⁸Flow
⁹Burst
¹⁰Packet
¹¹Physical Layer

تمامی رویدادهای شبکه اعم از ارتباطات، ارسال و دریافت بسته‌ها را می‌توان با استفاده از تئوری صف



شکل ۲.۱: بازگشایی بسته‌ها در مسیر یاب تا لایه شبکه

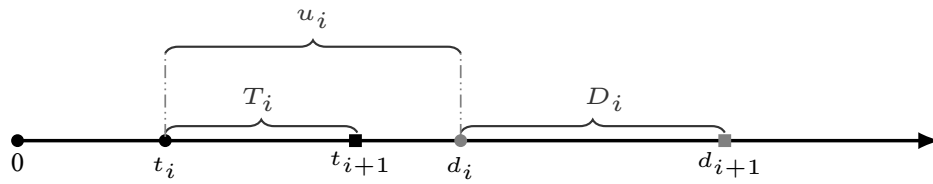
مدل نمود. دلیل مطرح نمودن این مباحث، تاثیر برخی از عناصر شبکه که رفتار آنها از قوانین صف تبعیت می‌کند بر روی ویژگی‌های آماری بسته‌ها می‌باشد. بدین‌سان ارتباط بین خروجی و ورودی سامانه‌های صف، نقش کلیدی در این مساله ایفاد خواهد نمود. از این‌رو مباحث مربوط به تئوری صف نقش مهمی در خصوص مدل کردن شبکه، ارسال و دریافت بسته‌ها و همچنین تاثیرات یک بسته در هنگام انتقال بر روی کانال دارد.

۳.۱ ساختار بسته‌های ترافیک شبکه

در حالت کلی ساختار داده‌های ترافیک شبکه که تحلیل‌گر ترافیک^{۱۲} از یک برنامه کاربردی در اختیار دارد، به دو گروه تقسیم‌بندی می‌شود.

- داده‌های مرتبط با سرآیند و محتوای بسته (در لایه‌های مختلف).
- داده‌های مرتبط با زمان تولید بسته و ویژگی‌های آماری آن.

¹²Traffic Analyser



شکل ۳.۱: پارامترهای t_i, u_i, d_i, T_i و D_i

۱.۳.۱ داده‌های مرتبط با سرآیند و محتوای بسته

داده‌های مرتبط با سرآیند و محتوای بسته، می‌تواند شامل محتویات بسته (در صورت عدم استفاده از رمزنگاری)، اندازه هر قسمت از بسته در لایه‌های مختلف و ویژگی‌های آماری هر بخش از بسته باشد. این دسته از داده‌ها، برخی مرتبط با برنامه کاربردی مورد نظر است، و برخی هیچ ارتباطی با برنامه کاربردی ندارد. به عنوان مثال، شناسه IP فرستنده از جمله داده‌هایی است که هیچ ارتباطی با برنامه کاربردی اجرا شده در سمت فرستنده ندارد. البته از داده‌های حاصل شده از بازگشایی لایه‌های مختلف یک بسته، تنها بخشی از آن‌ها مفید بوده و به نوعی به برنامه کاربردی تولید کننده آن وابسته می‌باشد.

۲.۳.۱ داده‌های مرتبط با زمان تولید بسته

در نظر بگیرید که یک برنامه کاربردی، بسته i ام خود را در لحظه t_i از لایه کاربرد^{۱۳}، رهسپار لایه ترابری^{۱۴} می‌کند. این بسته بعد از گذراندن u_i واحد زمانی، توسط تحلیل‌گر ترافیک در زمان $d_i = u_i + t_i$ دریافت می‌شود. u_i در حالت کلی یک متغیر تصادفی^{۱۵} است که تحت تاثیر عوامل بسیاری نظیر تاخیرهای ناشی از انتظار در صف مسیریاب‌ها و یا کارت‌های شبکه، مدت زمان پردازش بسته در گره‌های شبکه قرار می‌گیرد. زمان مابین تولید دو بسته متوالی (T_i) به صورت رابطه (۱.۱) تعریف می‌شود و زمان مابین دریافت دو بسته متوالی توسط تحلیل‌گر ترافیک، با پارامتر D_i نشان داده می‌شود، و به صورت رابطه (۲.۱) تعریف می‌شود.

$$T_i = t_{i+1} - t_i \quad (1.1)$$

$$D_i = d_{i+1} - d_i = T_i + u_{i+1} - u_i \quad (2.1)$$

شکل (۳.۱)، مفهوم پارامترهای یاد شده را به خوبی نشان می‌دهد. فرض کنید، $f(T_{j_1}, T_{j_2}, \dots, T_{j_n})$ به عنوان تابع چگالی توام مرتبه n ام بین متغیر تصادفی $T_{j_1}, T_{j_2}, \dots, T_{j_n}$ در نظر گرفته شود. لازم به ذکر

¹³Application Layer

¹⁴Transport Layer

¹⁵Random Variable

است که اگر متغیر تصادفی T_i از یک توزیع $i.i.d^{16}$ پیروی کند، با دانستن تابع توزیع مرتبه اول که با $f_T(t)$ نشان داده می‌شود، می‌توان توزیع تمامی مراتب را بدست آورد. ذکر این نکته ضروری است که اگر تابع توزیع زمان مابین تولید بسته‌ها در هر مرتبه‌ای موجود باشد، می‌توان تمامی اطلاعات آماری مربوط به یک برنامه کاربردی را استخراج نمود. به ندرت دو برنامه کاربردی را می‌توان یافت که تمامی اطلاعات آماری مشابه داشته باشند. لذا به نظر می‌رسد که اطلاعات آماری، احتمالاً یک امضای منحصر به فرد هر برنامه کاربردی خواهد بود [۱۳]. اما متأسفانه بدست آوردن امضای منحصر به فرد به این سادگی نخواهد بود. زیرا:

✗ به طور کلی با در نظر گرفتن رابطه (۲.۱) و داشتن $f(T_{j1}, T_{j2}, \dots, T_{jn})$ نمی‌توان توزیع $g(D_{j1}, D_{j2}, \dots, D_{jn})$ را بدست آورد. تحلیل گر نیاز به دانستن تابع توزیع u_i در هر مرتبه‌ای و به ازای هر i ای دارد.

✗ اگر T_i دارای توزیع $i.i.d$ نباشد، استخراج تمامی توابع توزیع در تمامی مراتب (تا بی‌نهایت مرتبه) به دلیل محدودیت پردازشی، امری غیر ممکن است. حتی اگر T_i دارای توزیع $i.i.d$ باشد، لزوماً D_i دارای توزیع $i.i.d$ نخواهد بود [۱۴، ۱۵].

✗ در شبکه اینترنت برای این که دریابیم که بسته‌ها با موفقیت به مقصد رسیده‌اند یا خیر از سازوکار ARQ^{17} ، استفاده می‌شود. ARQ موجب می‌شود که زمان ارسال بسته‌ها از سمت فرستنده تابعی از زمان دریافت ACK^{18} بسته‌های قبلی باشد. بدین سان توزیع زمان مابین خروج 19 بسته‌ها نه تنها به نوع برنامه کاربردی، بلکه به تمامی ساختار شبکه‌ی موجود بین فرستنده و گیرنده و میزان ترافیک شبکه نیز وابسته خواهد بود.

۴.۱ روش‌های شناسایی ترافیک شبکه

شناسایی جریان‌های جاری در شبکه بر اساس نوع برنامه کاربردی تولیدکننده آنها را شناسایی ترافیک شبکه گویند. این کار از طریق تحلیل ترافیک شبکه انجام شده و هدف آن به منظور ارائه سرویسی متناسب با نیازهای کاربران می‌باشد. برای طبقه‌بندی جریان‌های ترافیک شبکه بر اساس پروتکل تولیدکننده آنها، تکنیک‌های مختلفی وجود دارد. تاکنون روش‌های بسیاری برای شناسایی ترافیک 20 شبکه ارائه شده است [۱۶، ۱۷، ۱۸، ۱۹، ۲۰، ۱۱]. این روش‌ها را می‌توان در چهار دسته کلی زیر تقسیم‌بندی نمود [۲۱].

¹⁶independent identically distributed

¹⁷Automatic Repeat Request

¹⁸Acknowledgement

¹⁹InterDeparture Time

²⁰Traffic Identification

- روش‌های شناسایی بر پایه شماره درگاه
- روش‌های شناسایی بر پایه محتوا
- روش‌های شناسایی بر پایه رفتار کاربر
- روش‌های شناسایی بر پایه ویژگی‌های آماری

۱.۴.۱ روش‌های شناسایی بر پایه شماره درگاه

این روش شناسایی از این ایده منتج می‌شود که برنامه‌های کاربردی مختلف از پروتکل‌ها و شماره درگاه^{۲۱}‌های متفاوتی استفاده می‌کنند [۱۲]. جدول (۱.۱) نمونه‌ای از شماره درگاه‌های پروتکل‌های مختلف را نشان می‌دهد. روش کار به این صورت است که به هنگام عبور بسته‌ها از مسیریاب، قسمت سرآیند بسته

جدول ۱.۱: شماره درگاه‌های پروتکل‌های مختلف

شماره درگاه	برنامه کاربردی
80,8080	HTTP
443	HTTPS
25	SMTP
53	DNS
1080	SOCKS
113	IRC
21	FTP
110	POP3
6346	LimeWire

بررسی شده و با توجه به شماره درگاه‌های شناخته شده برنامه‌های کاربردی مختلف شناسایی می‌شوند. معایب این روش‌ها عبارتند از:

- بسیاری از برنامه‌های کاربردی شماره درگاه ثبت شده در IANA^{۲۲} ندارند. (مانند اکثر برنامه‌های نظیر به نظیر [۲۲])
- برخی برای دور زدن محدودیت‌ها از شماره دیگری استفاده می‌کنند (مثلاً HTTP با شماره درگاه غیر از ۸۰ در لینوکس)
- برخی از شماره‌های درگاه به صورت پویا انتخاب می‌شوند.
- رمزنگاری لایه IP اساساً مشاهده و بررسی شماره درگاه توسط تحلیل‌گر شبکه را ناممکن می‌نماید.

^{۲۱}Port Number

^{۲۲}Internet Assigned Numbers Authority

به طور نمونه آقای مور و همکارانش نشان داده‌اند که حداکثر ۷۰ درصد ترافیک شبکه توسط این روش قابل شناسایی می‌باشند و میزان دقت شناسایی توسط این روش به ۳۰ تا ۷۰ درصد کاهش یافته و در صورت اجرا در سطح بایت میزان دقت به ۳۰ درصد می‌رسد [۲۳، ۲۴].

۲.۴.۱ روش‌های شناسایی بر پایه محتوای بسته

در این شیوه محتوای بسته‌های شبکه مورد بازرسی قرار می‌گیرد. این روش شناسایی، به نوبه خود به دو زیردسته DPI²³ و SPI²⁴ دسته‌بندی می‌شود.

• در DPI، سعی می‌شود که محتوا^{۲۵} با یک امضای ثابت مقایسه گردد. شناسایی بر مبنای پروتکل و شماره درگاه، زیرمجموعه DPI محسوب می‌گردد. DPI به صورت گسترده در نرم‌افزارها و دیوارهای آتش^{۲۶} مورد استفاده قرار می‌گیرد [۲۵]. پیچیدگی بالا و عدم دستیابی به محتوای رمز شده برخی لایه‌ها، از بزرگترین اشکالات این روش است. جدول (۲.۱) نمونه‌ای از امضاهای پروتکل‌های مختلف را نشان می‌دهد.

• در SPI، ویژگی‌های آماری محتوا در نظر گرفته می‌شود. به عنوان مثال، آقای فینامور و همکاران یک روش شناسایی ترافیک بر مبنای خواص آماری تعدادی از بیت‌های اول بسته UDP²⁷ ارائه داده‌اند [۲۶].

جدول ۲.۱: امضای پروتکل‌های مختلف

برنامه کاربردی	امضای دیجیتال
HTTP	"http","Get","X09"
SMTP	"user","smtp","pass"
BitTorrent	0x13bittorrent porotocol
eDonkey	"0xe3","0xc5"
FTP	"ftp","x0d","220"

۳.۴.۱ روش‌های شناسایی بر پایه رفتار کاربر

دید دسته‌بندی روش‌های مبتنی بر رفتار^{۲۸} بسیار گسترده‌تر از دیگر روش‌ها است. در این شیوه، سعی می‌شود که در یک بازه زمانی بزرگ تمامی ترافیک رسیده از یک کاربر یا یک برنامه کاربردی مورد بررسی قرار گیرد و الگوی کلی ترافیک آن استخراج شود و در صورتی اجازه عبور بسته‌ها صادر می‌شود که رفتار

²³Deep Packet Inspection

²⁴Stochastic Packet Inspection

²⁵Payload

²⁶Firewall

²⁷User Datagram Protocol

²⁸Behavior

این بسته‌ها شبیه به یکی از رفتارهای موجود در سامانه باشد. برخی از کارهای تحقیقاتی در این زمینه در مراجع [۲۷، ۲۸] قابل مشاهده می‌باشند.

۴.۴.۱ روش‌های شناسایی بر پایه ویژگی‌های آماری

در این شیوه به ویژگی‌های آماری زمان و اندازه بسته‌های دریافتی، پرداخته می‌شود. مهم‌ترین گام در این قسمت، تشخیص تمایزهای آمار^{۲۹} بین برنامه‌های کاربردی مختلف است. مقالات بسیاری نیز در این حوزه مطرح شده است [۲۹، ۳۰، ۱۸]. در سال ۲۰۰۵ آقای مور و همکارانشان برای جریان بسته‌های ترافیک شبکه ۲۴۸ ویژگی آماری استخراج کرده و ارائه نموده‌اند [۳۱]. به عنوان نمونه می‌توان به کمترین، بیشترین و میانگین اندازه بسته‌ها اشاره نمود. شکل (۴.۱) توزیع اندازه بسته‌های شش برنامه کاربردی در دوره‌ها و ساختارهای متفاوت شبکه را نشان می‌دهد. همانطور که در این شکل‌ها مشاهده می‌شود، توزیع اندازه بسته‌ها برای هر برنامه کاربردی در دوره‌های متفاوت به یکدیگر شبیه بوده و در برنامه‌های کاربردی مختلف، متفاوت می‌باشد. از اینرو توزیع اندازه بسته‌ها را می‌توان به عنوان یکی از ویژگی‌های شناسایی کننده انتخاب نمود.

لازم به ذکر است که وجه تمایز اصلی بین دسته‌بندی آماری و SPI این است که در دسته‌بندی آماری ویژگی‌های آماری زمان تولید بسته‌ها و طول آن‌ها، بدون در نظر گرفتن محتوا، مورد بررسی قرار می‌گیرد؛ بدین سان این روش‌ها از لحاظ محاسباتی بسیار سبک هستند، چراکه دیگر نیازی به بازگشایی محتوای بسته وجود ندارد.

در این نوع از تحلیل، علم نظریه صف^{۳۰} نقش بسیار برجسته‌ای دارد، چرا که بسیاری از اتفاقاتی که در مسیر ارسال یک بسته تا رسیدن آن به مقصد رخ می‌دهد را می‌توان با یک صف مدل نمود. برخی از قوانین حاکم بر صف در این مورد می‌تواند مفید واقع شود، به عنوان مثال قانون پایستگی جریان^{۳۱}، که بر طبق آن در یک سامانه کلی نرخ ورود بسته‌ها به صف برابر با نرخ خروج از صف است [۳۲، قضیه ۵.۴.۳].

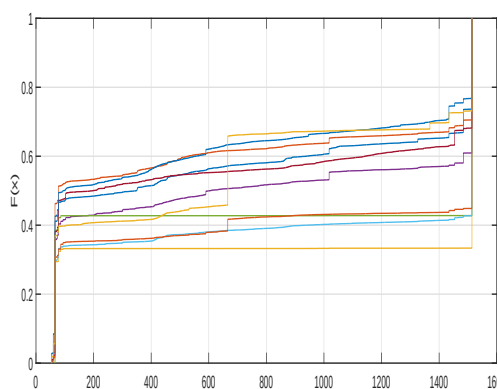
۵.۱ چالش‌های پیش‌رو

شناسایی ترافیک شبکه یکی از تکنیک‌های اصلی جهت مدیریت منابع شبکه و ارتقاء امنیت شبکه می‌باشد. در حقیقت جهت تحقق مدیریت شبکه و کنترل امنیت، مدیران شبکه بایستی هویت و وضعیت‌های برنامه‌های کاربردی را توسط شناسایی ترافیک معلوم نمایند. عواملی در روند توسعه شبکه‌های رایانه‌ای به

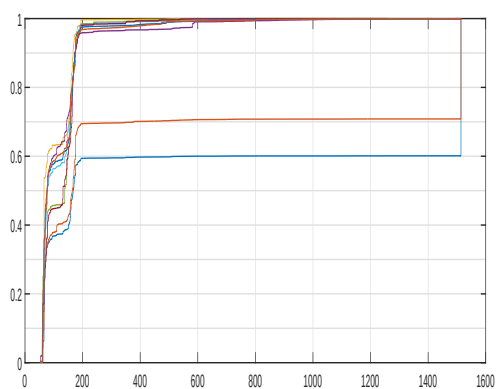
²⁹Statistical

³⁰Queueing Theory

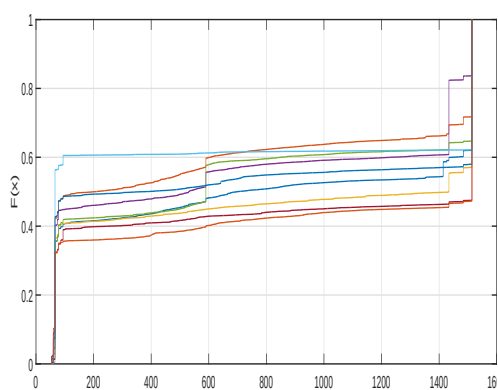
³¹Flow Conservation Law



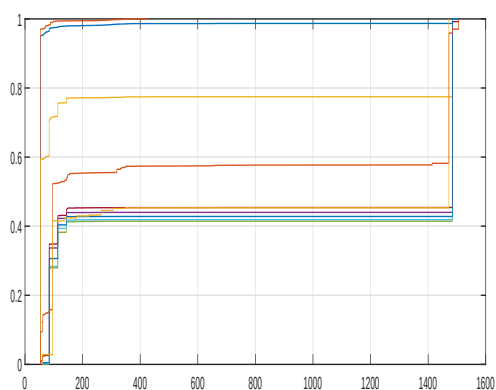
(ب)



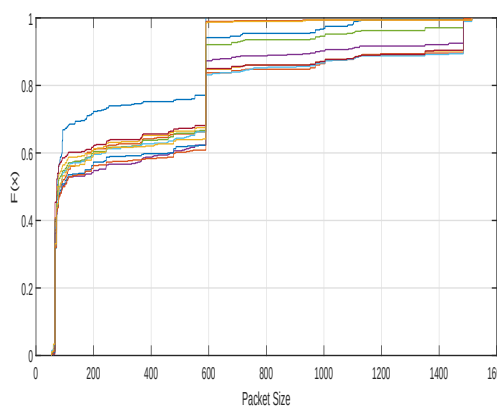
(آ)



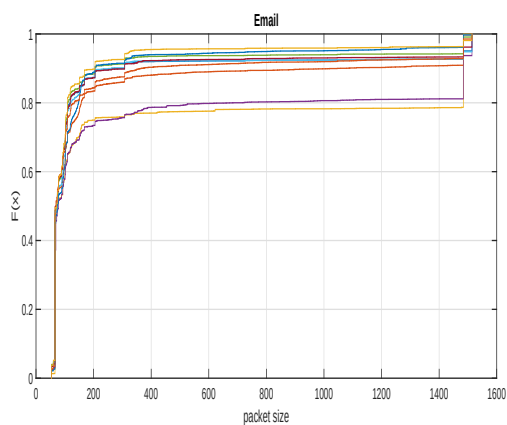
(د)



(ج)



(و)



(ه)

شکل ۴.۱: توزیع اندازه بسته‌های شش برنامه کاربردی در دوره‌ها و ساختارهای مختلف. (آ. Skype)، (ب. Safari)، (ج. Torrent)، (د. Firefox)، (ه. Mail)، (و. Thunderbird).

خصوص اینترنت، موجب چالش برانگیزتر شدن مبحث شناسایی ترافیک شده است. برخی از این عوامل به شرح زیر است:

● به روزرسانی برنامه کاربردی

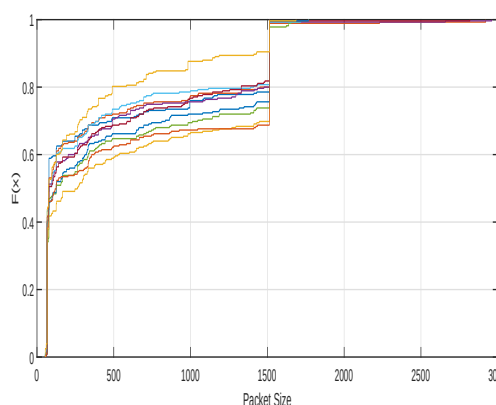
یکی از چالش‌های اصلی در حوزه روش‌های شناسایی ترافیک شبکه بروزرسانی برنامه‌های کاربردی، توسط شرکت ارائه دهنده این محصولات می‌باشد. این بروزرسانی‌ها به دلیل مواردی همچون: رفع خطا، افزایش سرعت سرویس‌دهی و تغییر ساختار صورت می‌گیرد. طبق بررسی‌های صورت گرفته، رفتار بعضی از ویژگی‌های آماری ارائه شده در مقالات در هنگام بروزرسانی برنامه‌های کاربردی تغییر کرده و به موجب آن عملیات شناسایی برنامه کاربردی با مشکل مواجه می‌شود. به طور نمونه شکل (۵.۱) توزیع اندازه سرآیندهای برنامه کاربردی Skype را در دوره‌ها و ساختارهای مختلف برای نمونه‌های سال‌های ۲۰۰۹ و ۲۰۱۶ نشان می‌دهد. همانطور که در شکل (۵.۱) مشاهده می‌شود، توزیع اندازه سرآیندها که به عنوان ویژگی آماری در بسیاری از مقالات معرفی شده است، در طی زمان و به روزرسانی برنامه کاربردی تغییر کرده است. همانطور که در شکل (۵.۱) مشاهده می‌شود، توزیع اندازه سرآیندهای ارسال شده در برنامه کاربردی Skype تا حدودی از توزیع نمایی پیروی می‌کند. به همین دلیل با تغییر توزیع اندازه سرآیند در دو ویرایش مختلف، پارامترهای تاثیر گذار در توزیع نمایی تغییر نموده و همچنین فراوانی اندازه سرآیندها نیز تغییر کرده است. در نتیجه اندازه سرآیندها را نمی‌توان به عنوان ویژگی اصلی تاثیر گذار جهت تمایز بین برنامه‌های کاربردی انتخاب نمود.

● یکی دیگر از مباحث چالش برانگیز که کمتر در کارهای گذشته به آن پرداخته شده است، تاثیر رخدادهای شبکه بر روی ویژگی‌های آماری می‌باشد. از جمله این رخدادهای می‌توان به موارد زیر اشاره نمود:

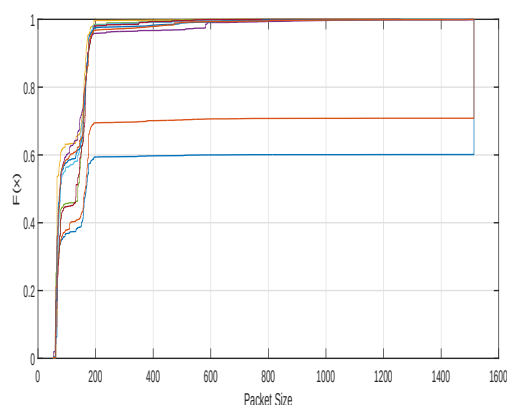
- کنترل ازدحام^{۳۲}: هنگامی که به بخشی از زیر شبکه تعداد بسیار زیادی بسته تحویل داده شود، کارایی آن کاهش یافته و در مواردی بسته‌ها از دست می‌روند. به این وضعیت ازدحام^{۳۳} گفته می‌شود. هنگام افزایش بی‌رویه ترافیک، مسیریاب‌ها قادر به انجام پردازش‌های لازم نبوده و بسته‌ها به دلایل مختلفی از جمله عدم حافظه کافی و پهنای باند کم شروع به از دست رفتن کرده و گاهی هیچ بسته‌ای به مقصد نمی‌رسد. در نتیجه یکی از مسائل تاثیر گذار بر روی ویژگی‌های آماری از جمله زمان دریافت بسته‌ها و همچنین ترتیب دریافت بسته‌ها وجود

³²Congestion Control

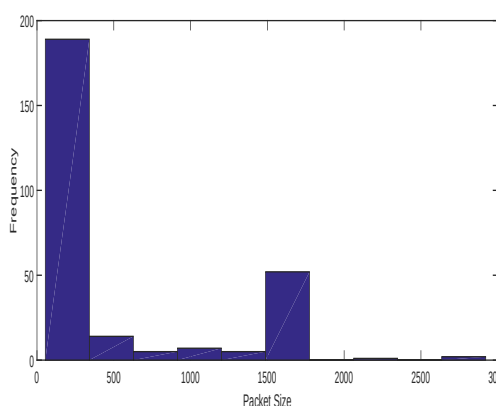
³³Congestion



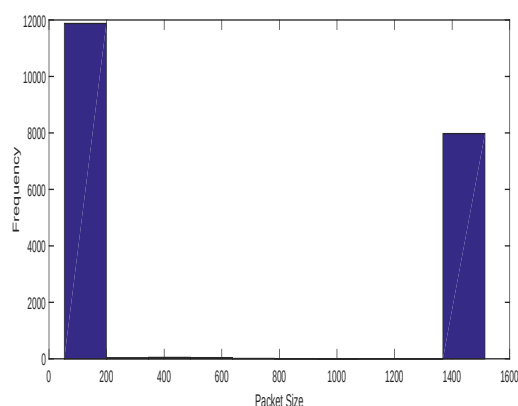
(ب)



(ā)



(د)



(ج)

شکل ۵.۱: توزیع و فراوانی اندازه سرآیندها برنامه کاربردی Skype در دوره‌ها و ساختارهای مختلف. (آ، ج. سال ۲۰۰۹)، (ب، د. سال ۲۰۱۶)،

ازدحام می‌باشد.

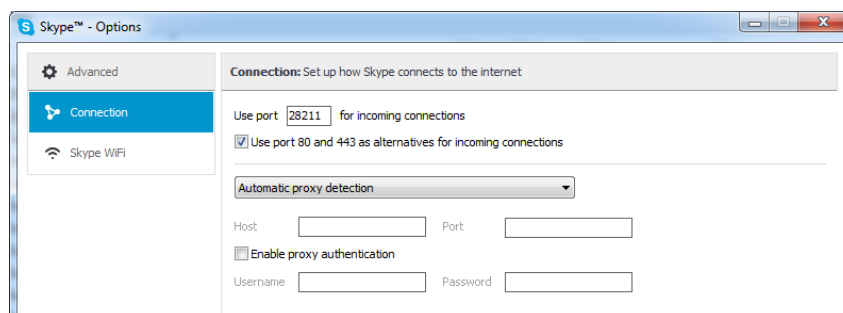
- کنترل جریان^{۳۴}: گاهی اوقات سرعت ارسال بسته‌ها توسط فرستنده موجب سرریزی بافر گیرنده می‌شود. در نتیجه گیرنده نمی‌تواند به صورت مداوم تایید دریافت بسته‌ها را به فرستنده ارسال کند، در نتیجه پروتکل TCP مجبور به کاهش سرعت می‌شود. در این موارد نیز در صورت عدم کنترل جریان، میزان سرعت ارسال بسته‌ها از طرف فرستنده می‌تواند ویژگی‌های آماری بسته‌ها را دستخوش تغییر قرار دهد [۳۳].

• با اعمال رمزنگاری^{۳۵} بر پیام‌های لایه کاربرد، شماره درگاه تنها شناسه باقیمانده از برنامه کاربردی است که در لایه ترابری قرار دارد. اما امروزه برخی از برنامه‌های کاربردی از شماره درگاه‌های پویا استفاده می‌کنند [۱۶]، که عملاً این روش تشخیص را نیز با مشکل مواجه می‌سازد. به این قابلیت

³⁴Flow Control

³⁵Encryption

اصطلاحاً شماره درگاه پویا^{۳۶} و یا شماره درگاه محرمانه^{۳۷} گفته می‌شود. به عنوان مثال Skype، از این قابلیت پشتیبانی می‌کند، و می‌تواند به صورت پویا شماره درگاه را در بازه ۸۰ تا ۴۴۳ تغییر دهد. شکل (۶.۱) نمایشی از این قابلیت در Skype Ver 6.11 را نشان می‌دهد.



شکل ۶.۱: نمایشی از قابلیت تنظیم شماره درگاه پویا در Skype Ver 6.11

- وجود داده‌های زائد

یکی از مشکلاتی که در مبحث شناسایی ترافیک شبکه با آن سروکار داریم، وجود بسته‌های زائد می‌باشد. منظور از بسته‌های زائد بسته‌های مربوط به سیستم عامل، دیوار آتش و اجزاء شبکه می‌باشد، که گاهی اوقات تعداد این بسته‌ها بسیار زیاد می‌شود. دلیل را می‌توان به نیاز بروزرسانی متناوب اجزاء شبکه در خصوص اجزاء دیگر نسبت داد. همچنین از دیگر موارد بسته‌های زائد می‌توان به بسته‌هایی اشاره نمود که ارسال آنها به هر دلیلی ناقص بوده و یا با موفقیت به مقصد نرسیده‌اند. در نتیجه گاهی اوقات مشاهده می‌شود که یک بسته چندین بار به صورت پی‌درپی ارسال می‌گردد.

- بالا رفتن نرخ تبادل اطلاعات در شبکه

این امر موجب می‌شود که تحلیل‌گر ترافیک مجبور باشد، که حجم بسیار زیادی از اطلاعات را در فرصتی محدود، پردازش کند. حتی شاید تحلیل‌گر ترافیک نتواند تمامی بسته‌های داده را پردازش کند، و به ناچار به روش‌های نمونه‌برداری و کاهش حجم داده روی آورد [۳۴].

- حفظ حریم خصوصی^{۳۸}

یکی دیگر از مسائل اصلی در شناسایی ترافیک شبکه، حفظ حریم خصوصی کاربران می‌باشد. این بدان معناست که در هنگام شناسایی ترافیک شبکه و برنامه‌های کاربردی بر عکس کاری که دیوارهای آتش و مانند آن انجام می‌دهند، محتوای بسته‌های ارسالی توسط برنامه‌های کاربردی

³⁶Dynamic Port Number

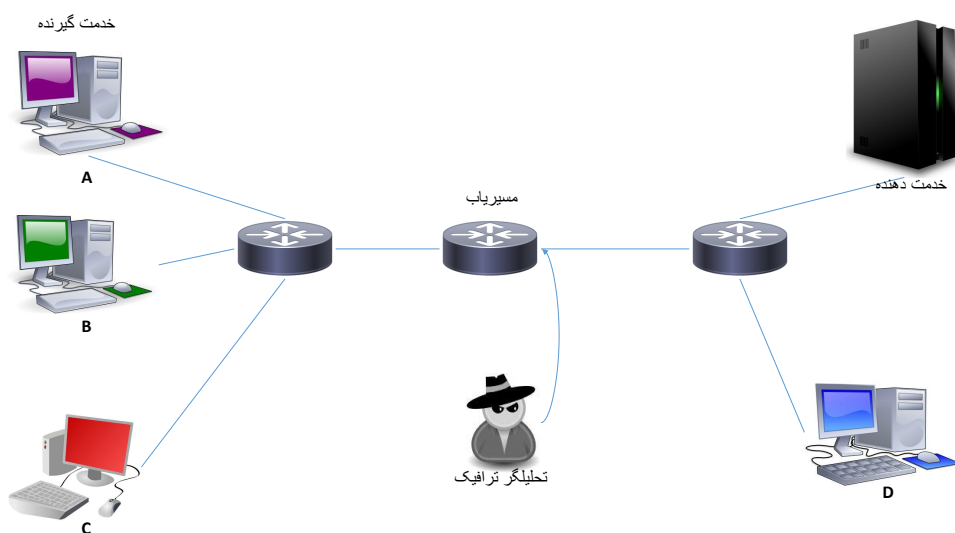
³⁷Private Port Number

³⁸Privacy

مختلف بازگشایی نشده و عملیات شناسایی بدون بررسی محتوا انجام گیرد [۳۵].

۶.۱ تعریف مساله

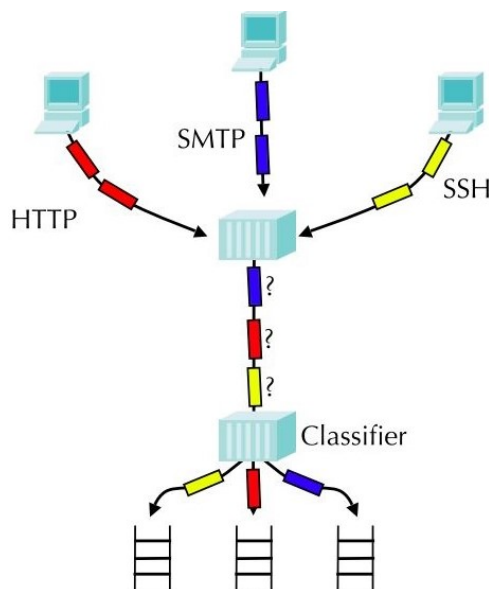
در نظر بگیرید که سه کاربر به نام‌های A، B و C، به شبکه اینترنت متصل هستند (شکل ۷.۱) و اطلاعات آن‌ها از طریق مسیریاب‌های مختلف به مقصد می‌رسد. ممکن است هر کدام از سه کاربر برنامه‌های کاربردی مختلفی را بر روی رایانه و یا تلفن همراه خود اجرا نمایند. بسته‌های تولید شده توسط این برنامه‌های کاربردی بر روی بستر شبکه با توجه به مسائل موجود در پروتکل TCP و همچنین ظرفیت پهنای باند و بافرهای گیرنده ارسال می‌شوند. همانطور که در بخش قبل مطرح گردید، این بسته‌ها شامل اطلاعات متعددی مربوط به خود داده، سرآیند داده و ویژگی‌های آماری مانند اندازه بسته و مدت زمان مابین ارسال بسته‌ها می‌باشند. شکل (۸.۱) نمایشی از ارسال این بسته‌ها را نمایش می‌دهد. تحلیلگر



شکل ۷.۱: اجرای برنامه‌های کاربردی مختلف بر روی شبکه

ترافیک در مسیر اتصال دو مسیریاب واقع شده و شروع به ضبط داده‌های این اتصال می‌کند. به این ترتیب با استفاده از روش‌های مطرح شده، قصد شناسایی برنامه‌های کاربردی در حال اجرا را دارد. بایستی دقت داشت که اطلاعات محدود به محتوای بسته‌ها نبوده و هر نوع اطلاعاتی که به نوعی به برنامه کاربردی در حال اجرا مرتبط است، می‌تواند برای تحلیل‌گر شبکه مفید باشد. به عنوان نمونه می‌توان به نوع پروتکل برنامه کاربردی اشاره نمود.

در فرآیند شناسایی ترافیک شبکه، تحلیل‌گر ترافیک بایستی بتواند نگاهی مناسب بین ترافیک‌ها و برنامه‌های کاربردی برقرار کند. در این پژوهش عملیات شناسایی برنامه‌های کاربردی با استفاده از



شکل ۸.۱: ارسال بسته‌های برنامه‌های کاربردی مختلف بر روی شبکه

روش‌های مبتنی بر ویژگی‌های آماری انجام می‌شود. بدین‌سان تمرکز ما بر روی تمایز ترافیک برنامه‌های کاربردی یک کاربر و نگاشت بین ترافیک‌ها و برنامه‌های کاربردی است. همچنین در این پژوهش جریان‌ها به صورت دو طرفه در نظر گرفته می‌شود. به همین منظور ویژگی‌هایی مقاوم و موثر در مقابل چالش‌های مطرح شده (بروزرسانی برنامه‌های کاربردی و رخداد اتلاف بسته‌ها در شبکه بین فرستنده و گیرنده) جهت شناسایی ترافیک برنامه‌های کاربردی ارائه می‌شود.

۷.۱ ضرورت توسعه روش شناسایی ترافیک شبکه

با رشد روزافزون استفاده از اینترنت و همچنین بالارفتن ترافیک شبکه ناشی از برنامه‌های کاربردی P2P³⁹ همچون Skype, Imo, Instagram, Telegram, Viber، نیازمندی‌هایی نیز در جهت مدیریت پهنای باند، امنیت شبکه و مواردی جهت سیاست‌گذاری‌های مختلف بوجود آمده‌اند. بر این اساس توسعه سامانه‌های شناسایی ترافیک شبکه امری ضروری می‌باشد. بعضی از این موارد در ادامه مطرح می‌گردد

📖 شناسایی حملات برپایه برنامه‌های کاربردی در مسیریاب [۳۶، ۳۷]: مسیریاب‌ها یکی از اجزای اصلی هر شبکه‌ای بوده و به دلیل آنکه کل ترافیک شبکه از آنها عبور می‌کند، همواره مورد توجه هکرها و سارقان اینترنتی قرار دارد. همچنین گاهی اوقات لازم است تا برنامه‌های کاربردی مخرب قبل ورود به شبکه داخلی شناسایی و در مورد نحوه برخورد با آنها تصمیم‌گیری شود.

📖 فیلترکردن برنامه‌های خاص توسط مدیر شبکه [۳۸]: گاهی اوقات سیاست کلی یک کشور، یک

³⁹Peer-to-Peer

سازمان یا شرکت حکم می‌کند که یک برنامه کاربردی خاص در شبکه‌ها فیلتر گردد. به همین دلیل لازم است تا در ابتدا بتوان برنامه‌های کاربردی در حال اجرا در بستر شبکه را طبقه‌بندی و شناسایی نمود تا در مراحل بعد امکان فیلتر کردن آن وجود داشته باشد.

🔖 رهگیری قانونی^{۴۰}: در مبحث شبکه‌های کامپیوتری و برنامه‌های کاربردی در حال اجرا بر روی بستر شبکه بخشی به نام رهگیری قانونی یا شنود قانونی وجود دارد که توسط آن اعمال نظارت قانونی و کنترل امنیت شبکه توسط مراجع قانونی انجام می‌شود. این امکانی است که تمام شبکه‌های مخابرات در سراسر جهان از آن استفاده می‌کنند. حال این که چه اطلاعاتی ضبط و نگهداری شوند یا چه کسانی ردیابی شوند، وابسته به سیاست گذاری سازمان مربوطه است.

🔖 تشخیص ناهنجاری^{۴۱}: شناسایی حملات در شبکه‌های کامپیوتری از دیدگاه اطلاعات مورد استفاده در مرحله یادگیری، به دو دسته تشخیص نفوذ و تشخیص ناهنجاری قابل تقسیم است. در تشخیص ناهنجاری، فعالیت‌های یک برنامه کاربردی با یک الگوی مشخص و سیاست گذاری شده توسط مدیر سیستم یا خود سیستم مقایسه می‌شود و در صورت وجود اختلاف، فعالیت آن برنامه کاربردی به عنوان ناهنجاری شناسایی می‌شود. در این روش پروفایلی برای هر برنامه کاربردی ایجاد می‌شود و در آن سطح توانائی‌ها و دسترسی مشخص می‌شود و عبور از این سطوح به معنای ناهنجاری در نظر گرفته می‌شود. تمامی این روش‌ها در صورتی امکان پذیر است که بتوان ترافیک برنامه‌های کاربردی را شناسایی نمود.

🔖 بالابردن کارایی^{۴۲} شبکه و شیوه‌های تخصیص منابع^{۴۳} به منظور بالابردن QoS^{۴۴} و QoE^{۴۵}: با کنترل منابع شبکه می‌توان با در نظر گرفتن رفتار مشتریان و اختصاص پهنای باند متفاوت برای آنها، نقشی جدی در انجام دقیق و با کیفیت تراکنش‌ها ایفا کرد. جهت کنترل کیفیت سرویس‌دهی، ابتدا بایستی برنامه‌های کاربردی شناسایی شود تا براساس الویت‌بندی تخصیص پهنای باند مناسب صورت گیرد.

🔖 محاسبه هزینه^{۴۶} کاربران: طبقه‌بندی ترافیک میزان استفاده کاربران از برنامه‌های کاربردی گوناگون و بالتبع منابع شبکه را معلوم می‌سازد. از این طریق می‌توان هزینه کاربران به خاطر استفاده از منابع و یا حتی در پاره‌ای از موارد نقض قوانین توسط کاربران را محاسبه نمود.

⁴⁰Lawful Interception

⁴¹Anomaly Detection

⁴²Performance

⁴³Resource Allocation

⁴⁴Quality of Service

⁴⁵Quality of experience

⁴⁶Charging

۸.۱ فرض‌های مسئله

شکل (۷.۱) نمایی ساده شده از مدل ساختار شبکه می‌باشد که در کل این نوشتار آن را مدنظر قرار خواهیم داد. البته در این میان فرض‌هایی را بایستی در نظر بگیریم. این فرض‌ها به شرح زیر خواهند بود.

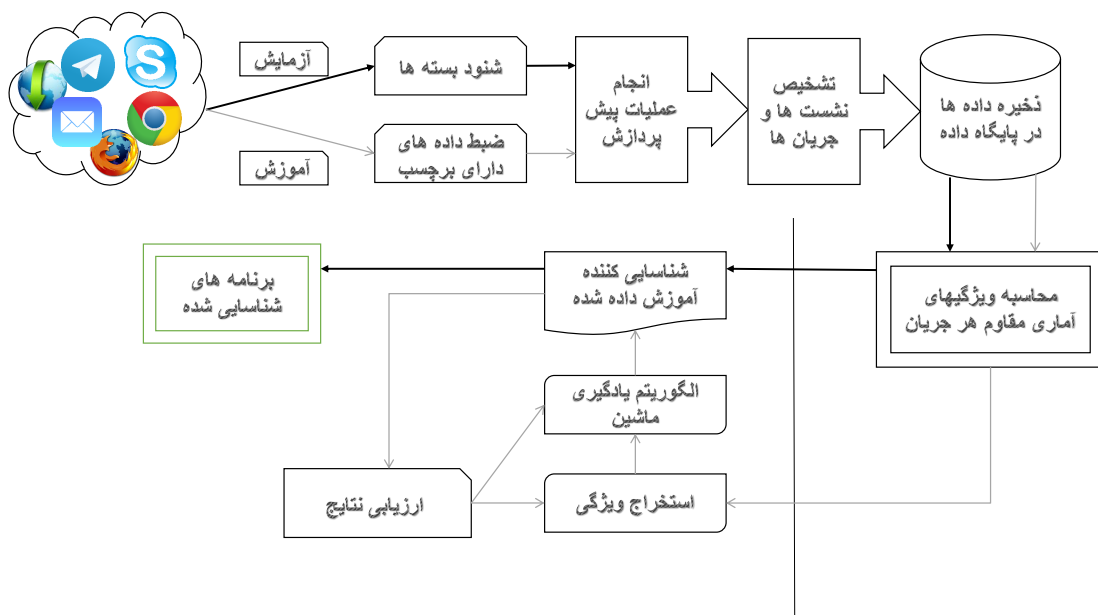
- 🔖 تحلیل‌گر شبکه بر روی مسیر یاب شبکه قرار داشته و از آنجا شروع به ضبط ترافیک شبکه می‌نماید.
- 🔖 نقطه تمرکز ما در این رساله، بر روی نگاشت بین ترافیک و برنامه کاربردی خواهد بود، و دو مرحله شناسایی کاربر و تفکیک ترافیک کاربر را انجام شده فرض می‌کنیم.
- 🔖 تحلیل‌گر ترافیک فقط به اطلاعات آماری بسته‌ها دسترسی داشته و به محتوای داده بسته‌ها هیچگونه دسترسی ندارد (حفظ حریم خصوصی).
- 🔖 بر روی مسیر یاب‌های شبکه، بسته‌ها تا لایه شبکه بازگشایی می‌شوند.
- 🔖 در این رساله هفت برنامه کاربردی IDM, Telegram, Skype, IE, Chrome, Firefox و Twitter مورد تحلیل و بررسی قرار می‌گیرند.
- 🔖 در این پژوهش جریان‌ها به صورت دو طرفه در نظر گرفته می‌شود.

۹.۱ مدل روش ارائه شده

روش ارائه شده شناسایی ترافیک برنامه‌های کاربردی بصورت شکل (۹.۱) مدل شده است. این مدل همانند تمامی روش‌های طبقه‌کننده نظارت شده، دارای دو فاز آموزش و آزمایش می‌باشد. دو مرحله اصلی در این مدل انجام عملیات پیش‌پردازش و استخراج ویژگی می‌باشد. در مرحله پیش‌پردازش داده‌های زائد حذف شده و جریان‌ها شناسایی می‌شوند و در مرحله استخراج ویژگی، ویژگی‌هایی مقاوم در برابر بروزرسانی برنامه‌های کاربردی و رخداد ائتلاف بسته در شبکه استخراج می‌گردد. لازم به ذکر است که در این رساله هدف ارائه روش جدیدی از الگوریتم‌های یادگیری ماشین نبوده و صرفاً از تکنیک‌های موجود استفاده شده است.

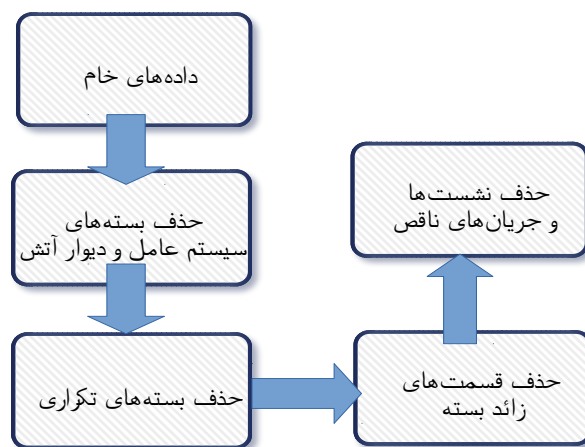
۱.۹.۱ عملیات پیش‌پردازش

یکی از مشکلات موجود در هنگام ضبط بسته‌ها، وجود بسته‌های زائد می‌باشد. منظور از بسته‌های زائد بسته‌های مربوط به سیستم عامل، دیوار آتش و اجزاء شبکه می‌باشد که گاهی اوقات تعداد این بسته‌ها بسیار زیاد می‌شود. دلیل این امر را می‌توان نیاز بروزرسانی متناوب اجزاء شبکه در خصوص اجزاء دیگر



شکل ۹.۱: نمای فرآیند کلی مدل ارائه شده

دانست. همچنین از دیگر موارد بسته‌های زائد می‌توان به بسته‌هایی اشاره نمود که ارسال آنها به هر دلیلی ناقص بوده و یا با موفقیت به مقصد نرسیده‌اند. در نتیجه گاهی اوقات مشاهده می‌شود که یک بسته چندین بار به صورت پی‌درپی ارسال می‌گردد. به همین منظور لازم است تا قبل از انجام هرگونه عملیات، بر روی بسته‌های دریافتی عملیات پیش پردازش انجام گیرد. در نتیجه در هنگام ضبط بسته‌ها لازم است تا فیلترهایی بر روی بسته‌ها اعمال گردند. شکل (۱۰.۱) نمایی از عملیات پیش پردازش را نشان می‌دهد. همانطور که در شکل مشاهده می‌شود، در این رساله چهار مرحله پیش پردازش برای



شکل ۱۰.۱: عملیات پیش‌پردازش بسته‌های ترافیک شبکه

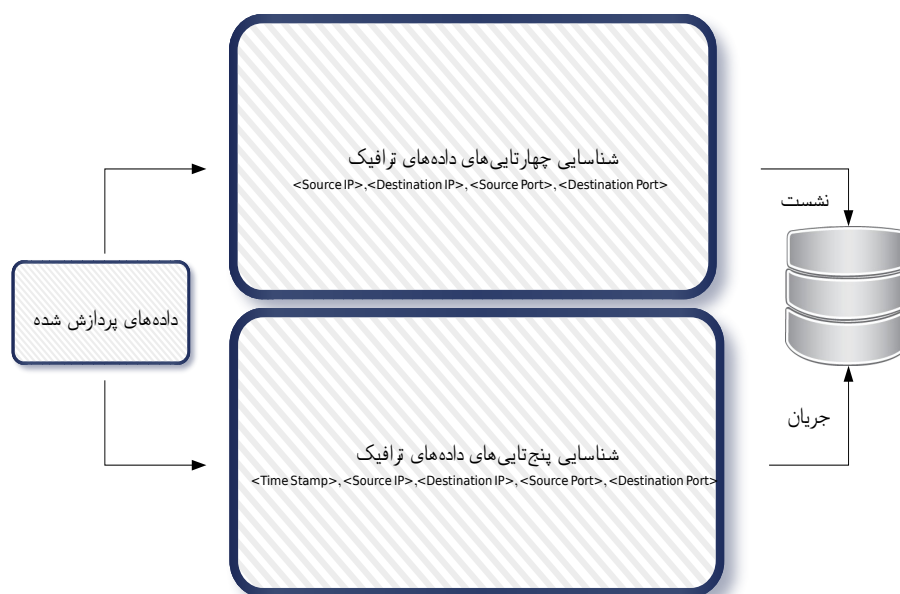
داده‌ها در نظر گرفته شده است. که این مراحل به ترتیب عبارتند از: (۱) حذف بسته‌های مربوط به سیستم عامل و دیوار آتش، (۲) حذف بسته‌های تکراری (بسته‌هایی که با موفقیت به مقصد نرسیده و ارسال آنها بارها تکرار شده است)، (۳) حذف قسمت‌های زائد و غیرضروری بسته‌ها و (۴) حذف جریان‌ها و نشست‌های ناقص (البته این قسمت از عملیات پیش پردازش در مرحله بعد و پس از شناسایی نشست‌ها انجام می‌گیرد).

۲.۹.۱ تشخیص نشست‌ها و جریان‌ها

همانطور که مطرح گردید، ترافیک شبکه دارای سطوح مختلفی می‌باشد. به طور خلاصه هنگامی که ارتباطی به جهت ارسال داده بین یک مبدا و مقصد برقرار می‌شود، یک نشست تولید می‌شود. یک نشست ممکن است در مدت زمان طولانی برقرار باشد به همین دلیل یک نشست به جریان‌هایی با فواصل زمانی کوتاه مدت تقسیم بندی می‌شود. در بعضی از پژوهش‌ها این مدت زمان را برابر ۳۰ دقیقه قرار داده‌اند [۲]. بسته‌های اولیه هر نشست دارای اطلاعات با ارزشی می‌باشند. به همین دلیل لازم است پس از آنکه عملیات پیش پردازش بر روی بسته‌ها انجام شد، نشست‌ها و همچنین جریان‌ها مشخص و جداسازی گردند. در اکثر مقالات مرتبط برای انجام عملیات شناسایی ابتدا جریان‌ها جدا شده و سپس به عملیات شناسایی پرداخته‌اند. برای نمونه آقای کروتی و همکاران پس از جداسازی جریان‌ها توزیع اندازه بسته‌های اول جریان‌ها و بعد از آن توزیع اندازه بسته‌های دوم جریان‌ها و همینطور تا آخر، با هم مقایسه شده‌اند [۱۳]. شیوه شناسایی نشست‌ها به این صورت است که آن بسته‌هایی که آدرس و شماره درگاه فرستنده و گیرنده آنها یکسان و یا بالعکس هستند، به عنوان یک نشست در نظر گرفته می‌شود. شیوه کار این مرحله در شکل (۱۱.۱) قابل مشاهده می‌باشد.

۳.۹.۱ محاسبه ویژگی‌های آماری مقاوم

یافتن ویژگی‌هایی که بتواند ترافیک یک برنامه کاربردی خاص را از سایر ترافیک سایر برنامه‌های کاربردی متمایز سازد، مهم‌ترین بخش در یک روش شناسایی ترافیک می‌باشد. فرض کنید که بسته‌های یک برنامه کاربردی دارای N ویژگی آماری باشند. استفاده از این N ویژگی جهت شناسایی برنامه کاربردی به دلایل محدودیت‌های پیاده‌سازی و حجم بالای پردازش بسیار پیچیده خواهد بود. همچنین در شناسایی برنامه کاربردی بکاربردن تمامی ویژگی‌ها لزوماً دقت شناسایی را افزایش نداده و حتی ممکن است در بعضی موارد موجب عدم کارآمدی روش نیز گردد. از طرفی دیگر همانطور که مطرح گردید، وجود سازوکارهای رسیدگی به صف‌ها و خدمت‌دهی به بسته‌ها در عناصر شبکه همانند پروتکل ARQ، موجب می‌شود تا

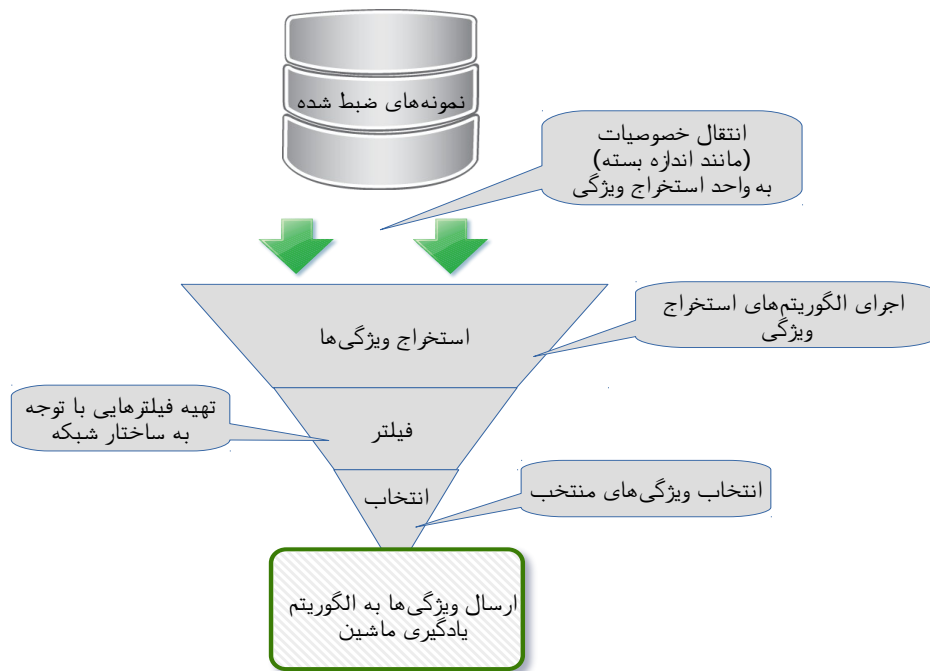


شکل ۱۱.۱: عملیات شناسایی نخست‌ها و جریان‌ها در داده‌های ترافیک شبکه

ویژگی‌های آماری بسته‌ها نه تنها به نوع برنامه‌های کاربردی، بلکه به تمامی ساختار شبکه‌ی موجود بین فرستنده و گیرنده، میزان ترافیک شبکه و رخداد‌های شبکه نیز وابسته باشد. لذا در روش ارائه‌شده و در مرحله استخراج ویژگی دو عامل را در نظر خواهیم گرفت. در مرحله اول با استفاده از تحلیل رفتاری ترافیک برنامه‌های کاربردی ویژگی‌های مقاوم و موثر در مقابل بروزرسانی برنامه‌های کاربردی استخراج شده و در مرحله بعد با تحلیل تأثیرات رخداد ائتلاف بسته‌ها بر این ویژگی‌ها ویژگی‌های مرتبه سومی جهت شناسایی ترافیک برنامه‌های کاربردی در شبکه ارائه می‌گردد. دو واحد محاسبه ویژگی و استخراج ویژگی در شکل (۱۲.۱) ارائه شده است.

۴.۹.۱ داده‌های ورودی

همانطور که در بخش قبل اشاره گردید، ورودی‌های روش ارائه شده به دو صورت داده‌های دارای برچسب و داده‌های ناشناخته می‌باشد. سطح ترافیک مورد استفاده، در سطح لایه شبکه بوده و ورودی روش، بسته‌های ارسالی توسط برنامه‌های کاربردی می‌باشد. به دلیل آنکه قرار است از ویژگی‌های آماری بسته‌ها استفاده شود، محتوای بسته‌ها دست نخورده باقی‌مانده و سرآیند بسته‌ها که شامل سرآیند لایه شبکه، سرآیند پروتکل TCP و سرآیند IP می‌باشد، به عنوان ورودی روش در نظر گرفته می‌شود. خروجی روش به صورت کلاس‌هایی از برنامه‌های کاربردی بوده و یک کلاس مخصوص داده‌های شناسایی نشده و یا بسته‌های مربوط به برنامه‌های کاربردی دیگر می‌باشد. شکل (۱۳.۱) و (۱۴.۱) نمونه سرآیند پروتکل IP



شکل ۱۲.۱: عملیات محاسبه و استخراج ویژگی از بسته‌های ترافیک شبکه

و TCP برای یک بسته می‌باشد.

```
Internet Protocol Version 4, Src: 245.234.7.169, Dst: 181.66.245.216
0100 .... = Version: 4
.... 0101 = Header Length: 20 bytes
Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
Total Length: 64
Identification: 0x0ff0 (4080)
Flags: 0x02 (Don't Fragment)
Fragment offset: 0
Time to live: 63
Protocol: TCP (6)
Header checksum: 0x8319 [validation disabled]
Source: 245.234.7.169
Destination: 181.66.245.216
[Source GeoIP: AS9394 China TieTong Telecommunications Corporation]
[Destination GeoIP: Peru, AS6147 Telefonica del Peru S.A.A., -12.043300, -77.028297]
```

شکل ۱۳.۱: نمونه سرآیند IP یک بسته

۱۰.۱ پایگاه داده‌های استفاده شده در رساله

۱.۱۰.۱ پایگاه داده UNIBS

این داده‌ها از دروازه^{۴۷} درون دانشگاه برسیکا جمع‌آوری شده است. اترنت^{۴۸} دانشگاه شامل چندین 1000Base-TX (که یکی از رایج‌ترین و سریع‌ترین اشکال اترنت فعلی است) تشکیل شده که مسیریابی توسط پردازشگرهای دو هسته‌ای لینوکس انجام می‌شود. داده‌ها از پیوند^{۴۹} 100Mb/s در دروازه لبه شبکه^{۵۰}

⁴⁷Gateway
⁴⁸Ethernet

⁴⁹Link
⁵⁰Edge

```

Transmission Control Protocol, Src Port: 60824 (60824), Dst Port: 80 (80), Seq: 0, Len: 0
Source Port: 60824
Destination Port: 80
[Stream index: 0]
[TCP Segment Len: 0]
Sequence number: 0 (relative sequence number)
Acknowledgment number: 0
Header Length: 44 bytes
Flags: 0x002 (SYN)
  000. .... = Reserved: Not set
  ...0 .... = Nonce: Not set
  .... 0... = Congestion Window Reduced (CWR): Not set
  .... .0.. = ECN-Echo: Not set
  .... ..0. = Urgent: Not set
  .... ...0 = Acknowledgment: Not set
  .... .... 0... = Push: Not set
  .... .... .0.. = Reset: Not set
  .... .... ..1. = Syn: Set
  .... .... ...0 = Fin: Not set
[TCP Flags: *****S*]
Window size value: 65535
[Calculated window size: 65535]
Checksum: 0x1ea4 [validation disabled]
Urgent pointer: 0
Options: (24 bytes), Maximum segment size, No-Operation (NOP), Window scale, No-Operation (NOP),
No-Operation (NOP), Timestamps, SACK permitted, End of Option List (EOL)

```

شکل ۱۴.۱: نمونه سرآیند TCP یک بسته

که متصل کننده به اینترنت بود به میزان ۷۲/۳ گیگابایت و به کمک نرم افزار Tcpdump جمع آوری شده است. این داده ها در دو روز و در ساعات شلوغی جمع آوری شده که روز اول تحت عنوان T_1 و روز دوم تحت عنوان T_2 ذخیره گردیده است. از آنجایی که دسترسی کامل به مسیریاب موجود بوده است، ۲۵۰ بایت اول هر فریم ذخیره شده تا با اعمال تطبیق الگو به محتوا به برنامه کاربردی که موجب جریان شده دست یابند و این امر با الگوریتم ground truth انجام شده است. به همین جهت برچسب جریان های مربوط به برنامه های کاربردی با اطمینان بالایی تهیه شده است. همچنین جریان های مربوط به SSL⁵¹ با توجه به شماره درگاه آنها تطبیق الگو شده اند. پس بسته های درگاه⁵² ۴۴۳ که مربوط به HTTPS و ۹۹۵ که مربوط به POPS3 است جدا شدند. سایر کلاس ها در جدول ۳.۱ نشان داده شده است.

جدول ۳.۱: پایگاه داده UNIBS

برنامه کاربردی	تعداد نمونه
Firefox	۳۵۱۶۴۱۰
Thunderbird	۶۵۶۶۰
Skype	۱۸۱۵۵۰
Safari	۳۴۸۲۶۵۰
Bittorrent	۱۲۴۴۲۴۰
Mail	۱۰۰۷۷۹۰

⁵¹Secure Sockets Layer

⁵²Port

۲.۱۰.۱ پایگاه داده جمع‌آوری شده

به دلیل آنکه اکثر پایگاه داده‌های موجود در اینترنت قدیمی بوده و اکثر دانشگاه‌ها حاضر به در اختیار گذاشتن پایگاه داده خود برای دیگران نیستند و همچنین به دلیل نیاز به ترافیک شبکه بروز، بر آن شدیم تا با وجود تمامی مشکلات این کار، پایگاه داده‌ای را تهیه نماییم. لذا پس از بررسی‌های گسترده، سامانه‌ای جهت ضبط بسته‌های ترافیک برنامه‌های کاربردی تهیه گردید. از مشکلات جمع‌آوری پایگاه داده می‌توان به وجود بسته‌های زائد بسیار زیاد و همچنین ضبط داده‌های دارای برچسب اشاره نمود. این داده‌ها از یک شبکه محلی و از طریق مسیریاب و از طریق پهنای باند با حجم ۱ Gb/s و توسط نرم‌افزار Tcpdump ضبط شده است. طی فرآیند ضبط داده‌ها که به مدت ۳ روز انجام شده و ترافیک هفت برنامه کاربردی ضبط شده است. روش ضبط به این صورت است که شبکه محلی دارای ۵۰ نود بوده و در هر یک از این نودها برنامه کاربردی خاصی اجرا شده است. عملیات ضبط از بستر مسیریاب انجام شده است و از آنجا که در هر نود یک برنامه اجرا شده است، عملیات برچسب زنی نام برنامه کاربردی پس از حذف داده‌های زائد با توجه به این نودها انجام شده است. از اینرو عملیات برچسب زنی داده‌ها با دقت بالایی انجام شده است. جدول ۴.۱ تعداد نمونه‌های جمع‌آوری شده از هفت برنامه کاربردی را به طور مجزا نشان داده است. پس از بررسی‌های لاگ نرم‌افزار Tcpdump و همچنین انجام آزمایش‌های متعدد مشخص گردید که میزان اتلاف بسته در این مرحله کمتر از ۲٪ می‌باشد.

جدول ۴.۱: پایگاه داده جمع‌آوری شده

تعداد نمونه	برنامه کاربردی
۱۶۳۶۰۷۰	Firefox
۱۳۰۲۱۳۰	Skype
۱۶۸۱۷۶۰	Telegram
۱۵۲۹۵۸۰	Download Manager
۱۳۰۲۶۱۰	Chrome
۱۳۰۵۰۳۰	IE
۱۲۰۴۳۳۰	Twitter

۱۱.۱ ساختار رساله

این رساله از پنج فصل تشکیل شده است. در ادامه ساختار رساله به صورت ذیل سازماندهی شده است:

در فصل دوم، از میان روش‌های ارائه شده، روش‌های شناسایی بر پایه ویژگی‌های آماری را به دلیل

اهمیت موضوع و همچنین حوزه کاری این رساله با جزئیات بیشتر مورد بررسی قرار گرفته و دلایل اهمیت ویژگی‌های آماری نیز مطرح شده است. همانطور که در فصل دوم مطرح خواهد شد، رهیافت روش‌های شناسایی برپایه ویژگی‌های آماری دارای دو مرحله (۱) استخراج ویژگی از داده‌های ترافیک شبکه و (۲) استفاده از الگوریتم‌های یادگیری ماشین می‌باشد. از اینرو در بخش دوم مروری بر کارهای انجام شده در این زمینه انجام شده و مقالات سال‌های ۲۰۰۵ تا ۲۰۱۷ مورد بررسی قرار خواهند گرفت.

✍ در فصل سوم، به موضوع تحلیل رفتاری پرداخته می‌شود. در این راستا ابتدا مقدمه‌ای در خصوص تحلیل رفتاری مطرح شده و سپس خصوصیات رفتاری برنامه‌های کاربردی در لایه شبکه مورد تحلیل و بررسی قرار می‌گیرد. پس از تحلیل و بررسی خصوصیات رفتاری برنامه‌های کاربردی چهار ویژگی آماری استخراج شده بیان می‌گردد. در ادامه به موضوع ائتلاف بسته‌ها و تاثیرات آنها بر ویژگی‌های آماری و همچنین دقت شناسایی برنامه‌های کاربردی پرداخته می‌شود. از اینرو مجدداً مجموعه داده‌ای حاوی رخداد ائتلاف بسته جمع‌آوری شده و پس از آن ویژگی‌های نوینی با استفاده از ویژگی‌های استخراج شده از تحلیل رفتاری برنامه‌های کاربردی استخراج می‌گردد.

✍ در فصل چهارم، پنج معیار اصلی مطرح شده در مقالات مختلف جهت ارزیابی یک روش شناسایی ترافیک شبکه مورد بررسی قرار گرفته و با توجه به این معیارها، ویژگی‌های ارائه شده در فصل سوم را مورد ارزیابی قرار می‌دهیم. به همین منظور، با استفاده از مباحث تئوری اطلاعات نشان داده شده است که این ویژگی‌ها از اهمیت بسیار بالایی برخوردار هستند و در آخر نیز میزان صحت شناسایی برنامه‌های کاربردی توسط این ویژگی‌ها بر روی دو پایگاه داده برای چالش‌های روزرسانی برنامه‌های کاربردی و رخداد ائتلاف بسته مورد بررسی و ارزیابی قرار می‌گیرد.

✍ در فصل پنجم به جمع‌بندی رساله خواهیم پرداخت و پیشنهاداتی برای ادامه کار مطرح شده است.

۱۲.۱ جمع‌بندی

در این فصل مقدمه‌ای در خصوص ساختار شبکه و نحوه تبادل داده در آن مورد بررسی قرار گرفت و توضیحاتی در خصوص روش‌های شناسایی ترافیک شبکه ارائه گردید. همچنین چالش‌های موجود برای ارائه یک سامانه شناسایی ترافیک شبکه مطرح و ضرورت انجام این مسئله مورد بررسی قرار گرفت. در نهایت به تعریف مساله، فرض‌های موجود، مدل ارائه‌شده در این رساله و همچنین پایگاه داده‌های مورد استفاده پرداختیم.

فصل دوم

مروری بر روش‌های شناسایی ترافیک شبکه بر پایه ویژگی‌های آماری

۱.۲ مقدمه

همانطور که در فصل اول مطرح شد، روش‌های موجود جهت شناسایی ترافیک برنامه‌های کاربردی را می‌توان در چهار گروه تقسیم‌بندی کرد: (آ) روش‌های مبتنی بر شماره درگاه، (ب) روش‌های مبتنی بر بررسی محتوای بسته‌های ترافیک، (ج) روش‌های مبتنی بر رفتار کاربر، (د) روش‌های مبتنی بر ویژگی‌های آماری بسته‌های ترافیک در حال عبور. در این میان به دلیل مشکلات مطرح شده، روش‌های مبتنی بر تحلیل آماری بسته‌ها دارای اهمیت بالایی می‌باشند. همچنین در این روش‌ها نیاز به بررسی محتوای بسته‌های ترافیک نبوده و با هدف حفظ محرمانگی اطلاعات مبادله شده انجام می‌شود. از اینرو در این بخش به مرور روش‌های ارائه شده در خصوص شناسایی ترافیک شبکه با استفاده از ویژگی‌های آماری پرداخته می‌شود.

۲.۲ رهیافت روش‌های شناسایی ترافیک بر پایه ویژگی‌های آماری

به طور کلی عمده کارهای انجام شده در حوزه شناسایی و طبقه‌بندی ترافیک را می‌توان در دو مرحله تقسیم‌بندی نمود.

- ارائه یک یا چند ویژگی^۱ برای تمایز بین ترافیک برنامه‌های کاربردی مختلف.
- پیشنهاد روشی نوین با استفاده از ویژگی‌های ارائه شده و استفاده از الگوریتم‌های یادگیری ماشین جهت شناسایی برنامه‌های کاربردی

۱.۲.۲ یافتن ویژگی‌های تمایز دهنده

به طور کلی هر چه ابعاد یا همان تعداد ویژگی‌های مساله مورد کاوش بالاتر رود، باعث خواهد شد رکوردها در فضای جستجو پراکنده‌تر شوند. استخراج ویژگی و کاهش ابعاد فضای نمونه نقش مهمی در مسائل مربوط به یادگیری ماشین ایفا می‌کنند. یافتن ویژگی‌هایی که بتواند ترافیک یک برنامه کاربردی را از سایر برنامه‌های کاربردی متمایز سازد، مهم‌ترین و اثرگذارترین بخش یک روش شناسایی ترافیک است. X را بسان یک موجودیت در بین تعداد N موجودیت در نظر گرفته و می‌توان K ویژگی برای هر یک از اعضای مجموعه یاد شده در نظر گرفت. به عنوان مثال اگر X نامی برای یک مدل معین از بین تمامی ماشین‌های تولید شده در کل دنیا باشد. برای تک تک ماشین‌های موجود در جهان می‌توان تعداد بسیار زیادی ویژگی در نظر گرفت، مثلاً رنگ، بیشینه سرعت، شتاب، تعداد سیلندر، تعداد درب.

¹Feature

اگرچه دانستن تمامی K ویژگی برای هر عضو مجموعه، مقصود را حاصل می‌نماید، اما به دلیل محدودیت‌های پیاده‌سازی دستیابی و اجرای پردازش بر روی تمامی K ویژگی به بسیار پیچیده خواهد بود. اگر نیاز به شناسایی هر چه سریعتر مدل ماشین X باشد، باید ویژگی‌هایی ارائه شوند که با دانستن آن میزان ابهام مسئله کاسته شود. مثلاً دانستن بیشینه سرعت ماشین و شتاب آن به احتمال زیاد اطلاعات بیشتری نسبت به دانستن رنگ ماشین در بر خواهد داشت. به عبارت دیگر ویژگی‌هایی نیاز است که میزان تمایز بیشتری بین نمونه‌هایی مجموعه N تایی ایجاد کند. جهت یافتن ویژگی‌های تمایزدهنده می‌بایست موارد زیر در نظر گرفته شود:

❖ استخراج ویژگی‌ها. باید دقت داشت که هر چه تعداد ویژگی‌ها بیشتر شود، لزوماً دقت بالا نخواهد رفت، و از سوی دیگر حجم پردازشی با افزایش تعداد ویژگی بیشتر خواهد شد. پس باید ویژگی‌ها را به دقت انتخاب نمود.

❖ تعریف معیارهایی برای میزان تمایز ویژگی‌ها و میزان اطلاعاتی که هر ویژگی در اختیار قرار می‌دهد.

❖ تبدیل فضای ویژگی‌های مناسب به فضای کوچکتر. پس از انتخاب ویژگی‌هایی که تمایزدهندگی بیشتری ایجاد می‌نمایند، ممکن است بازهم فضای این دسته ویژگی‌ها بسیار بزرگ باشد. بدین منظور روش‌های مختلفی مانند PCA^2 ، ICA^3 [۳۹] وجود دارد که می‌توان بدون از دست دادن اطلاعات زیاد، فضای ویژگی‌ها را کوچکتر نیز نمود. کوچک شدن فضای ویژگی علاوه بر افزایش سرعت پردازش، حجم حافظه مورد نیاز را نیز کاهش داده و در پاره‌ای از موارد دقت تمایز بین ویژگی‌ها را نیز بالا می‌برد.

هنگام استفاده از روش‌های شناسایی بر پایه ویژگی‌های آماری، استخراج ویژگی‌ها بدون بررسی محتوای بسته‌ها انجام می‌شود. در سال‌های اخیر، روش‌های متعددی برای استخراج ویژگی‌های آماری از بسته‌های شبکه ارائه شده است. در مقاله [۲۴] به طور کلی تعداد ۲۴۸ ویژگی برای بسته‌های شبکه تعریف شده که در مراجع مختلف زیرمجموعه‌های متفاوتی از این ویژگی‌ها انتخاب شده است، اما دلیل انتخاب این زیرمجموعه مطرح نشده و فقط به نتایج شبیه سازی اتکا شده است.

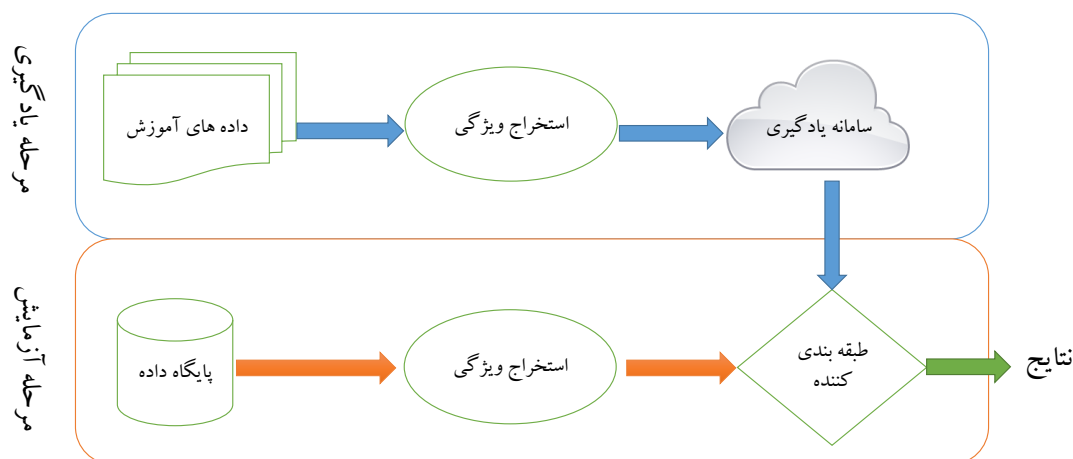
۲.۲.۲ تکنیک‌های یادگیری ماشین

بعد از یافتن یک یا چند ویژگی مناسب، گام بعدی استفاده از الگوریتم‌های یادگیری ماشین جهت تعیین سطح آستانه ویژگی‌ها، جهت ایجاد تمایز و شناسایی برنامه‌های کاربردی می‌باشد. بدین منظور، ویژگی‌های

²Principal Component Analysis

³Independent Component Analysis

یاد شده به یک الگوریتم یادگیری ماشین داده می‌شود، و این الگوریتم می‌تواند برطبق یادگیری که به مرور زمان انجام می‌دهد، با استفاده از ویژگی‌های یاد شده و یادگیری خودش، عملیات شناسایی ترافیک را انجام دهد. برحسب نحوه یادگیری، الگوریتم‌های یادگیری ماشین را می‌توان به چندگروه با نظارت، نیمه نظارتی، بدون ناظر و یادگیری تقویتی تقسیم‌بندی نمود [۴۰]. در شکل ۱.۲ نمایی از الگوریتم یادگیری بانظارت نشان داده شده است.



شکل ۱.۲: الگوریتم‌های یادگیری ماشین از نوع بانظارت

تاکنون الگوریتم‌های یادگیری ماشین بسیار مختلفی پیشنهاد شده است، که هر کدام از جنبه‌هایی بر دیگری برتری دارد. به عنوان برخی از الگوریتم‌های مشهور یادگیری ماشین می‌توان به ماشین یادگیری پشتیبان، شبکه‌های عصبی مصنوعی، شبکه بیز، یادگیری درخت تصمیم، یادگیری تقویتی، الگوریتم‌های خوشه‌بندی (K-means, DBSCAN, EM) اشاره نمود [۴۱].

۳.۲ کارهای انجام شده در زمینه شناسایی بر پایه ویژگی‌های آماری

در ادامه این بخش مروری بر مقاله‌هایی که در سال‌های اخیر جهت شناسایی ترافیک شبکه با استفاده از ویژگی‌های آماری ارائه شده‌اند، انجام خواهد شد.

Kim و همکارانش روشی برای طبقه‌بندی ترافیک ارائه کردند که بر پایه مدل مارکف عمل کرده و از معیار تعیین همگرایی و واگرایی Kullback-Leibler (یک معیار شباهت در بین توزیع‌های احتمالاتی) جهت بررسی ترافیک برنامه‌های کاربردی استفاده شده است. این روش با دقت ۹۰٪ قادر به شناسایی ترافیک برنامه‌های کاربردی در شرایط تداخل بسته‌ها می‌باشد [۴۲]. در این روش از یادگیری بانظارت استفاده

شده و با استفاده از اندازه چهار بسته اول هر جریان برای هر برنامه کاربردی یک مدل مارکوف تهیه شده است. بطور نمونه در این مقاله نشان داده شده است که داده های دو برنامه کاربردی *SMTP* و *IMAP* دارای تداخل زیاد بوده و دارای حالت های شبیه به هم هستند. از معیار F-measure برای ارزیابی روش پیشنهادی استفاده شده است. در فاز یادگیری برای هر برنامه کاربردی یک مدل مارکف تشکیل شده و در فاز آزمایش ابتدا یک مجموعه ای از جریان ترافیک را به نام جعبه جریان (Bag of Flow) تشکیل داده و مطرح کرده است که این گروه ها دارای شماره درگاه یکسان هستند. سپس برای هر گروه یک مدل مارکف تشکیل داده است. سپس از معیار واگرایی Kullback-Leibler استفاده کرده و میزان واگرایی هر کدام از مدل های مارکف فاز آموزش با مدل های مارکف فاز آزمایش را تعیین نموده است. گروه هایی که میزان واگرایی کمی را با هر برنامه کاربردی دارند به آن برنامه کاربردی اختصاص داده می شوند.

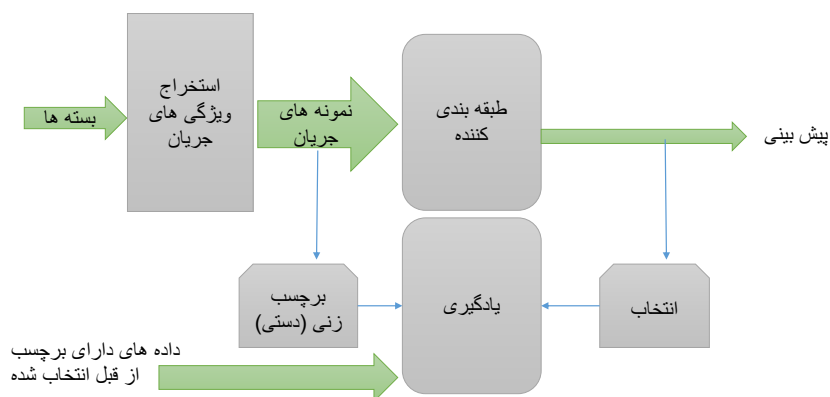
Muehlstein و همکارانش نشان دادند که با بررسی ترافیک شبکه نوع سیستم عامل، نوع مرورگر و نوع برنامه کاربردی داده های HTTP و داده های رمز شده HTTPS را می توان شناسایی نمود [۷]. نویسندگان ادعا کردند که در ترافیک رمز شده، برای اولین بار موفق به شناسایی سیستم عامل، مرورگر و برنامه کاربردی شدند. در آن تحقیق، پایگاه داده ای حاوی ۲۰ هزار نشست مورد ارزیابی قرار گرفت. که شامل ترافیک سیستم عامل های ویندوز، اوبونتو، آی او اس (IOS) و مرورگر های کروم، اینترنت اکسپلورر، فایرفاکس و سافاری، و برنامه های کاربردی یوتیوب، فیسبوک و توییتر بوده است. در این روش، هر نشست که شامل پنج تایی (شماره درگاه مبدا و مقصد، آدرس مبدا و مقصد، پروتکل مربوطه) است، به یک سه تایی (سیستم عامل، مرورگر و برنامه کاربردی) نگاشت می شود. در این مقاله از الگوریتم یادگیری SVM و هسته RBF استفاده شده است. ویژگی های مورد استفاده در این کار بصورت جدول (۱.۲) می باشد.

Loo و همکارانش الگوریتمی بر پایه خوشه بندی k-means افزایشی جهت یادگیری نمونه های دارای برچسب و بدون برچسب ارائه داده اند و جهت سنجش میزان شباهت نمونه ها از دو معیار فاصله اقلیدسی و منتهن استفاده کرده است. آزمایش بر روی ۶۷۱ هزار جریان انجام شده و میزان دقت خوشه بندی برابر ۹۴٪ گزارش شده است. همچنین سرعت اجرای الگوریتم با استفاده از معیار فاصله منتهن سه برابر بهتر از فاصله اقلیدسی ارزیابی شده است [۱]. مدل ارائه شده جهت انجام عملیات خوشه بندی به صورت شکل (۲.۲) ارائه شده و همچنین ویژگی های مورد استفاده به صورت جدول (۲.۲) می باشد.

Qin و همکارانش مدلی به نام مدل "جریان دو طرفه" ارائه داده اند که می تواند خصوصیات رفتاری متقابل بین نودهای مختلف را ضبط کند. از توزیع احتمالی اندازه محتوای هر بسته (PSD) موجود در این

جدول ۱۰.۲: ویژگی‌های آماری مورد استفاده در روش Muehlstein [۷]

ویژگی‌های پایه	ویژگی‌های ارائه شده
تعداد بسته‌های ارسال شده	اندازه اولیه پنجره TCP
تعداد کل بایت ارسال شده	میزان تغییر مقیاس پنجره TCP
کمترین مدت زمان مابین ارسال بسته‌ها	تعداد روش‌های فشرده‌سازی SSL
بیشترین مدت زمان مابین ارسال بسته‌ها	تعداد افزونه‌های SSL
میانگین مدت زمان مابین ارسال بسته‌ها	تعداد روش‌های رمزنگاری SSL
انحراف معیار مدت زمان مابین ارسال بسته‌ها	طول شماره (ID) نشست SSL
میانگین بسته‌های ارسال شده	حداکثر توان ارسال بسته‌ها
انحراف معیار بسته‌های ارسال شده	میانگین توان دریافت بسته‌ها
تعداد بسته‌های دریافت شده	بیشترین توان دریافت بسته‌ها
تعداد کل بایت دریافت شده	کمترین توان دریافت بسته‌ها
کمترین مدت زمان مابین دریافت بسته‌ها	انحراف معیار توان دریافت بسته‌ها
بیشترین مدت زمان مابین دریافت بسته‌ها	تعداد توده‌های ارسالی
میانگین مدت زمان مابین دریافت بسته‌ها	تعداد توده‌های دریافتی
انحراف معیار مدت زمان مابین دریافت بسته‌ها	کمترین توان ارسال بسته‌ها
میانگین بسته‌های دریافت شده	میانگین توان ارسال بسته‌ها
انحراف معیار بسته‌های دریافت شده	انحراف معیار توان ارسال بسته‌ها
میانگین مقدار زمان حیات بسته‌ها	میانگین مدت زمان مابین دریافت توده‌ها
کمترین بسته ارسال شده	کمترین مدت زمان مابین دریافت توده‌ها
کمترین بسته دریافت شده	بیشترین مدت زمان مابین دریافت توده‌ها
بیشترین بسته ارسال شده	انحراف معیار مدت زمان مابین دریافت توده‌ها
بیشترین بسته دریافت شده	میانگین مدت زمان مابین ارسال توده‌ها
تعداد کل بسته‌ها	کمترین مدت زمان مابین ارسال توده‌ها
کوچکترین اندازه بسته‌ها	بیشترین مدت زمان مابین ارسال توده‌ها
بزرگترین اندازه بسته‌ها	انحراف معیار مدت زمان مابین ارسال توده‌ها
میانگین اندازه بسته‌ها	تعداد بسته‌های در حال عبور
واریانس اندازه بسته‌ها	بیشترین اندازه قطعه TCP

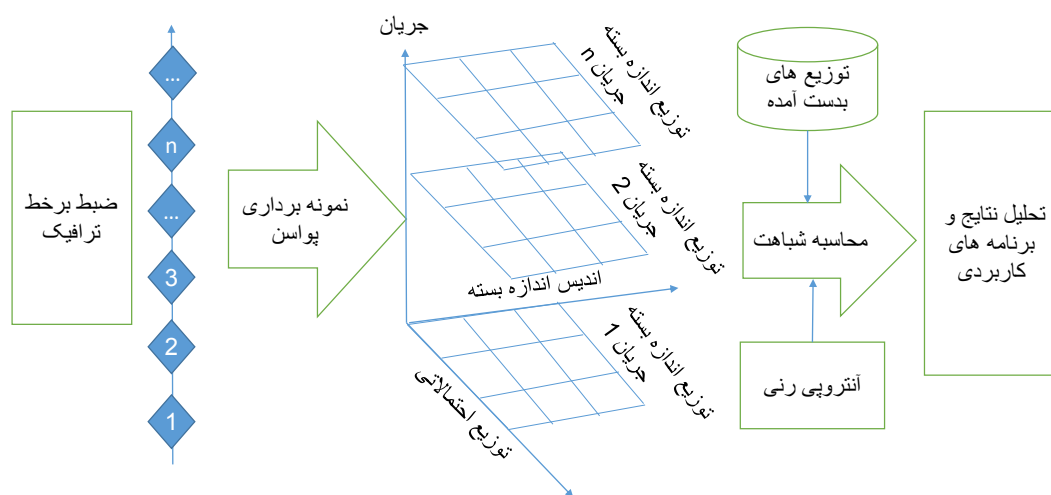


شکل ۲.۲: مدل ارائه شده در روش Loo [۸]

جدول ۲.۲: ویژگی‌های آماری مورد استفاده در روش Loo [۱]

ویژگی‌ها
شماره درگاه ارسال
شماره درگاه دریافت
مقدار کل بایت در بسته‌های IP
مقدار بایت در بسته‌های IP ارسالی
مقدار بایت در بسته‌های IP دریافتی
مقدار کل بایت در بسته‌های Ethernet
مقدار بایت در بسته‌های Ethernet ارسالی
مقدار بایت در بسته‌های Ethernet دریافتی
مقدار کل بایت در بسته‌های کنترلی
مقدار بایت در بسته‌های کنترلی ارسالی
مقدار بایت در بسته‌های کنترلی دریافتی

مدل به عنوان ویژگی استفاده شده است [۲]. بسته‌های (رفت و برگشت) دارای مبدا و مقصد^۴ یکسان به عنوان یک جریان دو طرفه در نظر گرفته می‌شوند. در مرحله بعد نشان داده شده است که توزیع اندازه بسته‌های مربوط به برنامه‌های کاربردی مختلف بایکدیگر متفاوت می‌باشد. سپس از Renyi Cross Entropy برای محاسبه شباهت بین PSD یک جریان دو طرفه با PSD برنامه‌های کاربردی شناخته شده استفاده می‌شود. در این مقاله جهت پردازش داده‌های حجیم از روش Poisson Sampling استفاده شده است.

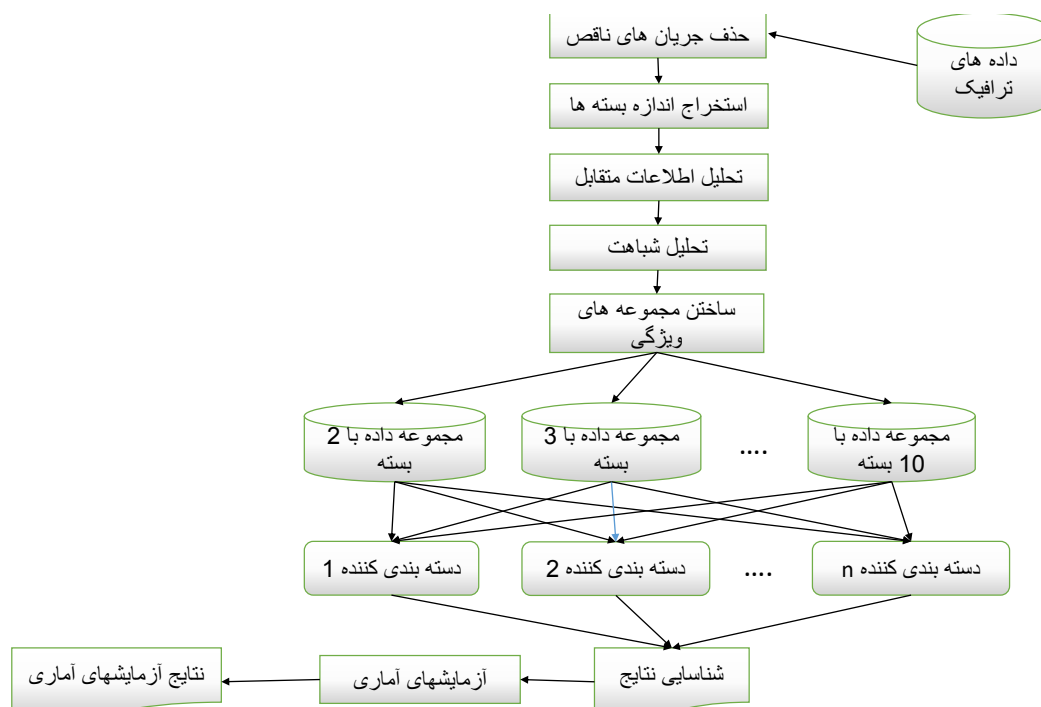


شکل ۳.۲: مدل ارائه شده در روش Qin [۲]

در مقاله [۲] در ابتدا نویسنده سوالی با عنوان اینکه چه تعداد بسته جهت شناسایی ترافیک شبکه

⁴Destination

مورد نیاز می‌باشد، مطرح نموده است. در همین راستا برای بدست آوردن تجربی تعداد بسته‌های مورد نیاز، از سه پایگاه داده استفاده نموده و برای هر جریان، اندازه ۱۰ بسته اول را استخراج نموده است که اندازه ۱۰ بسته را به عنوان بردار ویژگی‌های هر جریان مربوط به یک برنامه کاربردی در نظر گرفته است. در ادامه با حفظ ترتیب بسته‌ها، ابتدا توسط اطلاعات متقابل^۵ میزان اطلاعات^۶ که n بسته اول حمل می‌کند را مورد بررسی قرار داده و سپس همبستگی^۷ هر جفت بسته مجاور^۸ به یکدیگر جهت پیدا کردن فراوانی ویژگی‌ها اندازه‌گیری شده است. در مرحله بعد برای هر بردار ویژگی $F = (f_1, f_2, \dots, f_n)$ مربوط به یک جریان، تعداد نه بردار ویژگی تهیه شده است. اولین بردار حاوی دو ویژگی اول، دومین بردار حاوی سه ویژگی اول، ...، تا آخر. سپس هر کدام از این بردارها به ۱۱ الگوریتم طبقه‌بندی^۹ داده می‌شود. در مرحله آخر نتایج حاصل از ۱۱ الگوریتم طبقه‌بندی که هر کدام نه بردار ویژگی را طبقه‌بندی کرده‌اند، توسط الگوریتم‌های آماری مورد بررسی قرار گرفته و تعداد بسته‌هایی که بهترین نتایج را تولید کرده‌اند مشخص شده است. میانگین دقت شناسایی برنامه‌های کاربردی بدست آمده از این روش ۹۵ درصد گزارش شده است. مدل پیشنهادی به صورت شکل (۴.۲) می‌باشد.



شکل ۴.۲: مراحل تشخیص تعداد بسته در روش Peng [۳]

علی اکبریان و همکاران با استفاده از الگوریتم‌های یادگیری ماشین نظارت شده سعی در نشان دادن بهترین تعداد ویژگی برای شناسایی ترافیک داشته‌اند [۸]. همچنین از الگوریتم‌های درخت تصمیم،

^۵Mutual Information

^۶Information

^۷Correlation

^۸Adjacent

^۹Classification

شبکه عصبی مصنوعی، یادگیری بیز، Bagging و Boosting جهت طبقه بندی ترافیک استفاده کرده اند. بر اساس نتایج این مقاله، اگر تعداد M کلاس برنامه کاربردی داشته باشیم، تعداد ویژگی ها در بهترین حالت می تواند به $M - 1$ کاهش یابد. برای ارزیابی روش تعداد هشت برنامه کاربردی در نظر گرفته شده است. همانطور که در جدول (۳.۲) مشاهده می شود، الگوریتم Boosting در تعداد هفت ویژگی بهترین جواب را دارا می باشد.

جدول ۳.۲: نتایج الگوریتم های مورد استفاده در روش علی اکبریان [۸]

Boosting	Bagging	NB	ANN	DT	Size
91.9126	90.6452	64.3942	60.6980	87.1978	248
91.9153	90.4321	87.7442	87.3266	87.0666	100
91.4593	90.8567	87.5600	86.5876	85.9945	90
91.0192	90.0192	86.8429	87.2666	86.5648	80
91.7730	91.7493	87.0098	87.5000	86.1770	70
91.0000	90.3682	87.9334	87.5090	86.0173	60
91.6222	91.6433	87.2107	87.5000	87.0894	50
91.0029	91.0012	87.0126	86.9531	87.1807	40
91.3333	92.7573	87.9987	87.0118	86.7016	30
92.2731	92.0019	87.4532	86.3595	87.0210	20
92.1235	92.1800	87.3452	87.4564	86.8613	10
92.9982	91.2344	87.4088	87.6597	86.4507	7
88.0111	87.8782	63.3334	56.0000	83.8221	5

در مقاله [۴۳] یک روش جدید جهت طبقه بندی ترافیک با استفاده از شبکه عصبی پیشخور (feed-forward) ارائه شده است که معایب روش های مبتنی بر شماره درگاه و محتوا را برطرف می کند. در این مقاله از الگوریتم استخراج ویژگی برپایه شباهت بیشتر جهت انتخاب ویژگی های مناسب استفاده شده است. از آنجا که روش های مختلفی برای آموزش شبکه موجود می باشد، در این مقاله از تنظیم بیزی استفاده شده است. تنظیم بیزی میزان ترکیب خطی مربع خطاها و مربع پارامترهای شبکه را کم می کند. این کار به جهت جلوگیری از بیش برآزش (over-fitting) پایگاه داده انجام می شود. تابع هدف به صورت زیر می باشد: $F = aE_d + (1 - a)E_w$ که E_d مجموع مربعات خطا و E_w مجموع مربعات وزن های شبکه و α پارامتری جهت مشخص نمودن نسبت وزن دهی بین E_d و E_w می باشد. مرحله استخراج ویژگی دارای دو مرحله می باشد. ۱- تشخیص ارتباط بین یک ویژگی و یک نوع از برنامه های کاربردی ۲- تشخیص افزونگی بین ویژگی های مختلف. در انتها نیز برای نشان دادن نتایج خروجی روش پیشنهادی با خروجی طبقه بندی کننده بیزی مقایسه شده است و نشان داده شده است که این

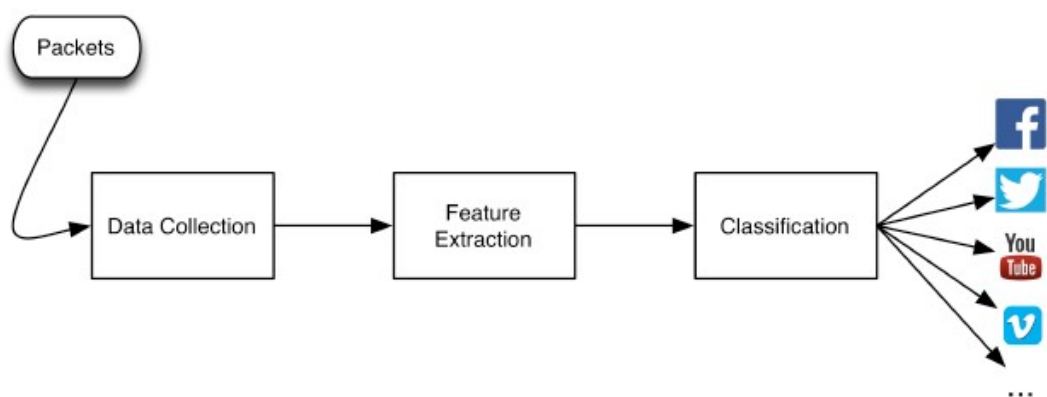
روش در میزان دقت و زمان آموزش از طبقه‌بندی بیزی بهتر عمل می‌کند.

هدف مقاله [۹] ارزیابی مقاومت روش‌های یادگیری ماشین جهت طبقه‌بندی ترافیک رمز شده شبکه از جمله ssh و skype می‌باشد. منظور از مقاومت در اینجا این می‌باشد که وقتی طبقه‌بندکننده ما با داده‌های یک شبکه آموزش دید، در فاز آزمایش بر روی داده‌های شبکه‌های دیگر به خوبی جواب دهد. برای این منظور از پنج الگوریتم یادگیری AdaBost, SVM, naive Bayesian, RIPPER, C4.5 استفاده شده است. در داده‌های مورد استفاده آدرس فرستنده و گیرنده و همچنین شماره درگاه فرستنده و گیرنده مشخص نبوده و به صورت رمز می‌باشد. این مقاله مدلی جهت شناسایی ترافیک SSH از غیر SSH و شناسایی ترافیک Skype از ترافیک غیر Skype ارائه کرده است. نتایج کار نشان دهنده این موضوع می‌باشد که الگوریتم C4.5 بهتر از باقی الگوریتم‌ها در خصوص شناسایی SSH و Skype عمل کرده است. ویژگی‌های مورد استفاده در جدول (۴.۲) ارائه شده است.

جدول ۴.۲: ویژگی‌های آماری مورد استفاده در روش alshammari [۹]

ویژگی‌ها	
پروتکل	مدت زمان جریان
تعداد بسته‌های ارسال شده	تعداد بایت ارسال شده
تعداد بسته‌های دریافت شده	تعداد بایت دریافت شده
کمترین مدت زمان مابین ارسال بسته‌ها	کمترین مدت زمان مابین دریافت بسته‌ها
انحراف معیار مدت زمان مابین ارسال بسته‌ها	انحراف معیار مدت زمان مابین دریافت بسته‌ها
میانگین مدت زمان مابین ارسال بسته‌ها	میانگین مدت زمان مابین دریافت بسته‌ها
بیشترین مدت زمان مابین ارسال بسته‌ها	بیشترین مدت زمان مابین دریافت بسته‌ها
کوچکترین اندازه بسته ارسالی	کوچکترین اندازه بسته دریافتی
بزرگترین اندازه بسته ارسالی	بزرگترین اندازه بسته دریافتی
انحراف معیار اندازه بسته ارسالی	انحراف معیار اندازه بسته دریافتی
میانگین اندازه بسته ارسالی	میانگین اندازه بسته دریافتی

Yamansavascular و همکارانش سعی داشتند که برنامه‌های کاربردی همچون توییتر، فیسبوک و اسکایپ بر روی دو پایگاه داده UNJ و پایگاه داده جمع آوری شده حاوی تعدادی برنامه کاربردی شناسایی نمایند [۴]. برای ارزیابی نتایج از چهار الگوریتم دسته‌بندی J48, Random forest, K-NN و Bayesnet استفاده کردند. نتایج حاصل از انجام آزمایش توسط الگوریتم K-NN با مجموعه ۱۱۱ ویژگی برابر ۹۳/۹۴٪، و با استفاده از Random forest دقت شناسایی برابر ۸۷/۹۰٪ شده است. کلیات انجام کار در شکل (۵.۲) نشان داده شده است.



شکل ۵.۲: مراحل شناسایی ترافیک شبکه در روش Yamansavascular [۴]

جهت شناسایی موثر ترافیک شبکه به صورت برخط روشی با استفاده از درخت شبکه عصبی انعطاف-پذیر (FNT) و استفاده از سه مجموعه داده جهت ارزیابی این روش ارائه شده است [۴۴]. در این روش از ویژگی‌های اندازه بسته‌ها و میانگین آنها، بیشترین مقدار، کمترین مقدار و انحراف معیار اندازه بسته‌ها در جریان‌ها استفاده شده است، و توانست با بررسی شش بسته ابتدایی برنامه کاربردی را شناسایی نماید. Ertam و همکارانش برای شناسایی ترافیک شبکه از روش ELM استفاده نموده‌اند. در این روش ابتدا از شبکه عصبی با یک لایه مخفی استفاده شده که در آن وزن نرون‌های ورودی و مخفی به صورت تصادفی تعیین می‌شود. میزان دقت دسته‌بندی توسط تغییر تعداد لایه‌های مخفی و همچنین تغییر پارامتر تابع هدف (تابع موجک) به ۹۵٪ رسیده است [۱۰]. جدول ۵.۲ ویژگی‌های آماری استفاده شده را نشان می‌دهد.

جدول ۵.۲: ویژگی‌های آماری مورد استفاده در روش Ertam [۱۰]

ویژگی‌ها	
شماره درگاه سرور	شماره درگاه کلاینت
تعداد بسته‌های ارسال شده	تعداد بسته‌های دریافت شده
کوچکترین اندازه بسته ارسالی	میانگین اندازه بسته ارسالی
میانگین اندازه بسته دریافتی	کوچکترین اندازه بسته دریافتی
بزرگترین اندازه بسته ارسالی	بزرگترین اندازه بسته دریافتی
انحراف معیار اندازه بسته ارسالی	انحراف معیار اندازه بسته دریافتی

جدول (۶.۲) مروری بر روش‌های شناسایی ترافیک در سال‌های ۲۰۰۵ تا ۲۰۱۳ داشته و میزان دقت هرکدام به همراه پایگاه داده مورد استفاده را مشخص نموده است.

جدول ۶.۲: کارایی روش‌های ارائه شده

پایگاه داده	معیار ارزیابی	دقت شناسایی (درصد)	برنامه‌های کاربردی	ویژگی‌های مورد استفاده	روش شناسایی استفاده شده	مقاله مرجع
NLANR and Waikato trace	میزان درستی	N/A	A mixture of HTTP, SMTP, FTP (control), NTP, IMAP, DNS ...	طول بسته (حداقل، حداکثر)، مدت زمان اتصال، مدت زمان مابین ارسال بسته‌ها، اندازه بایت، مدت زمان بیکاری یا منتظر بودن	EM	[۴۵]
Proprietary Hand Classified Traces	میزان درستی	96.29	A large range of Database, P2P, Bulck, Mail, Services	مدت زمان جریان، شماره درگاه، مدت زمان مابین ارسال بسته‌ها، اندازه بسته، پهنای باند مورد استفاده بر اساس آنتروپی، تبدیل فوریه مدت زمان مابین ارسال بسته‌ها	Bayesian Techniques	[۴۶]
Proprietary traces	میزان درستی	90	eDonkey, FTP, HTTP, Kazaa, NTP, POP3, SMTP, SSH, HTTPS, POP3S	طول بسته‌های جریان‌های رفت و برگشت	Simple KMeans	[۴۷]
NLANR, USC/ISI, CAIDA	پیچیدگی زمانی	94	WWW, Telnet, Chat (Messenger), FTP, P2P (Kazaa, Gnutella), Multimedia, SMTP, POP, IMAP, NDS, Oracle, X11	مدت زمان جریان، مقدار بایت پنجره انتشار، تعداد بسته‌های داده، طول بسته‌ها، مدت زمان مابین ارسال بسته‌ها، اندازه بسته‌های توده	Naive Bayes with Kernel Estimation, Decision Tree J48 and Reduced Error Pruning Tree	[۴۸]
Proprietary Hand Classified Traces	میزان درستی و پردازش	91.87	www, mail, web, p2p, ftp, im, voip	Correlation based Feature Selection algorithm and Consistency based Feature selection algorithm	MLP, RBF, C4.5, Bayes Net and Naïve Bayes	[۴۹]
Captured at Ohio University	میزان درستی	95.21	www, mail, web, p2p, Ftp, IM	Fast Correlation-Based Filter Algorithm	C4.5 Decision Tree	[۵۰]

مقاله مرجع	روش شناسایی استفاده شده	ویژگی‌های مورد استفاده	برنامه‌های کاربردی	دقت شناسایی (درصد)	معیار ارزیابی	پایگاه داده
[۵۲]	K-Means, DBSCAN and AutoClass	تعداد کل بسته‌ها، میانگین اندازه بسته‌ها، میانگین اندازه قسمت داده بسته‌ها، میزان بایت انتقال داده شده، میانگین مدت زمان مابین ارسال بسته‌ها	www, mail, web, p2p, ftp, im, voip	92.4	میزان درستی	NLANR and a self-collected 1-hour trace from the University of Calgary
[۵۳]	Supervised Naive Bayes	طول بسته (حداقل)، حداکثر، میانگین، انحراف معیار، طول بسته‌های ورودی (حداقل، حداکثر، میانگین، انحراف معیار)، مدت زمان مابین ارسال بسته‌ها (کمترین، بیشترین، میانگین، انحراف معیار)، محاسبه تعداد بسته‌های متوالی در هر جریان	Online Game (Enemy Territory) traffic, Others (HTTP, HTTPS, DNS, NTP, SMTP, Telnet, SSH, P2P ...)	95.5	میزان درستی	Traces collected at an online game server in Australia and provided by University of Twente, Netherland
[۵۴]	Naive Bayes and Decision Tree	طول بسته (حداقل)، حداکثر، میانگین، انحراف معیار، طول بسته‌های ورودی (حداقل، حداکثر، میانگین، انحراف معیار)، مدت زمان مابین ارسال بسته‌ها (کمترین، بیشترین، میانگین، انحراف معیار)، محاسبه تعداد بسته‌های متوالی در هر جریان	Online Game (Enemy Territory) traffic, Others (HTTP, HTTPS, DNS, NTP, SMTP, Telnet, SSH, P2P ...)	97.3	میزان درستی	Traces collected at an online game server in Australia and provided by University of Twente, Netherland
[۵۵]	K-Means	تعداد کل بسته‌ها، میانگین طول بسته‌ها، میانگین سائز مربوط به داده هر بسته، مقدار بایتی که انتقال داده شده، مدت زمان جریان، میانگین مدت زمان مابین ارسال بسته‌ها	Web, P2P, FTP, Others	95	میزان درستی	Self-collected 81-hour campus traces between April 6-9, 2006

مقاله مرجع	روش شناسایی استفاده شده	ویژگی‌های مورد استفاده	برنامه‌های کاربردی	دقت شناسایی (درصد)	معیار ارزیابی	پایگاه داده
[۱۳]	Protocol fingerprints and Anomaly score	اندازه بسته‌ها، مدت زمان مابین ارسال بسته‌ها، ترتیب بسته‌ها	TCP applications (HTTP, SMTP, POP3, SSH)	96.10	میزان درستی	6-month selfcol-lected traces at the edge gateway of the University of Brescia data centre network
[۵۶]	Bayesian Neural Network	مدت زمان جریان، تعداد بسته‌ها، تعداد بایت ارسال شده، مدت زمان مابین ارسال بسته‌ها، اندازه فیلد کنترل لایه انتقال، تعداد بسته‌ها برای ارسال و دریافت، سائز بسته‌ها، پهنای باند موثر مورد استفاده، تبدیل فوری برای زمان ارسال ده بسته، مقدار بایت داده ارسال شده، تعداد بسته‌های تایید تحویل	A large range of Database, P2P, Bulck, Mail, Services, Multimedia, Web	96.3	میزان درستی	Proprietary hand classified traces
[۳۰]	Naive Bayes with Discretisation, Naive Bayes with Kernel Estimation, C4.5 Decision Tree, Bayesian Network and Naive Bayes Tree	پروتکل، مدت زمان جریان، حجم جریان به صورت بایت، اندازه بسته (کمترین، بیشترین، میانگین و انحراف معیار)، مدت زمان مابین ارسال بسته‌ها (کمترین، بیشترین، میانگین، انحراف معیار)	FTP(data), Telnet, SMTP, DNS, HTTP	94.13	میزان درستی	NLANR
[۵۷]	Naive Bayes and AutoClass	تعداد بسته‌ها، میانگین طول بسته‌ها، مدت زمان جریان، میانگین اندازه قسمت داده بسته‌ها، میانگین مدت زمان مابین ارسال بسته‌ها	HTTP, SMTP, DNS, SOCKS, FTP(control), FTP (data), POP3, Limewire	93.2	میزان درستی	NLANR

مقاله مرجع	روش شناسایی استفاده شده	ویژگی های مورد استفاده	برنامه های کاربردی	دقت شناسایی (درصد)	معیار ارزیابی	پایگاه داده
[۵۹]	C4.5 Decision tree algorithm	تعداد بسته ها (ارسال و دریافت)، میزان بایت (ارسال و دریافت)، اندازه بسته (کمترین، بیشترین، میانگین، انحراف معیار)، مدت زمان ارسال بسته ها (کمترین، بیشترین، میانگین، انحراف معیار)	torrent, dns, http, smtp, ssh, ssl	90	میزان درستی	UTM campus network
[۶۰]	ZeroR, PART, DecisionS-tump, J48, J48graft, LADTree, NBTree, RandomForest, RandomTree and REPTree	N/A	dns, ftp, ssh, mail, http, https, msn	90	میزان درستی و زمان پردازش	UTM campus network
[۶۱]	Bayes (Naïve bayes and bayes net), SVM, C4.5, Kmeans, DBSCAN	Fast Correlation-Based Filter algorithm	http, p2p, imap, pop3, smtp, mysql	90	میزان درستی و زمان پردازش	The Auckland dataset
[۶۲]	C4.5	N/A	N/A	95	میزان درستی و زمان پردازش	UTM campus network
[۶۳]	Bayesian Network, C4.5 and MLP	N/A	p2p, web, service, content delivery	95	میزان درستی و زمان پردازش و عملکرد	National Academic Network of Turkey
[۶۴]	RBF, Bayes Net and C4.5	N/A	p2p	76.67	میزان درستی و عملکرد	Proprietary Hand Classified Traces

۴.۲ جمع‌بندی

در فصل اول روش‌های شناسایی ترافیک شبکه به همراه مزایا و معایب هرکدام مطرح گردید و در این فصل با توجه به موضوع رساله روش‌های شناسایی بر پایه ویژگی‌های آماری مورد بررسی قرار گرفت. در ادامه روش‌های ارائه شده در زمینه شناسایی ترافیک شبکه بر پایه ویژگی‌های آماری از سال‌های ۲۰۰۵ تا ۲۰۱۷ مورد بررسی قرار گرفت و برای هر مقاله مدل مورد استفاده به همراه ویژگی‌های استفاده شده مطرح شد. یکی از مسائل حائز اهمیت و قابل توجه در مقالات مروری این امر است که در تمامی مقالات از پایگاه داده‌های استاندارد استفاده شده و در هیچ کدام از این مقالات عوامل تاثیر گذار بر ویژگی‌های آماری همچون: بروزرسانی برنامه‌های کاربردی و اتلاف بسته‌ها در نظر گرفته نشده است.

فصل سوم

تحلیل رفتاری ترافیک شبکه برای شناسایی برنامه‌های کاربردی

۱.۳ مقدمه

در فصل اول چالش‌های موجود در زمینه شناسایی ترافیک شبکه شرح داده شد و مطرح گردید که در این رساله هدف اصلی ارائه ویژگی‌های آماری مقاوم در برابر بروزرسانی برنامه‌های کاربردی و رخداد اتلاف بسته‌ها، جهت شناسایی موثر برنامه‌های کاربردی می‌باشد. به همین منظور در این فصل ابتدا رفتار برنامه‌های کاربردی مورد تحلیل و بررسی قرار گرفته و توسط این تحلیل رفتاری، ویژگی‌هایی مقاوم در برابر بروزرسانی برنامه‌های کاربردی استخراج می‌گردد. سپس به موضوع اثرات رخداد اتلاف بسته‌ها بر روی دقت شناسایی ترافیک شبکه پرداخته شده و با استفاده از فراوانی توزیع ویژگی‌های ارائه شده توسط تحلیل رفتاری، ویژگی‌های مرتبه سومی جهت مواجهه با رخداد اتلاف بسته‌ها ارائه می‌شود.

۲.۳ تحلیل رفتاری

هنگامی که یک برنامه کاربردی بر روی بستر شبکه اجرا می‌شود، با توجه به رفتار تکنیکی آن برنامه کاربردی، مجموعه‌ای از فعالیت‌ها پردازش و اجرا می‌گردد. این فعالیت‌ها همچون فرآیند ساخت، ارسال و دریافت بسته‌ها، جهت ارتباط و تعامل بین کاربران و سرویس دهنده‌ها تعریف و اجرا می‌گردند. جهت تحلیل رفتاری یک برنامه کاربردی، بررسی و ارزیابی فعالیت‌ها و تعاملات یک برنامه کاربردی در شبکه واقعی امری ضروری می‌باشد.

تحلیل رفتاری برنامه کاربردی (ABA^۱)، را می‌توان به دو صورت تحلیل ایستا و تحلیل پویا^۲ انجام داد، که هر کدام از این دو روش، رفتار برنامه‌های کاربردی را از جنبه‌های متفاوتی مورد بررسی و ارزیابی قرار می‌دهند (شکل (۱.۳)). در شکل ۱.۳ ابر واژه‌ای در خصوص تحلیل رفتاری برنامه‌های کاربردی رسم شده است و هدف آن نشان دادن میزان اهمیت تحلیل رفتاری برنامه کاربردی و سایر مفاهیم مهم دخیل در تحلیل رفتاری همچون مانیتورینگ می‌باشد. در تکنیک‌های تحلیل ایستان، رفتار و میزان ریسک برنامه کاربردی با تحلیل سورس کد و بدون اجرای برنامه کاربردی انجام شده و در تحلیل پویا، تحلیل رفتاری برنامه کاربردی بوسیله اجرای برنامه کاربردی و به ازای تمامی ورودی‌ها و خروجی‌ها و همچنین نحوه اجرای برنامه کاربردی صورت می‌پذیرد و توسط این عملیات می‌توان میزان کارایی^۳ و پایداری^۴ برنامه کاربردی را تحلیل نمود. تحلیل پویا، غالباً حالت‌های مختلف اجرای یک برنامه کاربردی را در نظر گرفته و با توجه به باز خورد^۵های برنامه، توصیفات در خصوص آن ارائه می‌نماید [۶۵، ۶۶].

¹Application Behavior Analysis

²Dynamic

³Efficiency

⁴Robust

⁵Feedback

به همین منظور، در تحلیل رفتاری ترافیک شبکه، مکانیزم انتقال و دریافت بسته‌ها، مدیریت جریان‌ها، مدت زمان ارسال و دریافت بسته‌ها، اندازه بسته‌ها، انواع پروتکل‌ها و میزان پهنای باند مورد استفاده تحلیل و بررسی می‌شود. تحلیل رفتار ترافیک شبکه وابسته به محتوای بسته‌ها نبوده و عملیات تحلیل رفتار در لایه IP در نظر گرفته می‌شود.



شکل ۱.۳: تحلیل رفتاری برنامه کاربردی

۱.۲.۳ خصوصیات رفتاری برنامه کاربردی

اطلاعات ترافیک شبکه را می‌توان از دستگاه‌های متفاوت شبکه بدون تاثیر بر کارایی و سرویس‌دهی شبکه دریافت و مورد تحلیل قرار داد. در این رساله رفتار هفت برنامه کاربردی فایرفاکس، کروم، اینترنت اکسپلورر، اسکایپ، تلگرام، مدیریت دانلود و توییتر با استفاده از تحلیل پویا مورد بررسی و ارزیابی قرار گرفته است. جهت تحلیل رفتار یک برنامه کاربردی خصوصیات آن برنامه اعم از: نحوه ارسال بسته‌ها، مدت زمان ارسال بسته‌ها، تعداد بسته‌های ارسال شده و همچنین مکانیزم‌های اجرایی همچون: نحوه ارسال بسته‌ها بررسی می‌گردد.

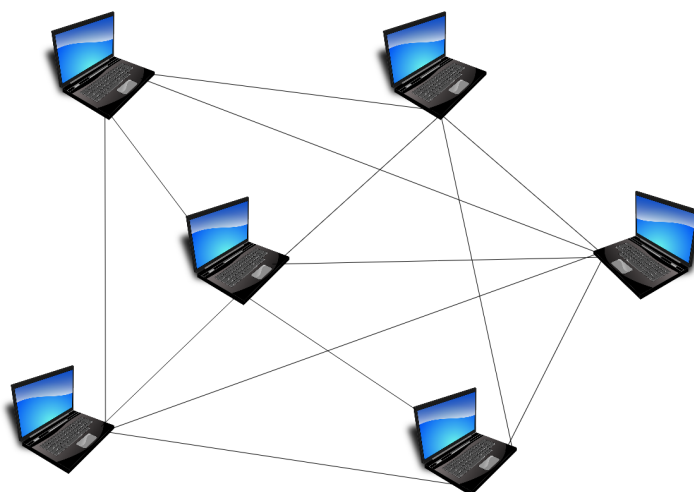
نکته



ذکر این نکته حائز اهمیت است که به دلیل تمرکز این رساله بر شناسایی ترافیک در لایه شبکه و عدم بررسی محتوای بسته‌ها، جهت تحلیل رفتار یک برنامه کاربردی خصوصیات و رفتار آن برنامه در لایه شبکه مورد بررسی قرار گرفته است.

در مقالات و کتب علمی، برنامه‌های کاربردی از لحاظ نوع عملکرد و سازوکار ارسال و دریافت بسته‌ها در

شبکه، به دو دسته برنامه‌های کاربردی P2P و برنامه‌های کاربردی (مشارتی^۶ - خدمت‌گزار^۷) تقسیم‌بندی می‌گردد. برنامه‌های کاربردی P2P دارای معماری توزیع‌شده بوده که حجم کاری و وظایف را در بین همتا^۸ها تقسیم‌بندی می‌نماید. در این برنامه‌ها ارتباطات به صورت غیرمترکز بوده که در آن همتاها دارای قابلیت‌های یکسانی هستند و می‌توانند نقش مشتری و خدمت‌گزار را داشته باشند (شکل ۲.۳). در طرف دیگر برنامه‌های کاربردی (مشارتی - خدمت‌گزار) قرار دارند که در این برنامه‌ها حجم کار در بین ارائه‌دهنده سرویس (خدمت‌گزار) و درخواست‌کننده سرویس (مشارتی) تقسیم می‌گردد. در این شبکه‌ها تمامی درخواست‌ها از طرف مشتری به سمت خدمت‌گزار ارسال شده و خدمت‌گزار پس از بررسی، سرویس‌های مورد درخواست را ارائه می‌دهد. شکل (۳.۳) نمایی از شبکه‌های مشارتی - خدمت‌گزار می‌باشد [۵۵].

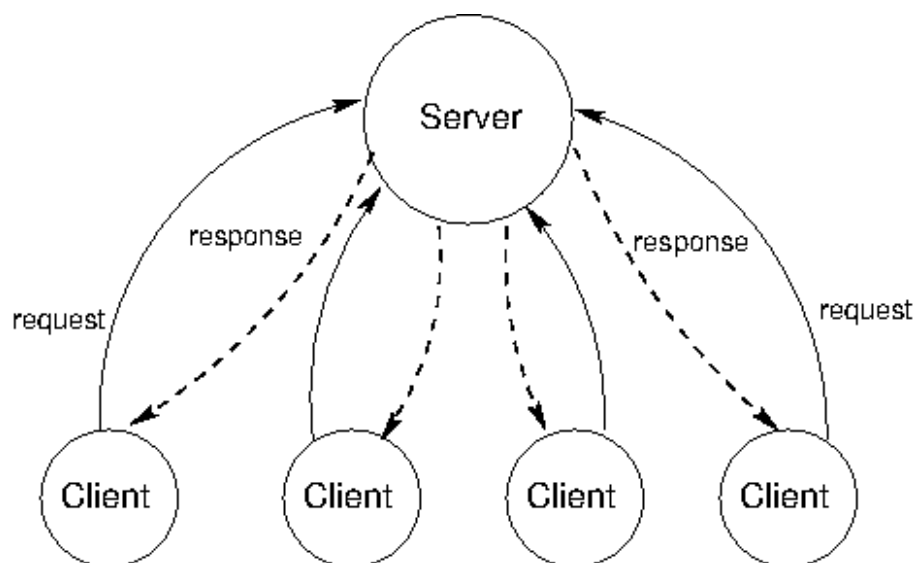


شکل ۲.۳: شبکه Peer2Peer

اسکایپ برنامه کاربردی اسکایپ از مکانیزم برنامه‌های کاربردی P2P استفاده کرده و از رمزنگاری انتها به انتها جهت ارتباط و ارسال و دریافت پیام‌ها استفاده می‌نماید. اسکایپ یک شبکه ارتباط کاربر به کاربر (Peer2Peer) است که سرویس‌های متعدد و همزمانی را از جمله ارتباط صوتی، ارتباط تصویری، ارسال و دریافت فایل و تبادل پیام ارائه می‌دهد و همچنین از پروتکل TCP^۹ جهت ارتباط با نودهای اصلی شبکه به منظور شناسایی، تبادل کلید نشست و تایید هویت استفاده کرده و از پروتکل UDP برای شناسایی

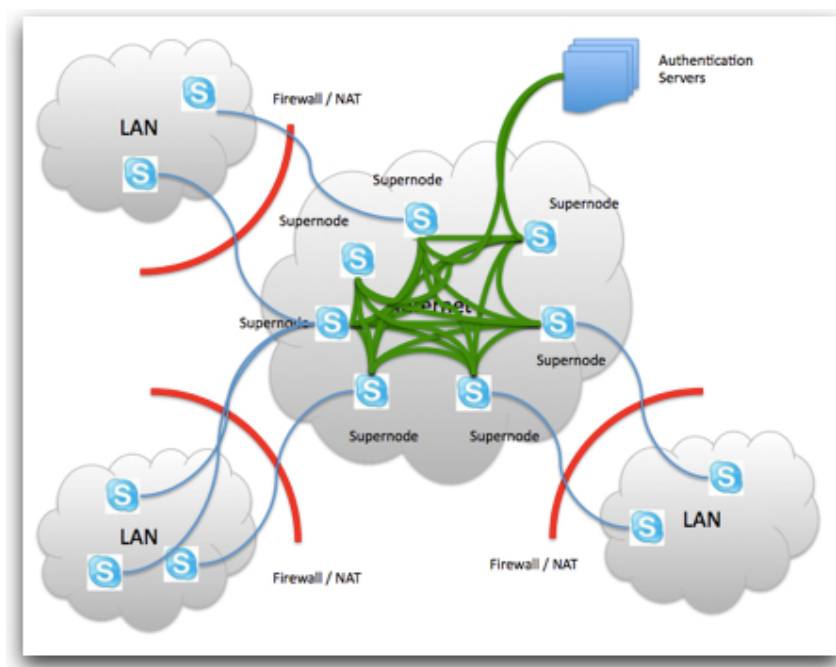
^۶Client
^۷Server

^۸Peer
^۹Transmission Control Protocol



شکل ۳.۳: شبکه مشتری - خدمت گذار

شرایط شبکه جهت تبادل برخط داده و بررسی نمودن وجود کاربر یا دیوار آتش در ¹⁰NAT شبکه استفاده می نماید. تمامی ارتباطات در شبکه اسکایپ از قوانین HTTPS تبعیت می کند [۶۷]. شکل (۴.۳) نمایی کلی از ساختار شبکه اسکایپ را نشان می دهد.

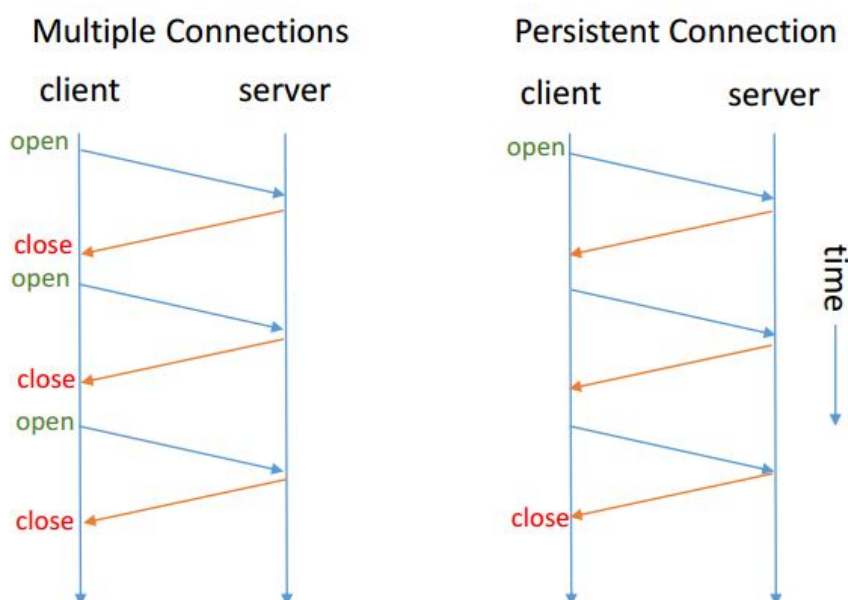


شکل ۴.۳: شبکه اسکایپ

¹⁰Network Address Translation

مرورگر وب مرورگر وب به نوعی برنامه کاربردی گفته می‌شود که برای دریافت، نمایش و مرور اطلاعات از وب مورد استفاده قرار می‌گیرد. یک منبع اطلاعات با یک URL شناخته می‌شود که می‌تواند یک صفحه وب، تصویر، ویدئو یا هر نوع اطلاعاتی دیگر باشد. در ویرایش قدیمی مرورگرها از اتصال چندگانه استفاده شده است. در حالی که ویرایش امروزه این مرورگرها از اتصال مداوم^{۱۱} استفاده نموده و پس از برقراری یک اتصال TCP، چندین درخواست و جواب HTTP را ارسال و دریافت می‌نماید. در صورتی که در نسخه‌های قدیمی برای هر درخواست یا جواب یک اتصال TCP برقرار می‌گردید. شکل (۵.۳) تفاوت اتصال چندگانه و اتصال مداوم را نشان می‌دهد.

تمامی مرورگرهای امروزی اعم از گوگل کروم، فایرفاکس، اینترنت اکسپلورر (نسخه ۵ به بعد)، اپرا (نسخه

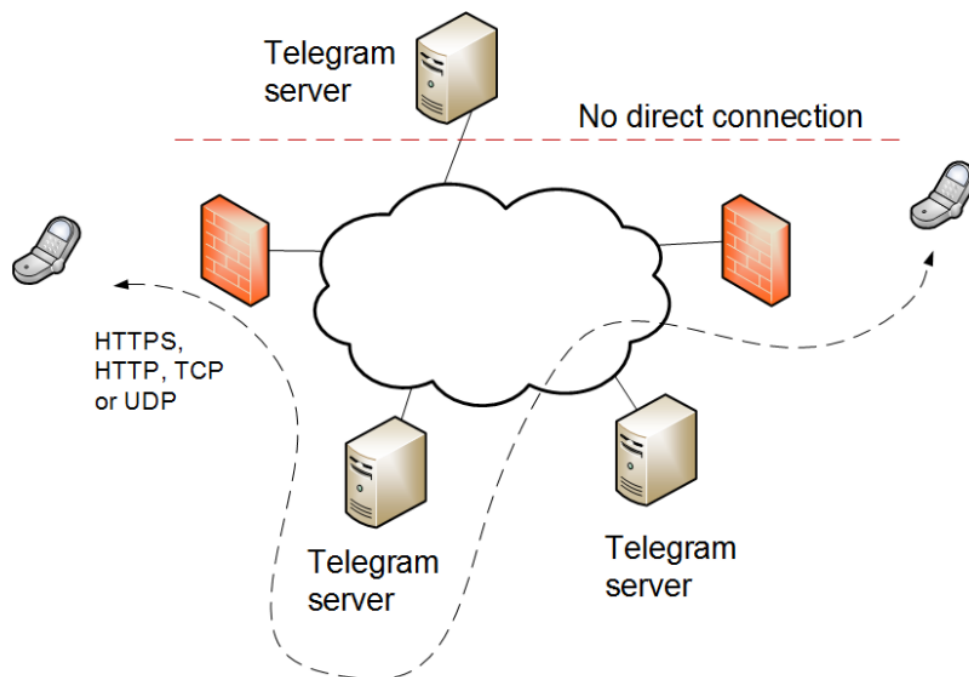


شکل ۵.۳: پروتکل TCP در اتصال مداوم و چند اتصالی

۴ به بعد) و سافاری از مدل اتصال مداوم استفاده می‌کنند. به طور پیش فرض اینترنت اکسپلورر در ویرایش شش و هفت از دو اتصال مداوم و در ویرایش‌های هشت به بعد از شش اتصال مداوم استفاده می‌نماید. مدت زمان تقریبی برای هر اتصال مداوم ۶۰ ثانیه بوده و پس از آن اتصال قطع می‌شود و این زمان را می‌توان تغییر داد. در فایرفاکس، تعداد اتصالات همزمان قابل تنظیم بوده و مدت زمان اتصال مداوم ۱۱۵ ثانیه (که قابل تغییر است) می‌باشد. این امر در گوگل کروم متفاوت بوده و معیار دیگری برای زمان بیکار بودن در نظر گرفته شده است. به این صورت که مرورگر در زمان بیکاری شروع به ارسال بسته‌های Ack به سمت سرور می‌نماید [۶۸].

¹¹Persistent Connection

تلگرام تلگرام یک سرویس پیام‌رسان متن‌باز چندسکویی مبتنی بر رایانش ابری است. رابط کاربری تلگرام برای تمامی چارچوب‌های تلفن همراه (اندروید، آیفون، اوپنتو و ویندوز) و سیستم‌های کامپیوتری ارائه شده است. کاربران در محیط برنامه کاربردی تلگرام قادر به ارسال متن، صوت، تصویر، چند رسانه، پیام و استیکر به صورت رمزشده می‌باشند. همچنین تلگرام مکانیزم رمزنگاری انتها به انتها را اجرا کرده است. نسخه‌های سمت کاربران این برنامه کاربردی کد باز بوده ولی نسخه این برنامه کاربردی در سمت سرویس دهنده قابل دسترس نمی‌باشد.¹² API‌های متفاوتی برای این برنامه کاربردی ارائه شده که هر کدام از این API‌ها به سرویس دهندگان آن متصل شده و عملیات سرویس دهی را انجام می‌دهند. همچنین ارتباطات تلگرام به صورت TCP و UDP بوده و از قوانین HTTPS تبعیت می‌کند. اندازه بسته‌ها در تلگرام کمتر از ۱۰۲۴ یا بیشتر از ۶۴ بایت می‌باشد. شکل (۶.۳) نمایی از ساختار شبکه برنامه کاربردی تلگرام را نشان می‌دهد.



شکل ۶.۳: شبکه تلگرام

مدیریت دانلود مدیریت دانلود، یک برنامه کاربردی است که برای انجام عملیات بارگذاری فایل‌ها مورد استفاده قرار می‌گیرد. این برنامه با مرورگر وب که برای مرور صفحات وب به کار می‌رود تفاوت

¹²Application Programming Interface

دارد. چون جابه‌جایی فایل‌ها میان رایانه و اینترنت، هدف فرعی مرورگر وب است. یک برنامه کاربردی مدیریت دانلود باید وسیله‌ای فراهم نماید تا هنگام رخداد خطاها، بدون اتلاف کاری عملیات دانلود را به اتمام رسانده، بازیابی نماید، و بتواند اختیار تکه‌تکه کردن فایل دانلودی را به دو بخش یا بیش‌تر به کاربر بدهد. این کار (تکه‌تکه‌سازی) در موازات دانلود انجام شده و به‌طور پنهانی فرآیند دانلود را حتی با وجود محدودیت‌های پهنای باند تسریع می‌کند. برنامه‌های مدیریت دانلود معمولاً یک یا چندتا از ویژگی‌های زیر را دارا می‌باشد:

- توقف موقتی دانلود فایل‌های بزرگ
- دانلود فایل‌ها در اتصال‌های ضعیف
- دانلود چند فایل از یک سایت به‌صورت خودکار با استفاده از قوانین ساده
- دانلودهای برگشتی خودکار (معکوس کردن)
- دانلودهای زمان‌بندی‌شده (شامل قطع خودکار اتصال و خاموش کردن رایانه)
- جستجو برای سایت‌های آینه، و دسترسی به اتصال‌های مختلف برای دانلود سریع‌تر همان فایل (دانلود تکه‌تکه)
- پهنای باند مصرفی قابل تغییر
- ساخت خودکار زیرپوشه

در برنامه کاربردی مدیریت دانلود، تمامی ارتباطات از قوانین TCP تبعیت کرده و توزیع اندازه بسته‌ها به دو قسمت بسته‌های دریافت اطلاعات و بسته‌های تایید دریافت تقسیم می‌شود. همچنین در این برنامه کاربردی تعداد جریان‌ها وابسته به تعداد دانلودهای همزمان بوده و نرخ دانلود (bit/sec) از برنامه‌های کاربردی دیگر (Skype، Telegram و Browsers) بالاتر می‌باشد.

همانطور که مطرح شد، ترافیک برنامه‌های کاربردی را می‌توان به دو دسته ترافیک برنامه‌های P2P و برنامه‌های Client - Server تقسیم‌بندی نمود. به همین منظور برنامه‌های کاربردی نامبرده شده در این دو دسته قرار گرفته و رفتار ترافیک آنها (با توجه به پایگاه داده‌های مطرح شده در فصل اول) با توجه به خصوصیات آماری لایه شبکه آنها مورد تحلیل و بررسی قرار گرفته است. پس از بررسی‌های جامع بر روی این خصوصیات رفتاری، نتایج حاصله به صورت جدول (۱.۳) ارائه شده است. در این جدول خصوصیات رفتاری لایه شبکه دو دسته برنامه‌های کاربردی P2P و Client-Server با یکدیگر مقایسه شده‌اند.

همانطور که مشاهده می‌شود، رفتار برنامه‌های کاربردی P2P و Client-Server در بعضی از خصوصیات

جدول ۱.۳: خصوصیات رفتاری برنامه‌های کاربردی P2P و Client-Server

خصوصیات	ترافیک (P2P)	ترافیک (Web)
میزان بایت ارسالی در جریانها	زیاد	کم
پارامتر خاصیت خودشبیهی	$0.5 >$	$0.8-0.7$
توزیع داده‌ها	Long Tail	On/Off, Heavy Tail
مدت زمان جریانها	$<100 \text{ ms}$	$> 100 \text{ ms}$
نرخ ارسال داده bit/sec	کم	زیاد
زمان بین ارسال و دریافت بسته‌ها	$<0.3 \text{ ms}$	$> 0.3 \text{ ms}$
کمترین زمان بین ارسال و دریافت بسته‌ها	0	0
بیشترین زمان بین ارسال و دریافت بسته‌ها	25-200 ms	25-200 ms
میانگین زمان بین ارسال و دریافت بسته‌ها	50-250 ms	0-100 ms
انحراف معیار زمان بین ارسال و دریافت بسته‌ها	50-150 ms	0-50 ms
اندازه بسته‌ها	$> 800 \text{ byte}$	$< 800 \text{ byte}$
کمترین اندازه بسته‌ها	60 byte	60 byte
بیشترین اندازه بسته‌ها	1514 byte	1514 byte
میانگین اندازه بسته‌ها	$10^5 - 6 * 10^5 \text{ byte}$	10^5 byte
انحراف معیار اندازه بسته‌ها	$<6 * 10^4 \text{ byte}$	$>2 * 10^4 \text{ byte}$
آنتروپی	زیاد	کم
نرخ ارسال بسته packet/sec	$3 * 10^5 >$	$2 * 10^5 <$
تعداد جریانها	1500 - 5500	500 - 4000
اندازه پنجره ارسال	$10^4 2.5 * 10^4 \text{ byte}$	$10^4 2.5 * 10^4 \text{ byte}$
کمترین اندازه پنجره ارسال	0 - 1 byte	0 - 1 byte
بیشترین اندازه پنجره ارسال	$< 4 * 10^4 \text{ byte}$	$< 4 * 10^4 \text{ byte}$
میانگین اندازه پنجره ارسال	$10^6 2 * 10^6 \text{ byte}$	$10^6 6 * 10^6 \text{ byte}$
انحراف معیار اندازه پنجره ارسال	010 ⁶ byte	$10^6 3 * 10^6 \text{ byte}$

(مانند اندازه بسته‌ها) کاملاً متفاوت بوده، و در بعضی از خصوصیات (مانند اندازه پنجره ارسال) تفاوت چندانی ندارد. نکته دیگر اینکه رفتار برنامه‌های کاربردی در دسته‌های P2P و Client-Server به یکدیگر شبیه می‌باشد.

نکته

پس از بررسی‌های به عمل آمده چهار ویژگی آماری اندازه بسته‌ها، مدت زمان مابین ارسال و دریافت بسته‌ها، مدت زمان جریانها و نرخ ارسال و دریافت بسته‌ها دارای بیشترین قابلیت تمایز دهنده در بین برنامه‌های کاربردی بوده و همچنین رفتار این ویژگی‌ها در ویرایش‌های مختلف یک برنامه کاربردی ثابت می‌باشد.



در فصل بعد ارزیابی این ویژگی‌های آماری استخراج شده با استفاده از اطلاعات متقابل مورد بررسی قرار گرفته و با استفاده از تکنیک‌های یادگیری ماشین به شناسایی برنامه‌های کاربردی با استفاده از این ویژگی‌ها پرداخته می‌شود.

۳.۳ اطلاعات متقابل

۱.۳.۳ تعریف مفهوم اطلاعات

هوای قطب شمال امروز سرد است. تقریباً این سخن اطلاعاتی را در بر ندارد، چون امر محتملی است. تا دو هفته دیگر زلزله شدیدی در تهران رخ خواهد داد. این سخن اطلاعات بسیار زیادی را در بر دارد، زیرا انتظار چنین رخدادی در ذهن وجود ندارد. در یک نتیجه‌گیری کلی می‌توان گفت که هر چه احتمال وقوع پیام کمتر باشد، آن پیام دارای اطلاعات بیشتری است. از دیدگاه دیگر هر چقدر ابهام برای رویداد یک اتفاق زیادتر باشد، آنگاه آن خبر از وقوع آن اتفاق اطلاعات بیشتری را به ارمغان می‌آورد.

فرض کنید یک منبع^{۱۳} اطلاعات رقمی شروع به تولید پیام بکند. هر پیام از اجزایی (حروف، ارقام و به طور کلی سمبل) تشکیل می‌گردد. با فرض اینکه، یک منبع اطلاعات با الفبای $s_1, s_2, \dots, s_N$ در اختیار داشته و هر یک از این N سمبل با یک احتمالی رخ می‌دهند. این احتمال‌ها نیز با $p_1, p_2, \dots, p_N$ نمایش داده شده و سمبل‌ها از یکدیگر مستقل هستند. نرخ تولید سمبل‌ها را نیز برابر Fs در نظر می‌گیریم. با توجه به مطالب بیان شده، سعی شده است که یک منبع اطلاعاتی، به طور کامل توصیف گردد. هر سمبل از منبع اطلاعات دارای مقداری اطلاعات است. میانگین اطلاعات تمامی سمبل‌ها اصطلاحاً إنتروپی آن منبع می‌باشد و مقدار إنتروپی با H نمایش داده می‌شود. اطلاعات یک پیام، تابعی از احتمال وقوع آن است. اما چه تابعی؟ با بیان چهار اصل موضوعه زیر اثبات می‌شود که تابع اطلاعات به چه صورت است. برای سادگی ابتدا حالت گسسته مورد بررسی قرار می‌گیرد، یعنی خروجی منبع اطلاعات دارای مثلاً M حالت است. نتیجه حاصله به سادگی قابل تعمیم به پیوسته نیز است.

توجه



اطلاعات و ابهام دو روی یک سکه معادل یکدیگرند. هر چه یک پدیده اطلاعات بیشتری داشته باشد، در صورت ندانستن آن اطلاعات، ابهام از پدیده بیشتر است.

¹³Source

۲.۳.۳ اصول موضوعه

۱. اگر متغیر تصادفی X دارای توزیع یکنواخت روی M مولفه باشد، $\Pr\{X = x_i\} = \frac{1}{M} \quad \forall 1 \leq i \leq M$ ، آنگاه $H(X)$ تابع افزایشی یکنوا از M است. یعنی $H(X) = f(M)$ اکیداً صعودی است. این اصل بدیهی است چرا که با افزایش تعداد حالات ممکن یک رویداد، ابهام ما و به طور معادل اطلاعات آن نیز بیشتر است. مثلاً در پرتاب تاس و سکه، ابهام بیشتری در مورد تاس وجود دارد.
۲. اگر منبع X دارای M حالت و منبع Y دارای L حالت باشد و دو منبع کاملاً مستقل و با توزیع یکنواخت باشند، اطلاعات کل دو منبع برابر مجموع اطلاعات تک تک منابع است. یعنی:

$$f(ML) = f(M) + f(L) \quad (۱.۳)$$

این اصل به طور شهودی قابل قبول است. به عنوان مثال فرض کنید یک سکه و تاس را به طور مستقل پرتاب می‌کنیم. قبل از مشاهده ابهامی به اندازه $f(12)$ داریم ($2 \times 6 = 12$) پس از دیدن سکه به اندازه ابهام ما از نتیجه پرتاب سکه ($f(2)$) کم می‌شود و تنها به اندازه نتیجه پرتاب تاس ($f(6)$) ابهام وجود دارد. یعنی $f(12) - f(2) = f(6)$.

۳. اصل دسته بندی: مسابقه بیست سوالی را در نظر بگیرید. فرض کنید تعداد کل پاسخ‌هایی که به یک سوال می‌توان داد M تا است. اگر تابع اطلاعات را بر حسب احتمالات تک تک پاسخ‌های ممکن به صورت $H(p_1, p_2, \dots, p_M)$ نشان داده و کل پاسخ‌ها قابل تفکیک به دو گروه (مثلاً جاندار و بی‌جان) باشد به طوری که پاسخ‌های شماره ۱ تا r در گروه A و بقیه پاسخ‌ها ($r+1$ تا M) در گروه B باشد، می‌توان نتیجه گرفت:

$$H(p_1, p_2, \dots, p_M) = H(p_A, p_B) + p_A H\left(\frac{p_1}{p_A}, \dots, \frac{p_r}{p_A}\right) + p_B H\left(\frac{p_{r+1}}{p_B}, \dots, \frac{p_M}{p_B}\right) \quad (۲.۳)$$

به طوری که $p_B = \sum_{i=r+1}^M p_i$ و $p_A = \sum_{i=1}^r p_i$ این اصل نیز با کمی تامل معقول و قابل استدلال است. ابهام در مورد جواب برابر ابهام از این که پاسخ در کدام گروه است به علاوه میانگین وزنی ابهام در مورد پاسخ‌های محتمل در هر گروه است. واضحاً رابطه فوق قابل تعمیم با گروه‌های بیشتری نیز هست.

۴. اصل پیوستگی: با فرض حالت باینری، $H(p, 1-p)$ تابعی پیوسته از p است. مثلاً فرض کنید احتمال بارش در امروز برابر ۰.۲۵ است. اگر این احتمال کمی تغییر کند مثلاً ۰.۲۵۰۱ گردد، واضحاً نباید تغییرات چشمگیری در اطلاعات موجود در این گزاره و ابهام ایجاد شود و این همان پیوستگی

است.

۳.۳.۳ انتروپی

به کمک چهار اصل فوق ثابت می‌شود تنها تابع ریاضی که اصول منطقی فوق را ارضاء می‌کند، تابع لگاریتمی است. این کمیت به صورت رابطه (۳.۳) تعریف می‌شود.

$$H(X) = H(p_1, p_2, \dots, p_M) \triangleq \sum_{i=1}^M -p_i \log_{\alpha} p_i \quad (3.3)$$

رابطه (۳.۳)، میزان اطلاعات خروجی X (که هر رویداد آن با احتمال p_i رخ می‌دهد)، را ارائه می‌دهد. از آنجایی که رابطه فوق میانگین وزنی اطلاعات تک تک رویدادها است، اطلاعات هر رویدادی مثل A با $I(A) = -\log_{\alpha} p_A$ نشان داده می‌شود. معمولاً تابع لگاریتم موجود در رابطه (۳.۳)، در مبنای دو ($\alpha = 2$) در نظر گرفته می‌شود. در این صورت واحد اطلاعات بیت نامیده می‌شود. البته دقت شود که این بیت با بیت به معنای رقم باینری اشتباه نشود.

واحد انتروپی که میانگین اطلاعات تک تک رویدادهای ممکن است بر حسب بیت بر سمبل است. مفهوم انتروپی مفهوم بسیار پرکاربردی است. انتروپی بیانگر این موضوع است که، هر سمبل خارج شده از منبع به طور متوسط دارای چه میزان اطلاعات است. منابع بایستی به گونه‌ای باشند، که انتروپی بیشتری داشته باشند. بیشینه انتروپی وقتی حاصل می‌شود که احتمال صدور سمبل‌ها با یکدیگر برابر باشند. در رابطه انتروپی^{۱۴} مشخصاً با فرض یکنواخت بودن داریم:

$$H(X) = f(M) = \log M \quad (4.3)$$

۴.۳.۳ انتروپی مشترک، شرطی و اطلاعات متقابل

اگر دو منبع خروجی‌های متناظر X و Y با توزیع احتمال p_{x_i} و p_{y_j} داشته باشند، می‌توان یک منبع مرکب فرض کرد که دارای خروجی (X, Y) و توزیع احتمال $p(x_i, y_j)$ است.

تعریف ۱.۳ انتروپی مشترک به صورت رابطه (۵.۳) تعریف می‌شود.

$$H(X, Y) = - \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) \quad (5.3)$$

¹⁴Entropy

اگر خروجی Y از یک منبع مشاهده شده برابر y_j باشد، ابهام از خروجی منبع دیگر که X است مشخصا با تابع احتمال آن به شرط دانستن y_j یعنی $p(x_i|Y = y_j)$ رابطه داشته و به صورت رابطه (۶.۳) تعریف می شود.

$$H(X|Y = y_j) = \sum_i p(x_i|Y = y_j) \log p(x_i|Y = y_j) \quad (۶.۳)$$

تعریف ۲.۳ **انتروپی شرطی** : ابهام از رویداد X به شرط دانستن رویداد Y برابر است با :

$$H(X|Y) = \sum_j H(X|Y = y_j) \quad (۷.۳)$$

تعریف ۳.۳ **اطلاعات متقابل** برابر اطلاعاتی است که X در مورد Y می دهد و به صورت اطلاعات متقابل یک معیار در تئوری اطلاعات بوده و از موارد استفاده آن می توان به استخراج ویژگی اشاره نمود [۶۹]. اطلاعات متقابل دو متغیر تصادفی X و Y یک معیار وابستگی متقابل متغیرها می باشد. در تئوری اطلاعات، اطلاعات متقابل به صورت رابطه (۸.۳) تعریف می شود [۳].

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \\ &= H(X, Y) - H(X|Y) - H(Y|X), \end{aligned} \quad (۸.۳)$$

آنتروپی و آنتروپی مشترک با توجه به تئوری شانون به صورت رابطه های (۹.۳، ۱۰.۳ و ۱۱.۳) تعریف می شود.

$$H(X) = - \sum_{x \in X} p(x) \log(p(x)) \quad (۹.۳)$$

$$H(Y) = - \sum_{y \in Y} p(y) \log(p(y)) \quad (۱۰.۳)$$

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log(p(x, y)), \quad (۱۱.۳)$$

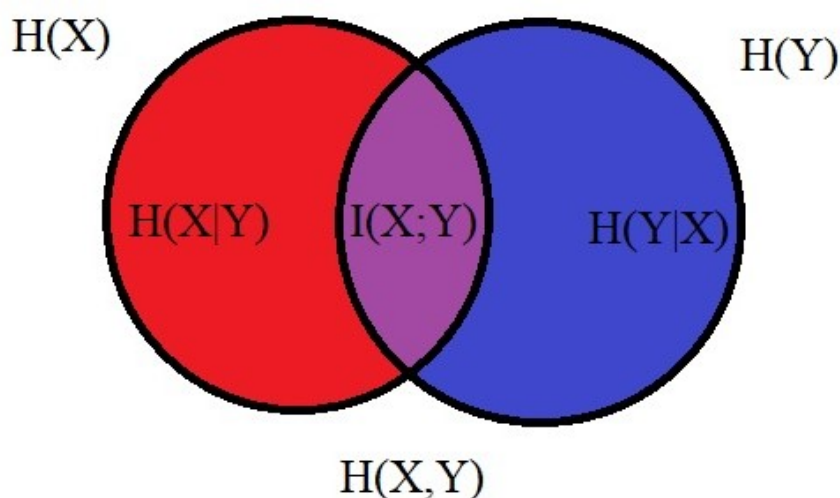
که $p(\cdot)$ توزیع احتمالاتی یک متغیر تصادفی می باشد. در نتیجه تعریف اطلاعات متقابل برای متغیرهای تصادفی گسسته به صورت رابطه (۱۲.۳) می باشد.

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (12.3)$$

نکته



برای به یاد داشتن روابط انتروپی دیاگرام زیر بسیار مفید است.



شکل ۷.۳: روابط بین انتروپی ها

۵.۳.۳ خواص

خواص بسیاری در مورد روابط انتروپی برقرار است. در اینجا به اهم آنها پرداخته شده است:  تابع انتروپی نامنفی است.

$$H(X) \geq 0 \quad (13.3)$$

 بیشترین میزان تابع انتروپی به ازای توزیع یکنواخت است.

$$H(X) \leq \log M \quad (14.3)$$

 حالت تساوی تنها در زمان استقلال دو متغیر تصادفی است.

$$H(X|Y) \leq H(X) \quad (15.3)$$

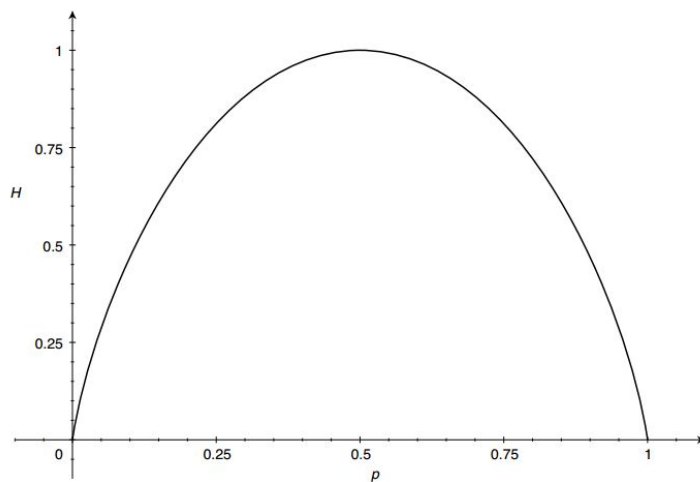
تابع انتروپی مشترک قابل تعمیم به n متغیر تصادفی است و داریم:

$$H(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n H(X_i) \quad (۱۶.۳)$$

حالت تساوی تنها در زمان استقلال تک تک متغیرهای تصادفی است.

$H(X)$ تابعی محدب از توزیع احتمالات است. و به طور مشخص برای متغیر باینری داریم:

$$H(X) = H(p, 1-p) = -p \log p - (1-p) \log(1-p) \quad (۱۷.۳)$$



شکل ۸.۳: انتروپی متغیر تصادفی باینری

$$I(X; Y) = H(X; Y) - H(X|Y) - H(Y|X) \quad (۱۸.۳)$$

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_1, \dots, X_{i-1}) \quad (۱۹.۳)$$

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z). \quad (۲۰.۳)$$

$$I(X_1, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_1, \dots, X_{i-1}). \quad (۲۱.۳)$$



در این رساله از اطلاعات متقابل جهت بررسی میزان اطلاعات و تمایزدهندگی ویژگی‌های استخراج شده نسبت به برنامه‌های کاربردی استفاده شده است. در فصل بعد این معیار برای ویژگی‌های استخراج شده به صورت مجزا و همچنین به صورت مجموعه‌ای مورد بررسی قرار گرفته است.

۴.۳ اتلاف بسته و تاثیر آن بر شناسایی ترافیک

۱.۴.۳ مفهوم کیفیت سرویس

مفهوم کیفیت سرویس همراه با توسعه شبکه‌های ارتباطی، تغییر زیادی نموده است. در روزهای آغازین ایجاد شبکه‌های کامپیوتری، ارسال بسته‌ها از مبدا به مقصد مهمترین و برجسته‌ترین هدف یک شبکه و دسترسی قابل اطمینان به شبکه مساله اصلی در زمینه کیفیت سرویس بود. امروزه با رشد سریع شبکه‌ها، مفاهیمی چون نیاز روزافزون به پهنای باند و پشتیبانی همزمان کلاسهای مختلف سرویس در درجه اول اهمیت قرار گرفته‌اند. در نتیجه کیفیت سرویس به فاکتوری کلیدی در بکارگیری شبکه‌ها و سرویس‌های امروزه تبدیل شده است. با وجود اینکه کیفیت سرویس اخیراً به مساله‌ای مهم در بین محققان شبکه تبدیل شده، هنوز هم ابهاماتی در شیوه فهم و تعریف آن وجود دارد. بطور کلی کیفیت سرویس به معنی ارائه سرویسی سازگار و قابل پیش‌بینی به منظور تأمین نیازهای کاربردهای مختلف است [۷۰].

کیفیت سرویس را از دو دیدگاه می‌توان بررسی نمود: کاربران شبکه و تأمین‌کنندگان سرویس شبکه. هر یک از این دو دارای اهداف کیفیت سرویس متفاوتی هستند. آنچه که یک کاربر شبکه نیاز دارد، دسترسی به حداکثر پهنای باند با کمترین هزینه ممکن است. از طرف دیگر، هدف تأمین‌کننده شبکه به حداکثر رساندن کارایی شبکه همزمان با برطرف کردن نیازهای کیفیت سرویس همه کاربران به صورت همزمان است. پارامترهای متداول کیفیت سرویس که برای توصیف کارایی شبکه بکار می‌روند عبارتند از:

- **پهنای باند (گذردهی)^{۱۵}**

تعداد بایت‌هایی که در یک دوره تناوب خاص از طریق شبکه انتقال می‌یابد.

- **تأخیر**

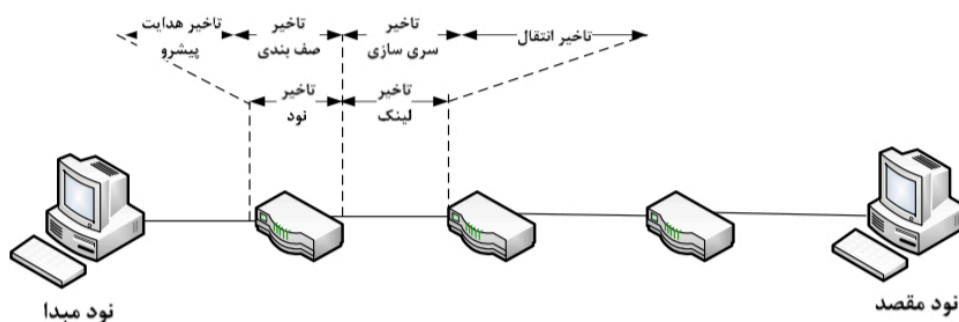
مقدار زمانی که به طور متوسط طول می‌کشد تا یک بسته از یک نقطه شبکه به نقطه دیگر آن

¹⁵Throughput

برود، تاخیر بسته را مشخص می کند. تاخیر انتها به انتهاهای بسته توسط پارامترهایی نظیر: تاخیر هدایت پیشرو^{۱۶}، تاخیر صف بندی، تاخیر انتقال و تاخیر سری سازی، مشخص می شود. در شکل (۹.۳) مثالی از نحوه محاسبه تاخیر آورده شده است.

نکته

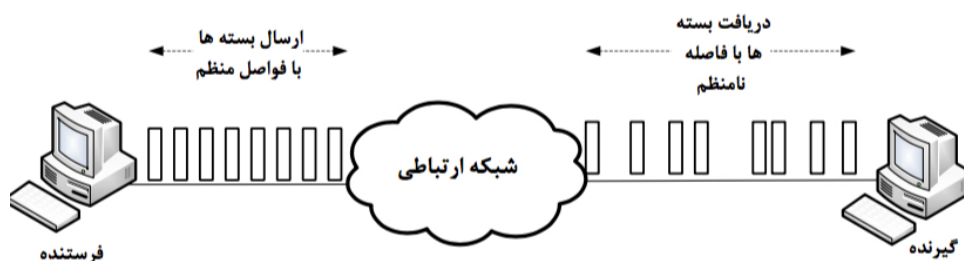
تاخیر هدایت پیشرو نشان دهنده مقدار زمانی است که طول می کشد تا مسیر یاب بسته را دریافت نموده و برای هدایت آن از طریق یک درگاه مناسب تصمیم گیری نماید. به عبارت دیگر تاخیر هدایت پیشرو نشان دهنده حداقل زمان لازم برای پردازش و هدایت یک بسته در مسیر یاب می باشد. مقدار زمان تاخیر فوق در محدوده ۱۰ تا ۱۰۰ میکروثانیه می باشد.



شکل ۹.۳: محاسبه تاخیر انتها به انتها

• واریانس تأخیر

تغییرات تاخیر بسته های متعلق به یک جریان ترافیکی باعث کاهش کیفیت سرویس های زمان حقیقی می شود. شکل (۱۰.۳) مثالی از اثر تغییرات تاخیر بر روی بسته های دریافتی در گیرنده را نشان داده است.



شکل ۱۰.۳: نحوه ایجاد تغییرات تاخیر در بسته های دریافتی در مقصد

¹⁶Forwarding delay

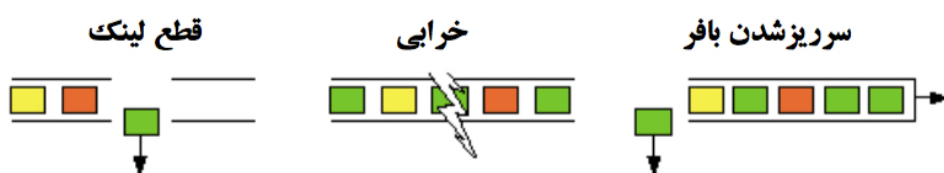


به دلیل استفاده برنامه‌های کاربردی از پروتکل TCP/IP، تغییرات تأخیر تأثیر چشمگیری بر رفتار این برنامه‌ها ندارد.

• اتلاف بسته

در شبکه‌های IP همانطور که در شکل (۱۱.۳) نشان داده شده است، سه عامل مهم باعث ایجاد اتلاف بسته‌ها می‌شوند که عبارتند از:

- قطع شدن کانال ارتباطی که باعث از بین رفتن بسته‌ها می‌شود.
- نویز کانال باعث خراب شدن برخی از فیلدهای بسته می‌شود. نودهای شبکه با بررسی مجموع مقابله‌ای بسته متوجه خراب بودن آن می‌شوند و اقدام به حذف بسته می‌نمایند.
- تراکم در شبکه باعث لبریز شدن بافر نودها و از بین رفتن بسته‌ها می‌شود.



شکل ۱۱.۳: عوامل اتلاف بسته در شبکه‌های کامپیوتری

مشکل قطع شدن کانال ارتباطی به دلیل استفاده از کانال‌های ارتباطی ذخیره در شبکه و همچنین بهره‌گیری از شبکه‌های بی‌سیم، از اهمیت چندانی برخوردار نمی‌باشد. همچنین با توجه به گسترش کانال‌های ارتباطی مطمئن و بدون نویز^{۱۷} نظیر فیبرنوری، احتمال خراب شدن بسته‌ها به خاطر وقوع نویز در کانال بسیار کم و قابل صرف نظر می‌باشد. بنابراین تنها عامل مهم خراب شدن بسته‌ها لبریز شدن بافر نودهای شبکه به خاطر وقوع تراکم در شبکه می‌باشد. از آنجائیکه اکثر ترافیک‌های امروزی از نوع انفجاری می‌باشند، شدت بار لینک‌های شبکه ثابت و از قبل قابل پیش‌بینی نمی‌باشد. چنانچه در نودهای شبکه تراکم بوجود آید، با حذف کردن بسته‌های ورودی از شدت تراکم کاسته می‌شود.

¹⁷Noise



اصلی‌ترین عامل تاثیرگذار بر رفتار برنامه‌های کاربردی رخدادهای اتلاف بسته^۱ می‌باشد.

Packet Loss

یکی از چالش‌های اصلی در روش‌های شناسایی برنامه کاربردی مبتنی بر ویژگی‌های آماری که کمتر در مقالات به آن پرداخته شده است، تاثیرپذیری آنها از رخداد ازدحام بسته‌ها و اتلاف بسته در هنگام انتقال در بستر شبکه می‌باشد [۱۳]. عموماً اتلاف بسته هنگام ازدحام در شبکه و عدم توانایی مسیریاب در رسیدگی به تمامی بسته‌های دریافتی، و دور ریختن آنها از صف، رخ می‌دهد. تحقیقات بسیاری در زمینه جلوگیری از رخداد اتلاف بسته و بازیابی بسته‌ها به جهت افزایش کیفیت سرویس‌دهی انجام شده است [۷۱، ۷۲].

در هنگام وقوع اتلاف بسته، دنباله‌ای از بسته‌ها از دست می‌روند، و در نتیجه ویژگی‌های آماری مورد استفاده جهت شناسایی برنامه کاربردی نیز تغییر می‌کنند. از جمله این تغییرات می‌توان به تغییر زمان ارسال و دریافت بسته‌ها، تغییر در ترتیب بسته‌ها، تغییر در نرخ ارسال و دریافت بسته‌ها و مدت زمان جریان‌ها اشاره نمود. در سال‌های اخیر روش‌های متعددی جهت شناسایی ترافیک شبکه ارائه شده است، ولی بررسی‌های انجام شده نشان می‌دهد که تاکنون راه‌حلی برای موضوع رخداد اتلاف بسته در شناسایی ترافیک شبکه ارائه نشده است و روش‌های ارائه شده جهت شناسایی ترافیک شبکه بر روی پایگاه داده استاندارد (بدون اتلاف بسته) مورد ارزیابی قرار گرفته‌اند [۴، ۵، ۷۳].

با توجه به شایع بودن رخداد اتلاف بسته در شبکه‌های رایانه‌ای، هدف این بخش بررسی رخداد اتلاف بسته، و تاثیر آن بر ویژگی‌های آماری و دقت شناسایی برنامه‌های کاربردی می‌باشد و در ادامه کار روشی جهت استخراج ویژگی‌های آماری مقاوم در برابر رخداد اتلاف بسته ارائه می‌شود. به همین منظور پایگاه داده‌ای از هفت برنامه کاربردی شامل کروم، فایرفاکس، اینترنت اکسپلورر، اسکایپ، تلگرام، مدیریت دانلود و توییتر با حالت‌های مختلف اتلاف بسته فراهم شده است تا تاثیر اتلاف بسته بر روی ویژگی‌ها و دقت شناسایی بررسی شود.

دلیل انتخاب این برنامه‌های کاربردی را می‌توان عمومیت آنها در استفاده‌های روزانه کاربران شبکه دانست و همچنین شناسایی این برنامه‌های کاربردی یکی از نیازهای ضروری مدیران شبکه‌ها می‌باشد. در ادامه کار با تحلیل فراوانی توزیع داده‌ها در چهار ویژگی اندازه بسته‌ها، فاصله زمانی مابین بسته‌های

ارسالی و دریافتی، مدت زمان جریان بسته ها و نرخ ارسال و دریافت بسته ها، ویژگی هایی مقاوم به رخداد اتلاف بسته جهت شناسایی برنامه های کاربردی استخراج می شود. سپس با اعمال این ویژگی ها به یک شبکه عصبی، عمل شناسایی ترافیک شبکه انجام می شود. نتایج حاصل از بررسی روش ارائه شده نشان می دهد که با استفاده از این ویژگی های استخراج شده، دقت شناسایی برنامه های کاربردی هنگام رخداد اتلاف بسته در شبکه نزدیک به دقت شناسایی در حالت عدم رخداد اتلاف بسته می باشد.

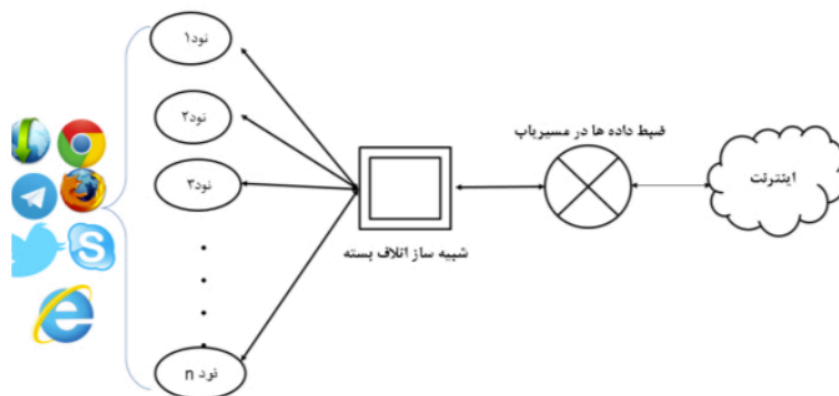
۲.۴.۳ اتلاف بسته ها

در بخش قبل، با استفاده از آنالیز و تحلیل رفتار برنامه های کاربردی، چهار ویژگی آماری (اندازه بسته ها، مدت زمان مابین ارسال و دریافت بسته ها، مدت زمان جریان ها و نرخ ارسال بسته ها) از بسته های ترافیک شبکه استخراج گردید. به منظور ارزیابی این روش، شناسایی ترافیک هفت برنامه کاربردی بر روی دو پایگاه داده UNIBS و پایگاه داده جمع آوری شده با خصوصیت عدم رخداد اتلاف بسته در فصل بعد انجام خواهد گرفت.

همانطور که قبلا اشاره شد، نکته اصلی و چالش برانگیز در این کار و همچنین مقالات اخیر در حوزه شناسایی ترافیک شبکه رخداد اتلاف بسته می باشد. از دست دادن بسته ها هنگامی رخ می دهد که یک یا چند بسته از داده های ارسال شده در بستر شبکه رایانه ای به مقصد نرسد [۷۴]. معمولا اتلاف بسته در اثر ازدحام در شبکه رخ می دهد و توسط پروتکل TCP قابل تشخیص است و عملیات ارسال مجدد بسته ها توسط این پروتکل جهت برقراری قابلیت اطمینان و افزایش کارایی در پیام رسانی انجام می شود. در شبکه ای که با اتلاف بسته همراه است پروتکل TCP ساز و کار ارسال و دریافت بسته ها در شبکه را با استفاده از پنجره ارسال تغییر می دهد و بسته های از دست رفته را مجددا با ترتیب خاصی ارسال می کند. از اینرو در این بخش تاثیر اتلاف بسته بر ویژگی های آماری و دقت شناسایی ترافیک برنامه های کاربردی بررسی می گردد. قدم اول در رسیدن به این هدف تهیه پایگاه داده ای از ترافیک شبکه می باشد که در آن، اتلاف بسته رخ داده باشد.

۱.۲.۴.۳ پایگاه داده

جهت بررسی تاثیرات اتلاف بسته لازم است مجموعه داده ای از ترافیک برنامه های کاربردی مختلف ضبط گردد که در آن اتلاف بسته با نرخ مختلفی رخ داده باشد. از اینرو با استفاده از معماری شکل (۱۲.۳) به شبیه سازی رخداد اتلاف بسته در شبکه پرداخته می شود که در آن برنامه های کاربردی در حال اجرا بوده و بسته های مربوطه در بستر آن ارسال می شوند.



شکل ۱۲.۳: معماری سیستم جهت شبیه سازی رخداد اتلاف بسته و ضبط داده ها

روش کار به این صورت می باشد که سیستم شبیه سازی رخداد اتلاف بسته ما بین ارسال و دریافت بسته ها در شبکه قرار می گیرد و تعداد ۵۰ نود در شبکه به صورت همزمان شروع به ارسال و دریافت بسته های برنامه های کاربردی می نمایند. بسته های ارسالی و دریافتی درون صف مربوط به این سیستم وارد می شوند. در هنگام رسیدگی به بسته های ورودی به صف، پدیده ازدحام رخ می دهد، و بسته ها به صورت کاملاً تصادفی و با نرخ تنظیم شده دور ریخته می شوند. به این ترتیب پروتکل TCP در مبدا ارسال بسته ها در صورتیکه بسته های برگشت را دریافت نکرده باشد، سرعت ارسال را کاهش داده و بسته های ارسالی را مجدداً ارسال می کند. از اینرو در شبکه اتلاف بسته رخ می دهد و بسته هایی که ضبط می شوند تحت تاثیر عملکرد پروتکل TCP جهت ارسال مجدد بسته ها قرار می گیرند.

داده های ضبط شده در اثر رخداد اتلاف بسته در شبکه طی آزمایشات مختلفی برای نرخ های اتلاف بسته صفر، ۱۰، ۲۰، ۲۵، ۳۰، ۴۰، ۵۰ و ۷۰ درصد انجام شده است. جدول (۲.۳) تعداد بسته های ضبط شده از هر برنامه کاربردی با درصد های مختلف اتلاف بسته را نشان می دهد. این برنامه های کاربردی جزئی از اصلی ترین برنامه های مورد استفاده به صورت مستمر می باشند. همچنین سعی شده است تا از گروه های مختلف (نظیر به نظیر، مشتری - خدمت گذار و ابزار دانلود) ترافیک تولید و ضبط گردد.

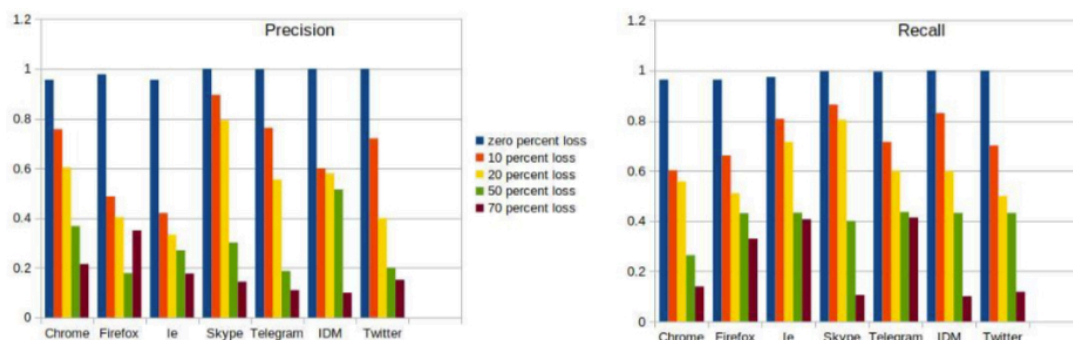
۳.۴.۳ تاثیرات اتلاف بسته بر ویژگی های آماری

پس از تهیه پایگاه داده، عملیات شناسایی برنامه های کاربردی با استفاده از الگوریتم Random Forest انجام گرفت (جزئیات انجام اینکار در فصل بعد آمده است). برای انجام اینکار، داده های فاقد اتلاف بسته جهت آموزش، و داده های با ۱۰، ۲۰، ۵۰ و ۷۰ درصد اتلاف بسته جهت آزمایش، مورد استفاده قرار گرفت. نتایج شناسایی هفت برنامه کاربردی فایرفاکس، کروم، اینترنت اکسپلورر، اسکایپ، تلگرام،

جدول ۲.۳: پایگاه داده جمع آوری شده با رخدادهای مختلف اتلاف بسته

برنامه کاربردی	% صفر	% ۱۰	% ۲۰	% ۲۵	% ۳۰	% ۴۰	% ۵۰	% ۷۰
Chrome	۲۴۵۳۵۹	۲۴۳۰۳۲	۲۳۳۱۲۴	۲۳۲۲۲۴	۳۲۵۹۵۰	۲۳۲۲۸۶	۲۳۵۹۵۰	۲۳۵۵۸۸
Firefox	۲۴۴۰۳۲	۲۱۱۱۲۴	۲۴۱۰۳۲	۲۱۷۶۷۴	۲۷۷۰۳۲	۲۴۰۳۴۱	۲۴۵۰۳۲	۲۳۲۵۸۸
Internet Explorer	۲۹۷۶۷۴	۲۳۱۳۵۹	۲۶۴۰۳۲	۲۳۳۶۷۴	۲۴۳۹۴۱	۲۷۰۵۸۸	۲۳۴۵۸۸	۲۴۹۴۴۲
Skype	۲۱۹۰۳۲	۲۲۹۰۳۲	۲۱۳۳۵۹	۲۴۹۴۳۲	۲۶۰۵۶۶	۲۲۵۶۹۱	۲۲۷۷۹۱	۲۵۵۲۸۶
Telegram	۲۸۷۷۹۱	۲۳۷۷۲۴	۲۳۵۳۵۹	۲۶۲۴۶۶	۲۲۱۷۹۱	۲۰۰۵۸۸	۲۹۲۲۸۶	۲۸۵۹۵۰
IDM	۲۴۱۱۴۱	۲۷۷۴۶۶	۲۳۵۲۲۴	۲۱۸۸۵۹	۲۱۲۷۹۱	۲۹۹۰۳۲	۱۹۲۲۸۶	۲۴۵۵۳۲
Twitter	۲۹۸۹۴۱	۲۵۸۹۵۰	۲۲۲۴۶۶	۲۶۹۱۶۶	۲۲۵۳۵۹	۲۱۳۵۵۹	۳۱۲۲۸۶	۲۱۱۰۳۲

مدیریت دانلود و توئیتر در شکل (۱۳.۳) نشان داده شده است. در شکل (۱۳.۳) نمودار مقادیر دقت

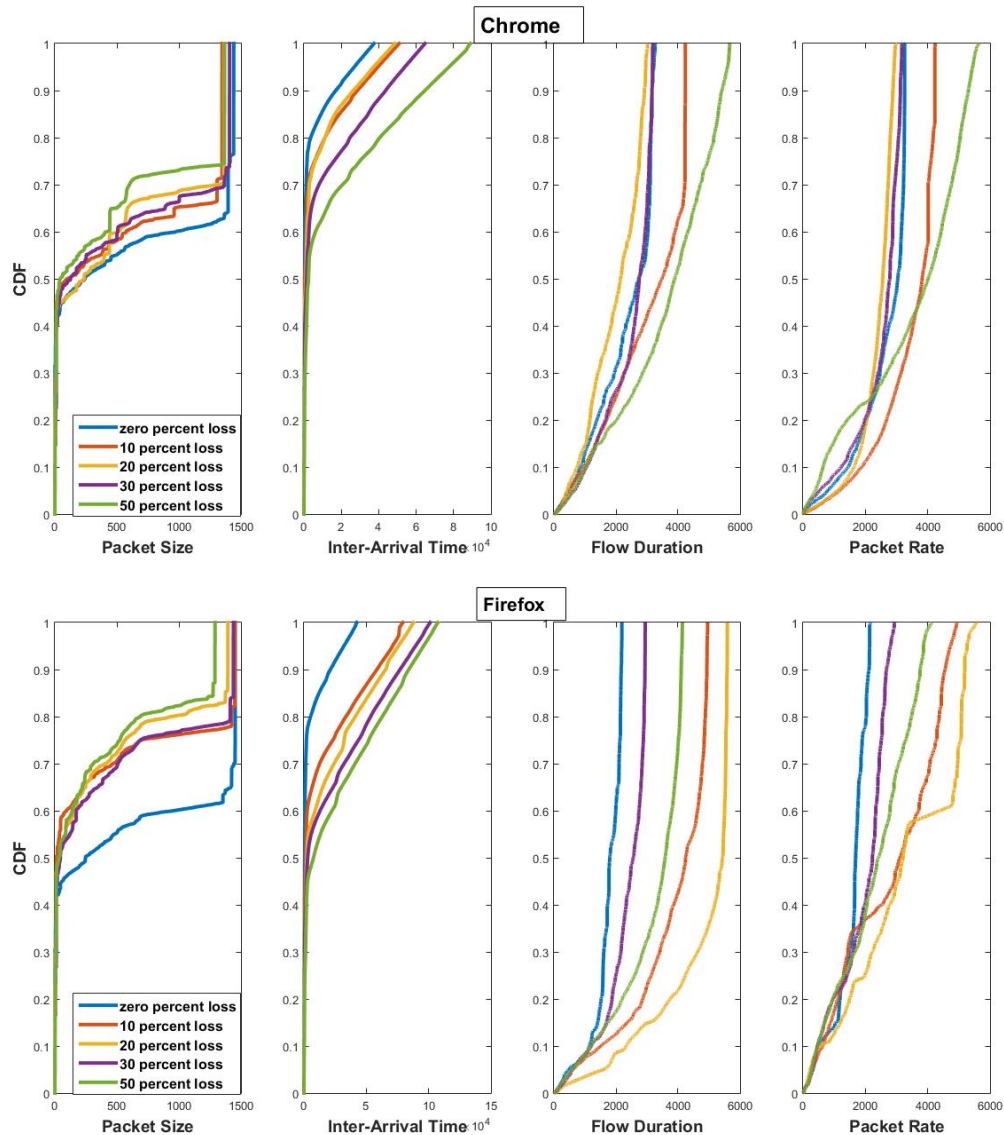


شکل ۱۳.۳: نتایج بدست آمده از شناسایی برنامه‌های کاربردی مختلف از پایگاه داده با درصدهای مختلف اتلاف بسته

و بازخوانی بدست آمده از این آزمایشات نشان داده شده است. همانطور که مشاهده می‌شود دقت شناسایی با رشد نرخ اتلاف بسته کاهش پیدا می‌کند. دلیل این امر را می‌توان اینطور بیان نمود که با از دست دادن بسته‌ها و ارسال مجدد آنها توسط پروتکل TCP، رفتار داده‌ها نیز تغییر می‌کند. برای نشان دادن رفتار داده‌ها و میزان تغییر رفتارهای نرخ‌های مختلف اتلاف بسته از تابع توزیع تجمعی (¹⁸CDF) می‌توان استفاده نمود. نمودار توزیع تجمعی چهار ویژگی اندازه بسته‌ها، مدت زمان مابین ارسال و دریافت بسته‌ها، مدت زمان جریان‌ها و نرخ ارسال بسته‌ها برای هفت برنامه کاربردی با نرخ‌های مختلف اتلاف بسته در شکل (۱۴.۳) نشان داده شده است. در بخش قبل، برنامه‌های کاربردی به دو دسته Client-Server و P2P تقسیم بندی شدند و نشان داده شد که رفتار ویژگی‌های آماری این دو دسته در داخل دسته‌ها به یکدیگر شبیه بوده و این رفتار مابین این دو دسته کاملاً متفاوت می‌باشد ([۷۵]). طبق بررسی‌های به عمل آمده، در برنامه‌های Client-Server هنگامی که میزان رخداد اتلاف بسته از ۵۰ درصد

¹⁸Cumulative Distribution Function

افزایش یابد، کارایی برنامه از بین رفته و صفحه مربوط به برنامه بارگذاری نمی‌شود.



این میزان اتلاف بسته در برنامه‌های P2P به میزان ۳۰ درصد کاهش می‌یابد. به بیانی دیگر در برنامه‌های کاربردی P2P هنگامی که میزان اتلاف بسته به بیش از ۳۰ درصد افزایش یابد پیام‌های به صورت صوت یا تصویر بارگذاری نمی‌شوند. به همین دلیل رفتار برنامه‌های Client-Server در بازه صفر تا ۵۰ درصد اتلاف بسته و برای برنامه‌های P2P بازه رخداد اتلاف بسته بین صفر تا ۳۰ درصد مورد بررسی قرار می‌گیرد.

در شکل (۱۴.۳) توزیع تجمعی داده‌های چهار ویژگی آماری مربوط به بسته‌های هفت برنامه کاربردی نشان داده شده است. همانطور که در این شکل قابل مشاهده است، توزیع داده‌ها در تمامی این ویژگی‌ها همزمان با رشد نرخ اتلاف بسته تغییر کرده و این امر موجب کاهش دقت شناسایی برنامه‌های کاربردی

می‌گردد. به همین منظور در بخش بعد روشی برای استخراج ویژگی‌های مقاوم و موثر در مقابل اتلاف بسته ارائه می‌شود تا بتوان دقت شناسایی برنامه‌های کاربردی در محیط‌های واقعی که عمدتاً با اتلاف بسته همراه هستند را به دقت شناسایی در محیط‌های ایده‌آل نزدیک نمود.

۴.۴.۳ ارائه ویژگی‌های مرتبه سوم

در بخش قبل و در شکل (۱۴.۳) توزیع تجمعی داده‌های چهار ویژگی آماری در حالت‌های مختلف اتلاف بسته برای هفت برنامه کاربردی نشان داده شد. در آن شکل نمودارهای توزیع تجمعی، با روند افزایشی خود از ۰ به ۱، به خوبی رفتار داده‌ها را نشان می‌دهند. برای تحلیل رفتار داده‌ها، فراوانی تابع توزیع تجمعی در بازه‌های کوتاه و یکسان را مورد بررسی قرار می‌دهیم.

در این بخش با بهره‌گیری از این ویژگی فراوانی تابع توزیع تجمعی رفتار داده‌های ترافیکی برنامه‌های کاربردی مختلف مورد بررسی قرار می‌گیرند. در شکل (۱۵.۳) فراوانی تابع توزیع تجمعی داده‌های مربوط به چهار ویژگی اندازه بسته‌ها، مدت زمان مابین ارسال و دریافت بسته‌ها، مدت زمان جریان‌ها و نرخ ارسال بسته‌ها برای هفت برنامه کاربردی با نرخ مختلف اتلاف بسته در بازه‌های کوتاه ۱۰۰ نشان داده شده است. همانطور که مشاهده می‌شود، نمودار فراوانی مقادیر تابع توزیع تجمعی برای هر یک از برنامه‌های کاربردی با نرخ مختلف اتلاف بسته به یکدیگر شبیه هستند، و در برنامه‌های کاربردی مختلف به خوبی قابل تمییز هستند. به عنوان نمونه نمودار توزیع داده‌های برنامه کاربردی Skype برای سه ویژگی اندازه بسته، مدت زمان مابین ارسال و دریافت بسته‌ها و مدت زمان جریان‌ها را در نظر بگیرید. نمودار فراوانی هر یک از این ویژگی‌ها با نرخ مختلف اتلاف بسته شباهت‌های بسیار زیادی با یکدیگر دارند. ولی اگر همین نمودارهای فراوانی را با نمودارهای فراوانی تابع توزیع برنامه کاربردی Chrome مقایسه نماییم، اختلاف زیادی با هم دارند.

قابل ذکر است که بردار ویژگی استخراج شده دارای طول ثابت ۱۰ برای هر یک از چهار ویژگی رفتاری استخراج شده می‌باشد. در نتیجه هر نمونه از برنامه‌های کاربردی دارای چهار بردار ویژگی به طول ۱۰ می‌باشد. در آخر نیز با بکارگیری این چهل ویژگی در شبکه عصبی به راحتی می‌توان برنامه‌های کاربردی مختلف را از هم تمییز داد.

اگر بخواهیم روش ارائه شده در خصوص استخراج ویژگی‌های مقاوم در برابر اثرات رخداد اتلاف بسته را به صورت خلاصه بیان نماییم، مراحل زیر جهت استخراج ویژگی لازم می‌باشد:

ویژگی‌های آماری ترتیب بسته‌ها، اندازه بسته‌ها، فاصله زمانی مابین بسته‌های ارسالی و دریافتی،

مدت زمان جریان‌ها و نرخ ارسال بسته‌ها، با توجه به رفتار برنامه‌های کاربردی استخراج می‌شود. پس از استخراج ویژگی‌های آماری برای هر برنامه کاربردی، بسته‌های هر جریان برنامه کاربردی به عنوان یک دسته در نظر گرفته شده و برای هر دسته بردار مقادیر تابع توزیع تجمعی (CDF) محاسبه می‌شود.

فراوانی CDFها در بازه‌های کوتاه محاسبه می‌شود.

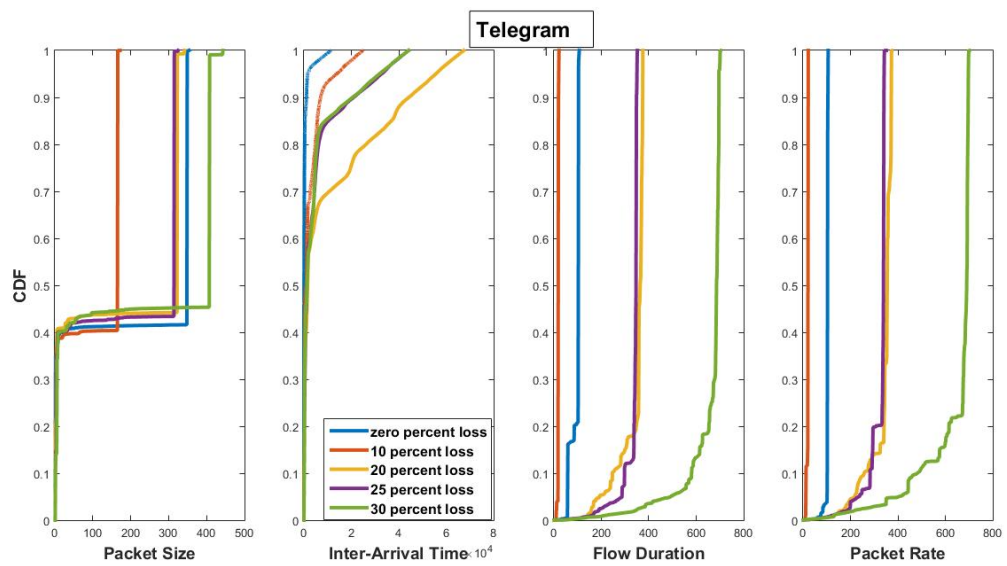
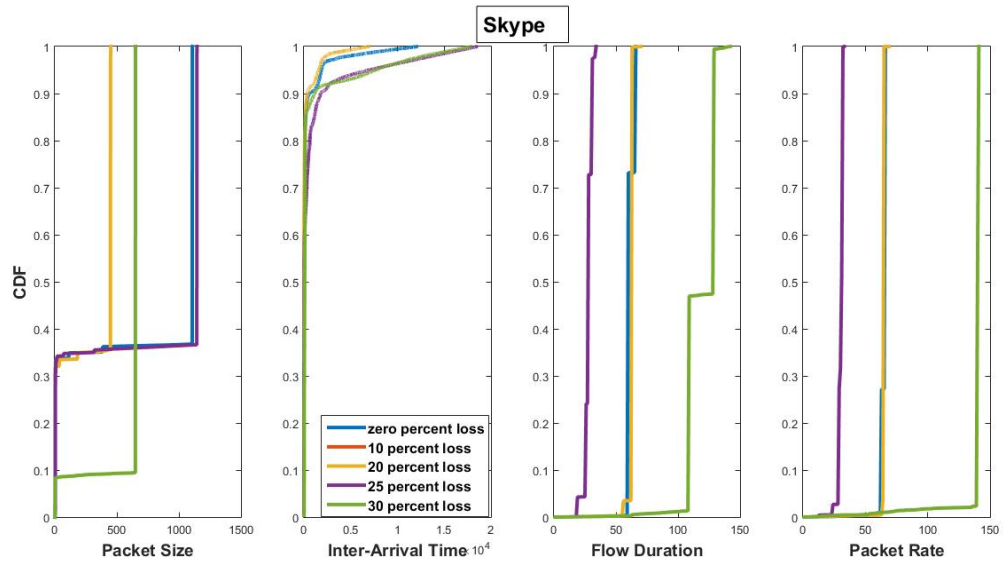
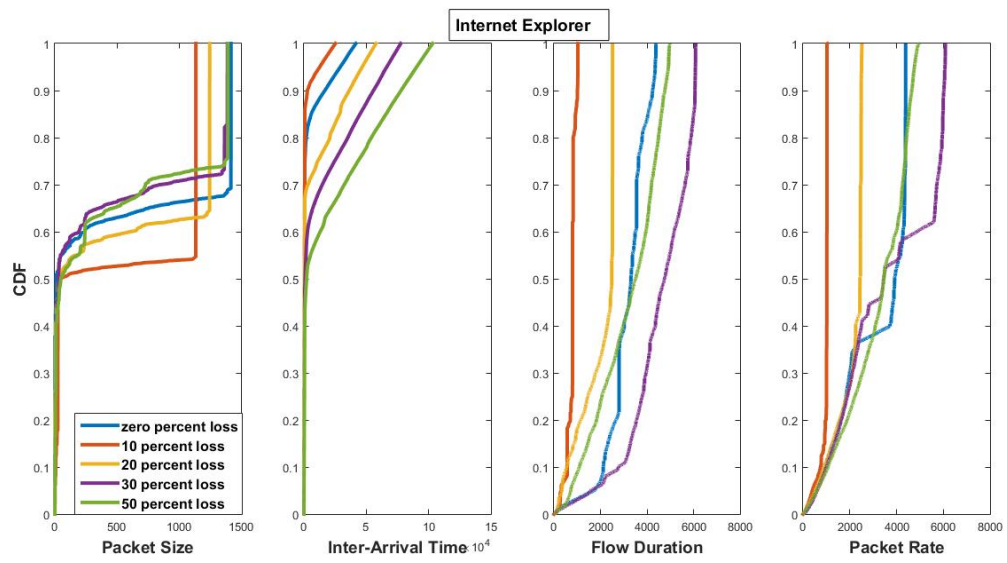
بدین ترتیب برای ترافیک شبکه هر برنامه کاربردی با احتساب تعداد نمونه‌های ورودی، تعدادی بردار ویژگی خواهیم داشت که هر کدام از این بردارها حاوی ۱۰ ویژگی (فراوانی تابع توزیع) می‌باشد.

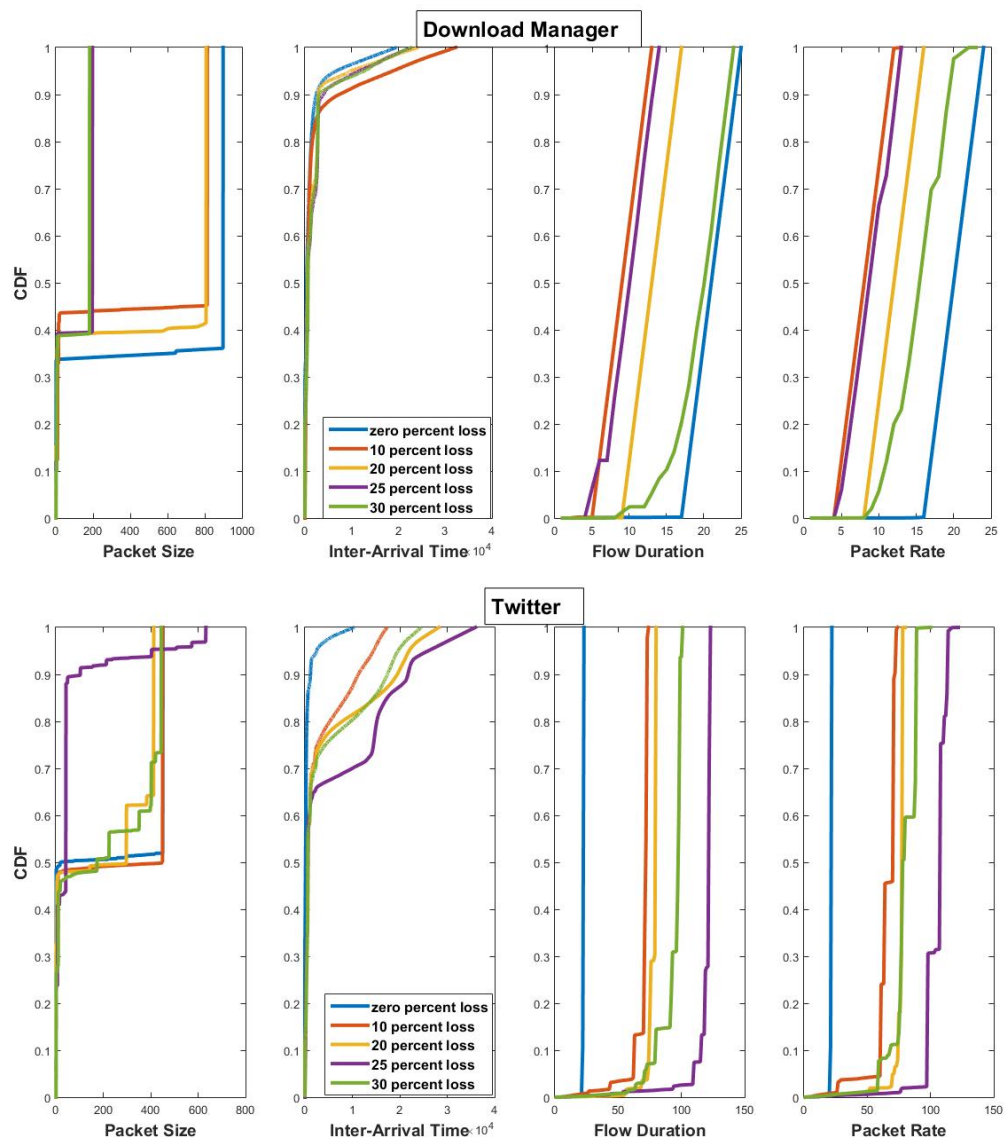
از شبکه عصبی جهت دسته‌بندی ترافیک شبکه استفاده می‌شود.

جزئیات بیشتر در خصوص شبکه عصبی و همچنین نتایج حاصله در فصل بعد شرح داده شده است.

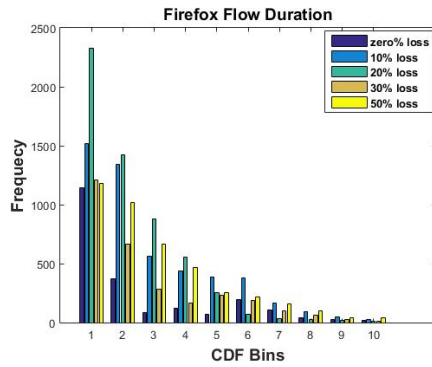
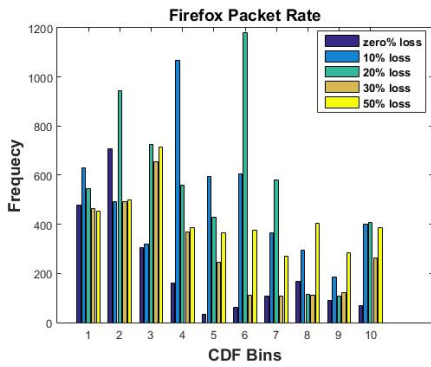
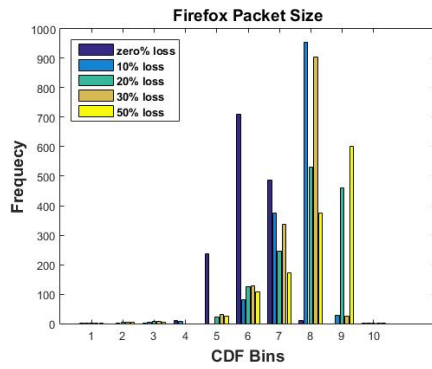
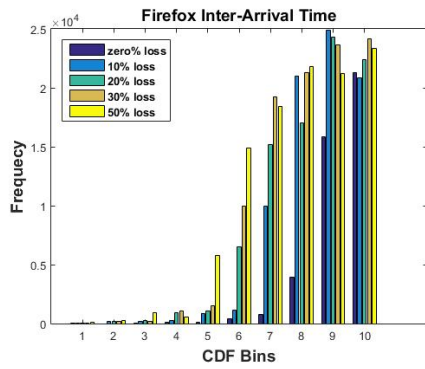
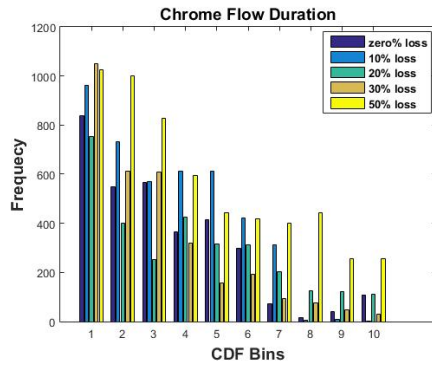
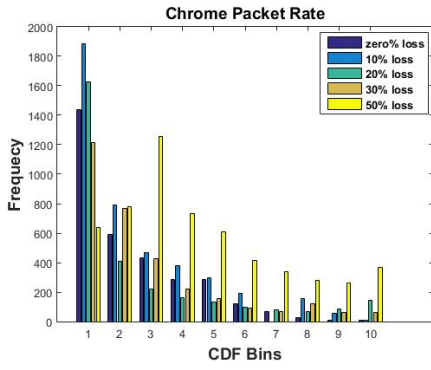
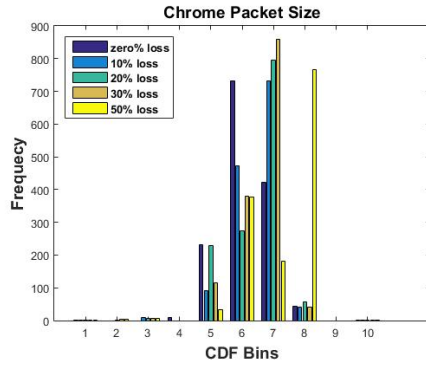
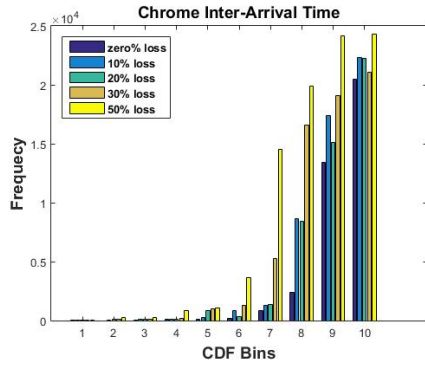
۵.۳ جمع‌بندی

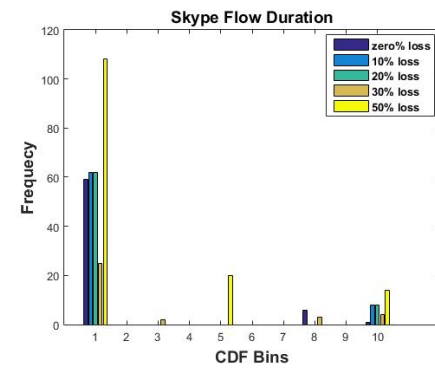
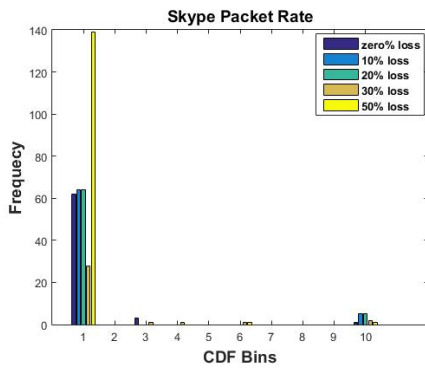
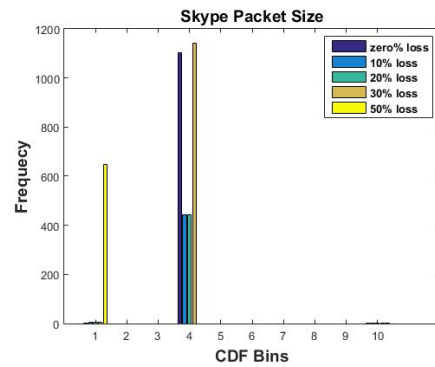
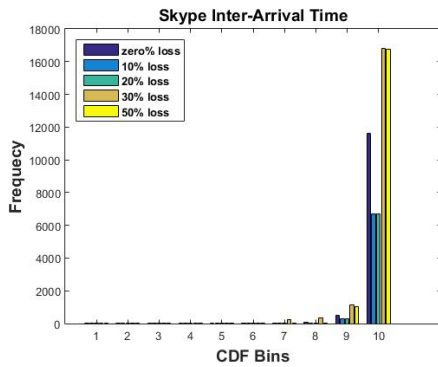
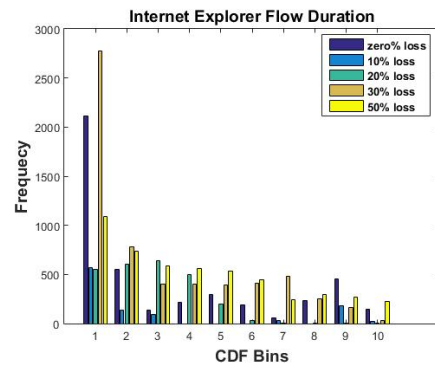
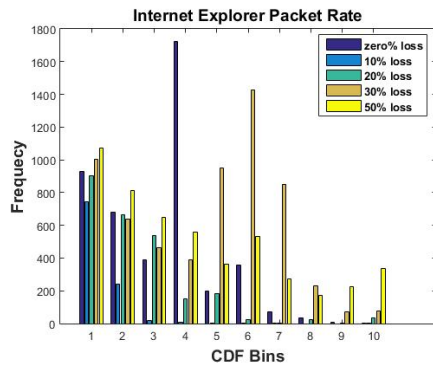
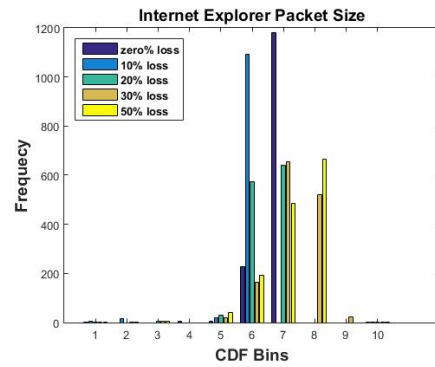
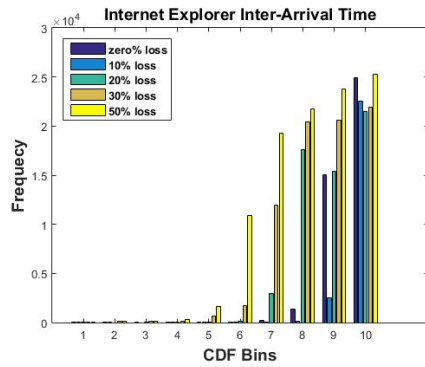
در این فصل ضمن بررسی مجدد چالش‌های موجود در شناسایی ترافیک شبکه به دو چالش اصلی بروز رسانی برنامه‌های کاربردی و رخداد اتلاف بسته‌ها پرداخته شد. بدین منظور ابتدا با تحلیل رفتاری برنامه‌های کاربردی موجود در دو پایگاه داده UNIBS (داده‌های مربوط به بروزرسانی سال ۲۰۰۹) و پایگاه داده جمع‌آوری شده (داده‌های مربوط به بروزرسانی سال ۲۰۱۶)، چهار ویژگی آماری مقاوم استخراج گردید. در ادامه به رخداد اتلاف بسته‌ها در شبکه و تاثیرات آن بر رفتار برنامه‌های کاربردی و همچنین دقت شناسایی پرداخته شد و ویژگی‌های مرتبه سومی (بدست آمده از چهار ویژگی اصلی استخراج شده) ارائه گردید.

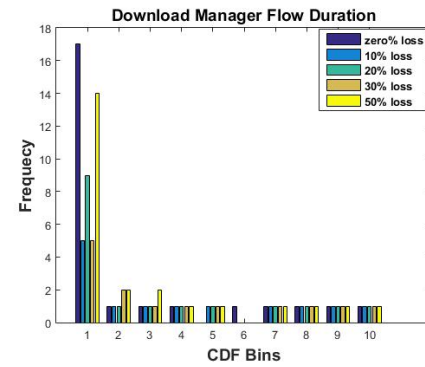
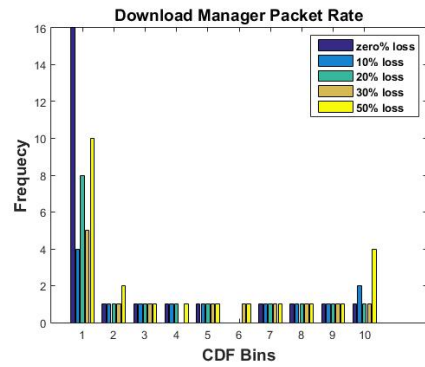
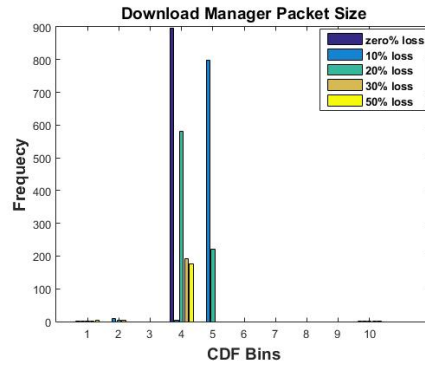
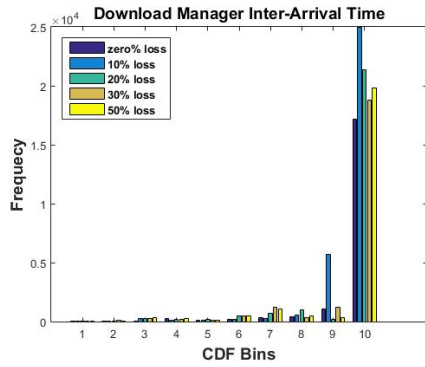
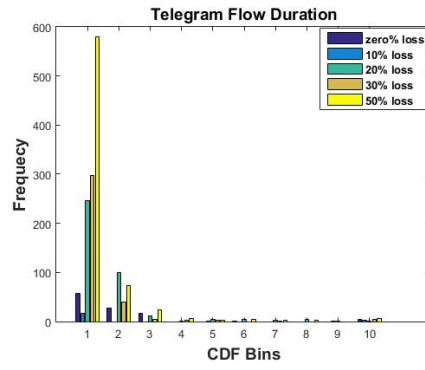
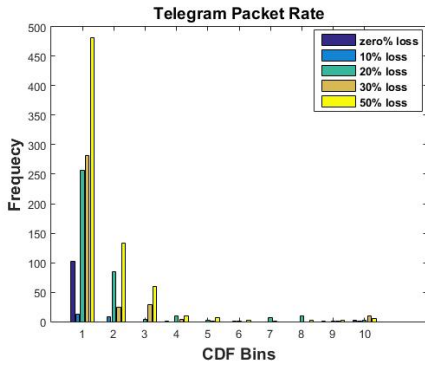
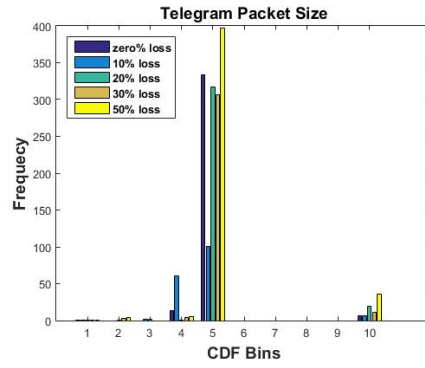
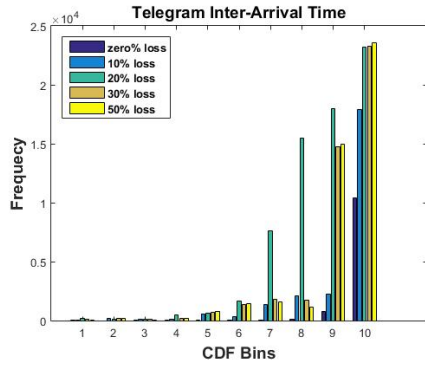


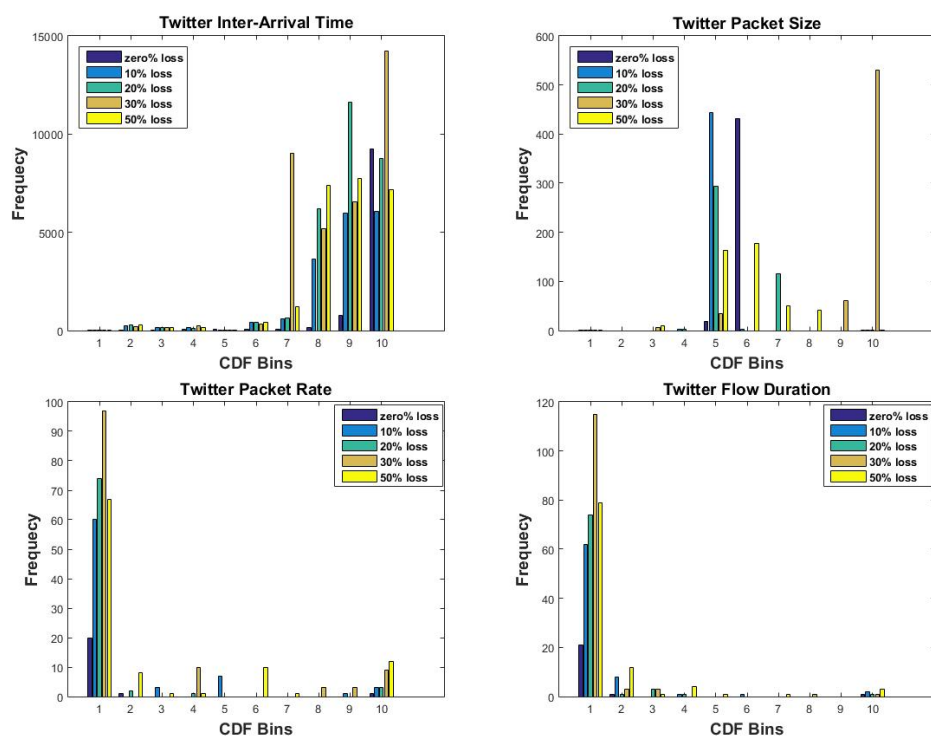


شکل ۱۴.۳: نمودار توزیع تجمعی چهار ویژگی آماری مربوط به هفت برنامه کاربردی با نرخ های مختلف اتلاف بسته









شکل ۱۵.۳: فراوانی تابع توزیع تجمعی چهار ویژگی مربوط به بسته‌های ترافیک هفت برنامه کاربردی در ده بازه یکسان از ۰ تا ۱

فصل چهارم

ارزیابی روش‌های ارائه شده

۱.۴ مقدمه

در فصل قبل چالش‌های تاثیرگذار بر دقت شناسایی برنامه‌های کاربردی مورد بررسی قرار گرفت و پس از آن چهار ویژگی آماری (ویژگی‌های مرتبه اول) جهت حل چالش بروزرسانی برنامه‌های کاربردی ارائه گردید و سپس به مسئله اتلاف بسته‌ها و تاثیرگذاری این رخداد بر ویژگی‌های آماری پرداخته شد و ویژگی‌های مرتبه سومی ارائه گردید. در این فصل با توجه به ویژگی‌های ارائه شده در فصل قبل به ارزیابی این ویژگی‌ها و همچنین بررسی میزان دقت شناسایی برنامه‌های کاربردی با توجه به ویژگی‌های ارائه شده پرداخته می‌شود. به همین منظور ابتدا مروری بر معیارهای ارزیابی روش‌های شناسایی برنامه‌های کاربردی پرداخته شده و سپس با استفاده از معیار اطلاعات متقابل میزان تمایزدهندگی ویژگی‌های ارائه شده بررسی شده و در ادامه با استفاده از تکنیک‌های یادگیری ماشین میزان دقت شناسایی برنامه‌های کاربردی برای چالش‌های بروزرسانی برنامه کاربردی و رخداد اتلاف بسته مورد ارزیابی قرار می‌گیرد.

۲.۴ معیارهای ارزیابی روش‌های شناسایی ترافیک

سازوکارهای مطرح شده در شناسایی ترافیک را می‌توان از جنبه‌های مختلفی مورد بررسی و ارزیابی قرار داد. در این بخش به معرفی برخی از این معیارها می‌پردازیم.

۱.۲.۴ عمق تشخیص

برخی از روش‌های شناسایی ترافیک می‌توانند نوع برنامه کاربردی تولیدکننده جریان ترافیک را مشخص کنند. برخی تنها مشخص می‌کنند که ترافیک تولید شده از نوع چند رسانه‌ای^۱ است و یا نوع دیگری است. برخی دیگر فقط می‌توانند نوع پروتکلی که در لایه کاربرد استفاده می‌شود را تعیین کنند، حال آن‌که ممکن است چندین برنامه کاربردی از یک پروتکل استفاده کنند. در تمامی حوزه‌های یاد شده کارهای تحقیقاتی مختلفی صورت گرفته است. مراجع [۱۹، ۷۶، ۷۷، ۷۸، ۷۹]، سعی کرده‌اند که ترافیک Skype را تشخیص دهند. در مراجع [۸۰، ۸۱، ۲۰]، سعی شده است که برنامه‌های کاربردی را در دو گروه کلی، مانند برنامه‌های کاربردی چند رسانه‌ای و غیرچند رسانه‌ای، طبقه‌بندی نمایند.

۲.۲.۴ هزینه محاسباتی

یکی از مهم‌ترین معیارها در مقایسه روش‌های شناسایی ترافیک، میزان هزینه محاسبات است. گشودن بسته در لایه‌های مختلف و همچنین تعداد نمونه مورد نیاز، دو موردی است که به شدت هزینه محاسبات

¹Multimedia

را تحت تاثیر قرار می‌دهد. از طرفی دیگر به جهت حفظ امنیت و شناسایی رفتار برنامه‌های کاربردی، برنامه کاربردی بایستی با کمترین تعداد نمونه شناسایی گردد. ارایه یک روش شناسایی ترافیک مناسب در شبکه‌های با نرخ انتقال بالا و تعداد کاربر زیاد، مستلزم پایین بودن هزینه محاسبات آن روش است.

۳.۲.۴ میزان دقت

فرض کنید که در K بار اجرای یک روش شناسایی ترافیک، تحلیل‌گر ترافیک P بار تشخیص داده است که برنامه کاربردی X ، استفاده شده است، و $N = K - P$ بار عدم استفاده از X ، تشخیص داده شده است. برحسب درست یا غلط بودن تشخیص تحلیل‌گر ترافیک، نتیجه را می‌توان در چهار گروه تقسیم‌بندی نمود.

- در تعداد TP مورد، برنامه کاربردی X واقعا استفاده شده، و تحلیل‌گر ترافیک نیز بدرستی استفاده از آن را تشخیص داده است.
- در FP مورد، برنامه کاربردی X استفاده شده است، اما تحلیل‌گر ترافیک به اشتباه عدم استفاده تشخیص داده است. اصطلاحا به این اشتباه، **خطای نوع اول** گفته می‌شود.
- در TN مورد، کاربر X را اجرا نکرده است، و تحلیل‌گر ترافیک نیز به درستی عدم استفاده، تشخیص داده است.
- در FN تعداد از نتایج، کاربر X را اجرا نکرده، اما تحلیل‌گر ترافیک به اشتباه اجرای آن را تشخیص داده است. به این نوع اشتباه، اصطلاحا **خطای نوع دوم**، گفته می‌شود.

برای ارزیابی میزان دقت یک روش شناسایی ترافیک، معیارهای بسیاری برحسب چهار نوع نتیجه یاد شده، تاکنون ارایه شده است [۸۲]. برخی از این معیارها به صورت عددی و برخی به صورت نموداری هستند. مهم‌ترین نمودار در این حوزه نمودار ROC^2 است. نمودار ROC از رسم نرخ TP (TPR) بر حسب نرخ FP (FPR)، به ازای تغییر یک پارامتر در روش شناسایی ترافیک مورد نظر، ایجاد می‌شود. دو پارامتر TPR و FPR به صورت رابطه (۱.۴) تعریف می‌شوند.

$$TPR = \frac{TP}{P} \quad FPR = \frac{FP}{P} \quad (۱.۴)$$

همچنین معیارهای درصد تشخیص درست هنگامی که ترافیک متعلق به برنامه کاربردی X می‌باشد و درصد کل ترافیک‌های موجود که متعلق به X هستند به صورت روابط (۲.۴) و (۳.۴) تعریف می‌شوند.

$$Recall = \frac{TP}{TP + FN} \quad (۲.۴)$$

²Receiver Operating Characteristics

$$\text{Precision} = \frac{TP}{TP + FN} \quad (3.4)$$

معیار دیگری نیز معرفی شده است که به $F - \text{Measure}(F_\beta)$ معروف می‌باشد و مقدار آن به کمک مقادیر Precision و Recall محاسبه می‌شود. این معیار نیز جهت ارزیابی کارایی سامانه شناسایی ترافیک برنامه کاربردی استفاده می‌شود. رابطه (۴.۴) نحوه محاسبه فرم معمول این معیار (به ازای $\beta = 1$) را نشان می‌دهد.

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4)$$

۴.۲.۴ شروع زمان تحلیل

فرض کنید که برنامه کاربردی X در زمان t_0 ، آغاز به ارسال بسته از طریق شبکه ارتباطی کند. تحلیل‌گر ترافیک از زمان t_c شروع به دریافت اطلاعات بسته‌های ارسالی از سوی برنامه کاربردی X می‌کند. یکی از قابلیت‌های مهم در هر روش شناسایی ترافیک، این است که آیا نیازی هست که تحلیل‌گر ترافیک از همان ابتدای اتصال وارد عمل شود ($t_0 = t_c$)، یا روش شناسایی ترافیک به گونه‌ای است که از هر کجای اتصال که او وارد شود ($t_c \geq t_0$)، به شرط داشتن میزان اطلاعات مورد نیاز، می‌تواند به صورت موفقیت‌آمیز کار تحلیل را به نتیجه برساند.

۵.۲.۴ تعداد بسته مورد نیاز

در کل در یک روش شناسایی ترافیک تعداد بسته‌های مورد نیاز شامل دو گروه بسته خواهند شد.

- در دسته اول منظور از تعداد بسته مورد نیاز، تعداد بسته جهت آموزش الگوریتم یادگیری می‌باشد. هنگام استفاده از الگوریتم‌های مبتنی بر یادگیری ماشین، تعدادی بسته به منظور یادگیری باید در ابتدای کار، به الگوریتم داده شود. این بسته‌ها را می‌توان ضبط یا تولید نمود و به عنوان ورودی الگوریتم در نظر گرفت. در ضمن عملیات یادگیری می‌تواند به صورت برون خط نیز انجام پذیرد.
- در دسته دوم منظور از تعداد بسته مورد نیاز، تعداد بسته‌هایی از یک جریان است که جهت عملیات شناسایی نیاز به بررسی آنها می‌باشد. جهت استخراج ویژگی‌های مورد نظر برای شناسایی ترافیک برنامه کاربردی، از این بسته‌ها استفاده می‌شود. برای مثال در روش‌های شناسایی مبتنی بر شماره درگاه، تحلیل‌گر ترافیک تنها به بازگشایی یک بسته نیاز دارد. اما در روش‌های آماری نیاز به تعداد بیشتری بسته خواهد بود.

در روش‌های آماری، معمولاً تحلیل‌گر ترافیک برای بدست آوردن ویژگی‌های مورد نظر، به دنبال تخمین یکسری پارامتر است. به عنوان نمونه می‌توان به تخمین میانگین و یا واریانس نرخ تولید بسته‌ها اشاره نمود. تمامی بسته‌های تولید شده توسط یک برنامه کاربردی مشخص را به عنوان یک جامعه آماری در نظر بگیرید. در نتیجه تحلیل‌گر ترافیک در این روش‌ها، هرچه اطلاعات تعداد بیشتری از بسته‌ها را بدست آورد، میزان تخمینش بهتر خواهد شد.

فرض کنید که تحلیل‌گر ترافیک می‌خواهد به ما اطمینان دهد که تخمین پارامترها تاچه حدی به پارامتر واقعی نزدیک است. او این کار را می‌تواند با ارایه دو پارامتر (ϵ, γ) انجام دهد. او می‌گوید که احتمال این که خطای تخمین کمتر از یک مقدار معین به نام ϵ باشد، از یک سطح به مانند γ بیشتر است. بیان ریاضی این عبارت به صورت رابطه (۵.۴) است.

$$\Pr[|\hat{T} - T| < \epsilon] \geq \gamma \quad (5.4)$$

♠ $\Pr[.]$: عملگر احتمال رخداد یک رویداد.

♠ γ : سطح اطمینان یک روش شناسایی ترافیک که مقداری است بین صفر تا یک.

♠ $\hat{T}$: مقدار تخمین زده شده.

♠ T : پارامتر واقعی منتج شده از تمامی اعضای جامعه.

نکته مهم در این مجال این است که رابطه (۵.۴) به طور کامل وابسته به تعداد نمونه‌گیری از جامعه است. لذا نوع روش تخمین و پارامترهای (ϵ, γ) بر روی تعداد بسته مورد نیاز برای تحلیل تاثیر مستقیم بر جای خواهند گذاشت. این مبحث به طور کامل در علم آمار مورد بررسی قرار می‌گیرد [۸۳، ۸۴]. لازم به ذکر است که تخمین بهتر برای رسیدن به تشخیص بهتر به تنهایی کافی نیست. بلکه علاوه بر تخمین بهتر، دقت خود روش شناسایی ترافیک از اهمیت فوق‌العاده‌ای برخوردار است. به عبارت دیگر با تخمین بهتر، تنها ورودی روش شناسایی ترافیک دقیق‌تر می‌شود، نه خود روش.

۳.۴ معیارهای ارزیابی روش ارائه شده

جهت ارزیابی روش‌های ارائه شده، پس از آموزش سامانه توسط داده‌های دارای برچسب و استفاده از الگوریتم‌های یادگیری ماشین، سامانه توسط داده‌های برخاسته از آزمایش قرار می‌گیرد. همانطور که در روش ارائه شده مطرح گردید، خروجی سامانه، برنامه‌های کاربردی شناسایی شده و شناسایی نشده می‌باشد. برای بررسی صحت عملکرد از معیار میزان دقت که در بخش قبل مطرح گردید، استفاده خواهد

شد. در این راستا معیارهای Recall، Precision و F1-Measure نیز استفاده شده است. جدول (۱.۴) نحوه محاسبه پارامترهای مربوط به رابطه‌های (۳.۴، ۲.۴، ۴.۴) را نشان می‌دهد.

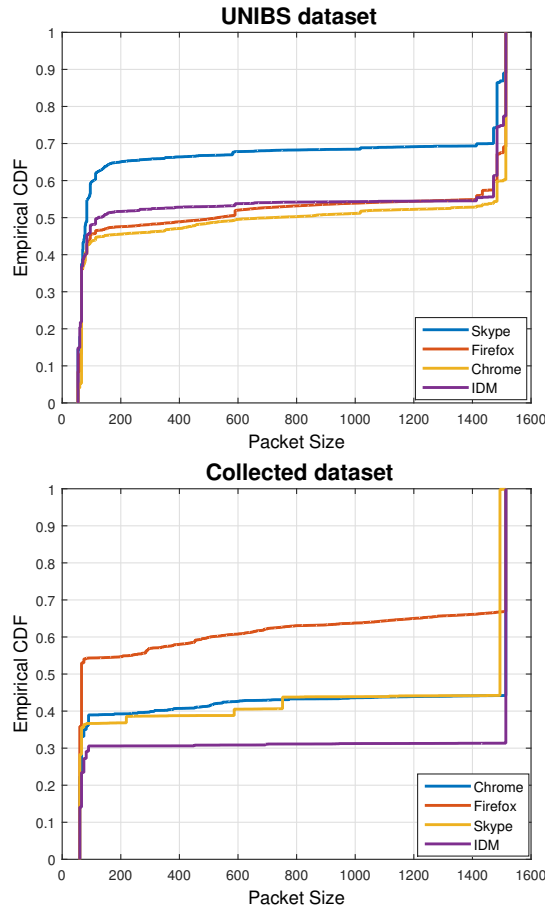
جدول ۱.۴: معیارهای ارزیابی روش ارائه شده

شناسایی صحیح شده	شناسایی صحیح نشده	شناسایی صحیح شده	شناسایی صحیح نشده
نمونه‌های درست یک برنامه کاربردی	نمونه درست یک برنامه کاربردی که اشتباه شناسایی شده True Positive (TP)	نمونه درست یک برنامه کاربردی که اشتباه شناسایی شده False Negative (FN)	نمونه نادرست یک برنامه کاربردی که اشتباه شناسایی شده True Negative (TN)
نمونه نادرست یک برنامه کاربردی	نمونه نادرست یک برنامه کاربردی که اشتباه شناسایی شده False Positive (FP)	نمونه نادرست یک برنامه کاربردی که اشتباه شناسایی شده True Negative (TN)	نمونه نادرست یک برنامه کاربردی که اشتباه شناسایی شده False Positive (FP)

۴.۴ ارزیابی ویژگی‌های رفتاری

در فصل قبل، با استفاده از تحلیل رفتاری، برنامه‌های کاربردی با توجه به P2P و Client-Server بودن آنها، مورد تحلیل و بررسی قرار گرفته و نشان داده شد که بعضی از خصوصیات رفتاری موجود در برنامه‌های کاربردی در این دسته‌ها متفاوت بوده و دارای خاصیت تمایز دهنده می‌باشند. در این بخش به ارزیابی این خصوصیات با استفاده از دو پایگاه داده UNIBS و پایگاه داده جمع‌آوری شده پرداخته می‌شود. نکته‌ای که لازم است مطرح گردد این می‌باشد که داده‌های جمع‌آوری شده از برنامه‌های کاربردی در پایگاه داده UNIBS مربوط به ویرایش سال ۲۰۰۹ این برنامه‌های کاربردی بوده و همچنین داده‌های پایگاه داده جمع‌آوری شده مربوط به ویرایش سال ۲۰۱۶ این برنامه‌های کاربردی می‌باشد. شکل (۱.۴) توزیع اندازه سرآیندها در این دو پایگاه داده را نشان می‌دهد. همانطور که مشاهده می‌شود، توزیع داده‌ها برای دو ویرایش مختلف از برنامه‌های کاربردی متفاوت می‌باشد.

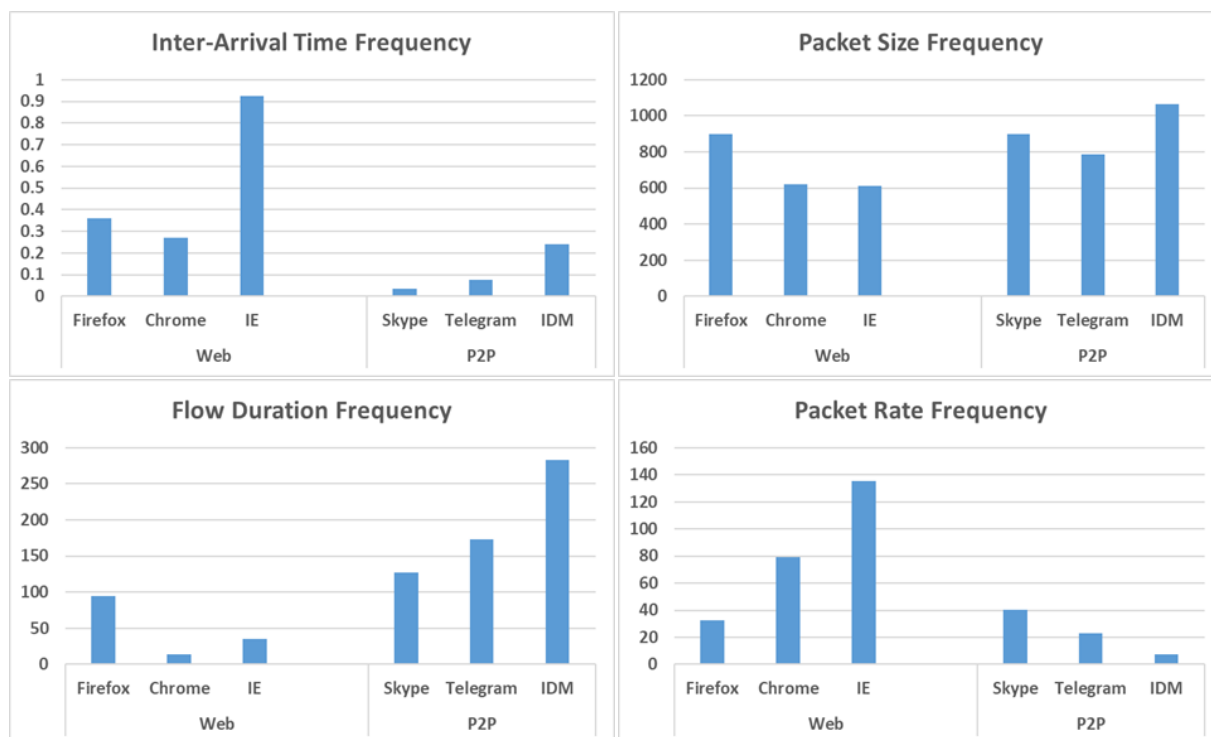
همانطور که در فصل قبل نشان داده شد، بعضی از خصوصیات رفتاری در برنامه‌های کاربردی دارای قابلیت تمایز دهنده می‌باشند. ویژگی‌های انتخاب شده از تحلیل رفتاری برنامه‌های کاربردی، مجموعه ترتیب بسته‌ها، اندازه بسته‌ها، مدت زمان مابین ارسال و دریافت بسته‌ها، مدت زمان جریان‌ها و نرخ ارسال و دریافت بسته‌ها در لایه شبکه می‌باشد. شکل (۲.۴ و ۳.۴) نمودار توزیع تجمعی و فراوانی این ویژگی‌ها را برای برنامه‌های کاربردی پایگاه داده جمع‌آوری شده نشان می‌دهد. همانطور که مشاهده می‌شود، رفتار جریان‌های برنامه‌های کاربردی در این ویژگی‌ها متفاوت بوده و می‌توان از این مجموعه ویژگی جهت شناسایی برنامه‌های کاربردی استفاده نمود. برای تشخیص قابلیت تمایز دهنده این



شکل ۱.۴: توزیع ویژگی آماری اندازه سرآیندها در پایگاه داده UNIBS (ویرایش قدیمی) و پایگاه داده جمع‌آوری شده (ویرایش جدید) برای برنامه‌های کاربردی

ویژگی‌ها در شناسایی برنامه‌های کاربردی، کیفیت تمایزدهندگی هر ویژگی را می‌توان با استفاده از اطلاعات متقابل قبل از آزمایش پایگاه داده ارزیابی نمود. اطلاعات متقابل، $I(X; Y)$ یک معیار غیر منفی می‌باشد که میزان وابستگی ویژگی X را نسبت به برنامه کاربردی Y نشان می‌دهد و هنگامی که این مقدار صفر باشد X و Y مستقل از هم می‌باشند [۸۵]. شکل (۴.۴ و ۵.۴) نمودار اطلاعات متقابل بین ویژگی‌های ارائه شده و برنامه‌های کاربردی را در دو پایگاه داده UNIBS و پایگاه داده جمع‌آوری شده را نشان می‌دهد.

آنتروپی برنامه کاربردی y_i ، $i \in \{1, \dots, K\}$ که K نشان‌دهنده تعداد برنامه‌های کاربردی که ویژگی‌های آماری جهت شناسایی آنها استخراج شده‌اند، می‌باشد را می‌توان توسط رابطه (۱۰.۳) محاسبه نمود. مقدار اطلاعات متقابل بین هر مجموعه از ویژگی‌ها و یک برنامه کاربردی کمتر از سه (حد بالا برای تعداد برنامه‌های کاربردی سه می‌باشد) بوده و هنگامی حاصل می‌گردد که شناسایی برنامه کاربردی کاملاً قابل پیش‌بینی باشد. همانطور که مشاهده می‌شود، مجموعه ویژگی آماری استخراج شده دارای بیشترین حد

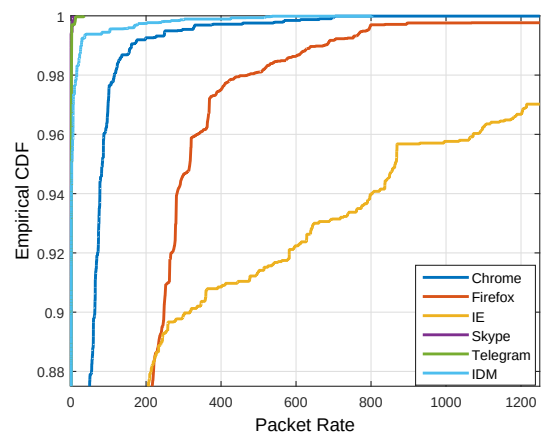
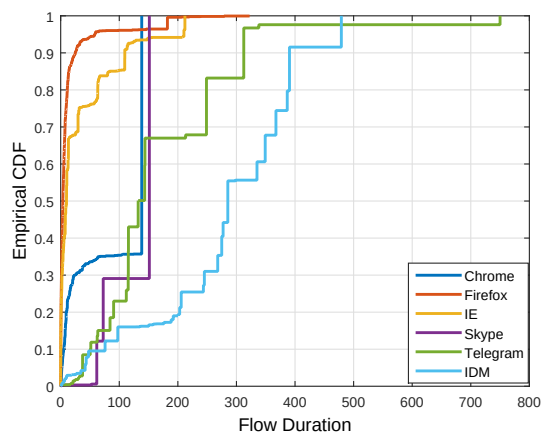
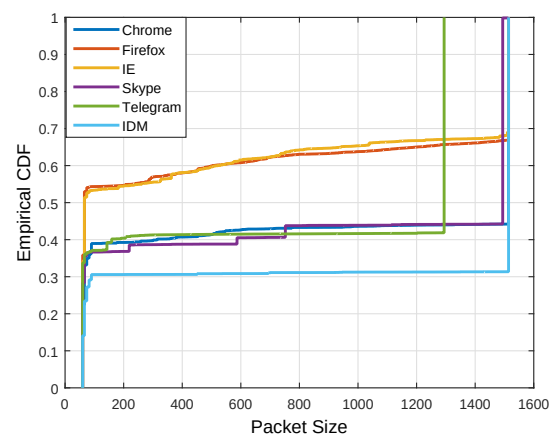
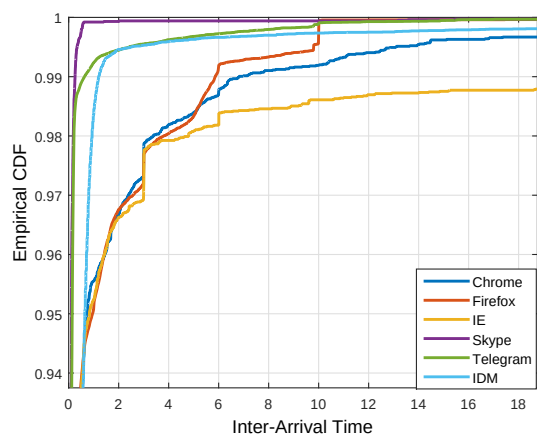


شکل ۲.۴: نمودار فراوانی چهار ویژگی آماری استخراج شده برای دو گروه برنامه کاربردی P2P و Client-Server

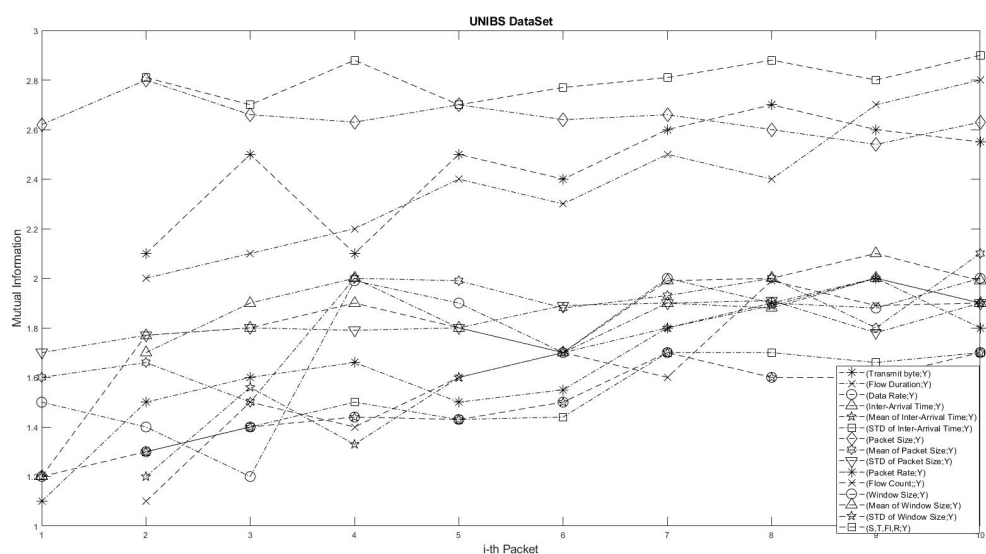
اطلاعات نسبت به برنامه‌های کاربردی در دو پایگاه داده می‌باشد و همچنین نشان داده شده است که هر کدام از این ویژگی‌ها به تنهایی دارای این حد از اطلاعات نسبت به برنامه‌های کاربردی نمی‌باشند. در بسته‌های اول و دوم و حتی در بعضی مواقع در بسته‌های سوم، بسیاری از ویژگی‌ها (از جمله ویژگی‌های استخراج شده) دارای مقدار نمی‌باشند. اگر به عنوان نمونه ویژگی مدت زمان مابین ارسال و دریافت بسته‌ها را در نظر بگیریم، این ویژگی تا زمانی که بسته‌های اول و دوم ارسال نشوند و بسته‌های برگشتی در مسیر یاب دریافت نشود مقداری ندارد.

در شبیه‌سازی انجام شده ما برای تمامی ویژگی‌هایی که مقدار ندارند، مقدار صفر را در نظر گرفتیم. پس طبیعی است که در بسته‌های اول و دوم مقادیر ویژگی‌های بسته‌های مربوط به برنامه‌های کاربردی مختلف دارای مقادیر شبیه به هم و شاید نادرست باشند. وقتی مقادیر ویژگی‌های مربوط به بسته‌های یک برنامه کاربردی در بسته‌های اول و دوم مبهم باشد، طبیعی است که میزان اطلاعات آنها به صورت انفرادی یا مجموعه‌ای از برنامه کاربردی کم یا اشتباه باشد ولی این امر در مورد اندازه بسته‌ها صدق نمی‌کند. به این دلیل که اندازه بسته‌ها از بسته اول مشخص بوده و مقادیر آن درست می‌باشد.

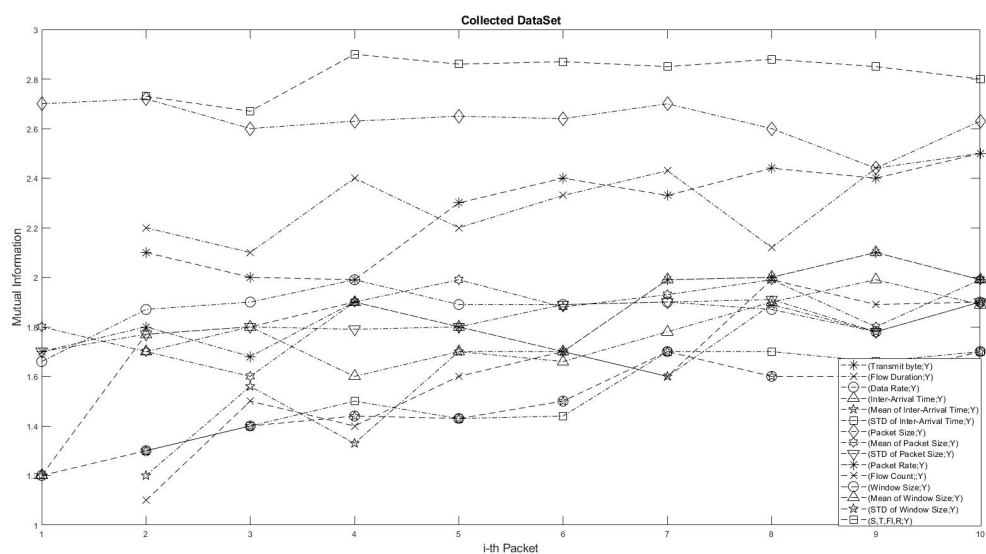
در خصوص میزان آنتروپی و اطلاعات متقابل در بسته‌های اول ترافیک شبکه چند مقاله مورد بررسی قرار گرفت [۲، ۳، ۸۶]. در تمامی این مقالات میزان اطلاعات متقابل برای ویژگی‌هایی که در بسته‌های



شکل ۳.۴: نمودار توزیع تجمعی چهار ویژگی آماری استخراج شده برای دو گروه برنامه کاربردی P2P و Client-Server



شکل ۴.۴: میزان اطلاعات متقابل هر ویژگی و برنامه کاربردی تولید کننده آن در پایگاه داده UNIBS (s=Size, t=Inter-Arrival Time, fl=Flow duration, r=Packet Rate)



شکل ۵.۴: میزان اطلاعات متقابل هر ویژگی و برنامه کاربردی تولید کننده آن در پایگاه داده جمع آوری شده (s=Size, t=Inter-Arrival Time, fl=Flow duration, r=Packet Rate)

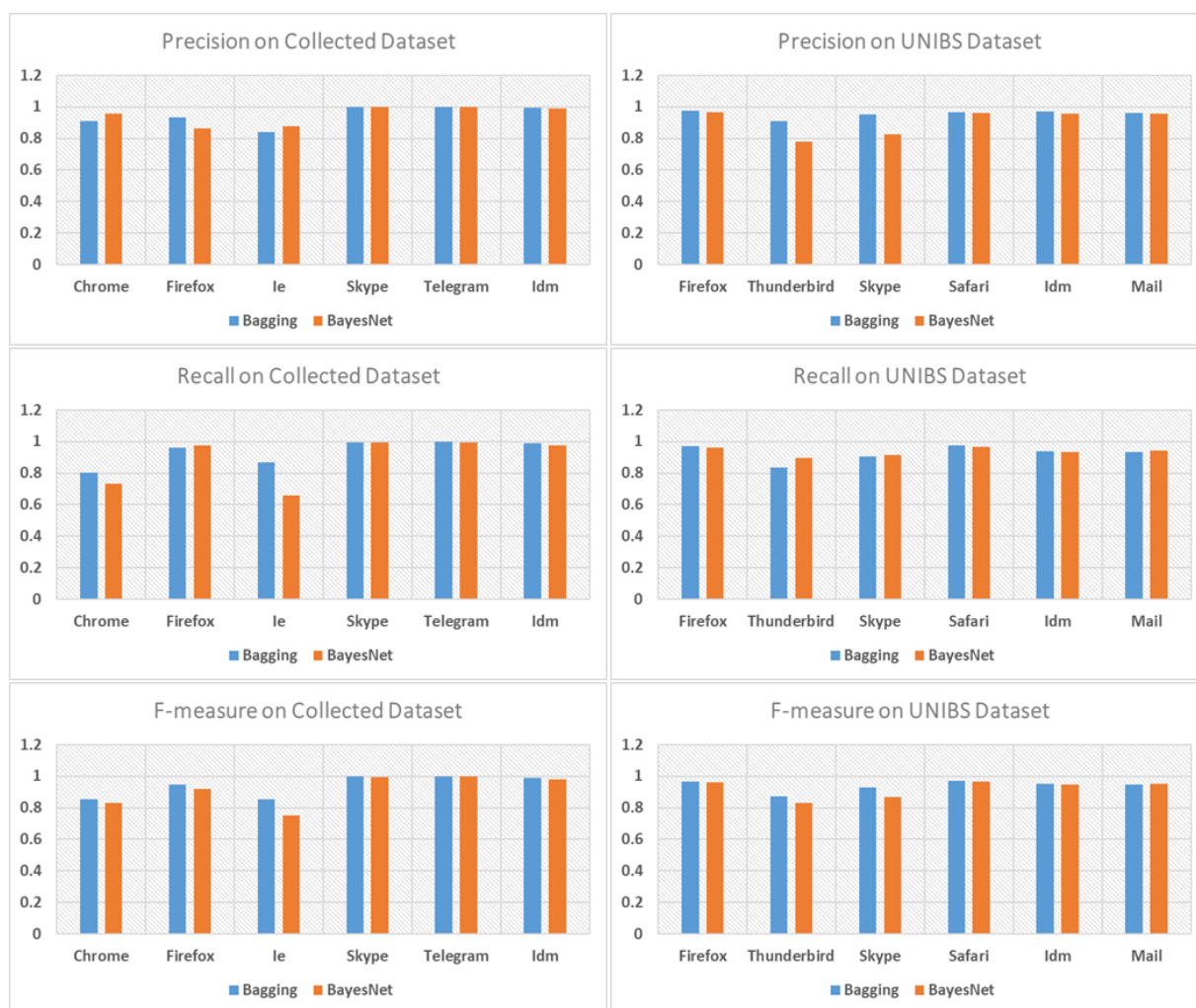
اول دارای مقدار نیستند در نظر گرفته نشده است و نمودار اطلاعات برای این ویژگی ها از بسته های دوم به بعد رسم شده است. همانطور که در بالا مطرح گردید در نتایجی که ما برای میزان اطلاعات ویژگی ها رسم کرده ایم، میزان اطلاعات متقابل برای ویژگی های بدون مقدار در بسته های اول نیز محاسبه شده است. پس از بررسی های انجام شده و اجرای چند باره کد شبیه سازی مشخص گردید که تابع تخمین

چگالی در هنگامی که ویژگی‌ها بدون مقدار یا با مقدار صفر هستند درست کار نکرده و دچار مشکل می‌گردد. لذا کدهای شبیه‌سازی شده مورد بازبینی قرار گرفت و میزان اطلاعات متقابل برای ویژگی‌هایی که در بسته‌های اول دارای مقدار نیستند از بسته دوم به بعد در نظر گرفته شد.

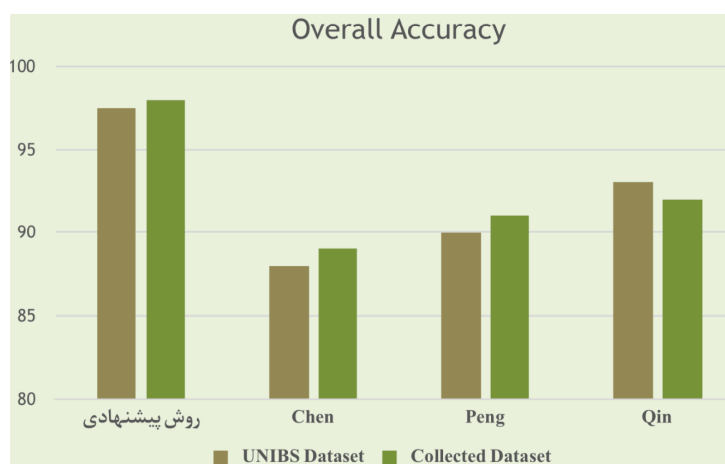
پس از آنکه با استفاده از معیار اطلاعات متقابل میزان تمایز دهنده‌گی ویژگی‌های ارائه شده نشان داده شد، به شناسایی برنامه‌های کاربردی با استفاده از تکنیک‌های یادگیری ماشین پرداخته می‌شود. به همین منظور در این رساله از دو تکنیک مهم طبقه‌بندی Bagging (with Random Forest classifier) و Bayesian Network استفاده شده است. داده‌های این دو پایگاه داده ممکن است دارای داده‌هایی با برچسب اشتباه، وابستگی داده‌ای، و داده‌های زائد باشند. از اینرو از طبقه‌بندهای Bagging (with Random Forest classifier) و Bayesian Network جهت بروزرسانی و اصلاح مقادیر داده‌ها با استنتاج احتمالاتی استفاده شده‌اند [۱۸، ۲۴].

در ادامه جهت اجرای تکنیک‌های یادگیری ماشین به صورت جامع و بهینه از نرم افزار Weka استفاده شده است [۸۷]. شکل (۶.۴) مقادیر Recall، Precision و F-Measure بدست آمده از آزمایش دو مجموعه داده با استفاده از تکنیک‌های یادگیری ماشین را نشان می‌دهد. بدلیل آنکه تعداد نمونه‌ها برنامه‌های کاربردی متفاوت می‌باشد، هر برنامه کاربردی به صورت جداگانه بررسی شده است. در ادامه کار جهت ارزیابی ویژگی‌های ارائه شده، میزان دقت شناسایی برنامه‌های کاربردی با نتایج سه روش ارائه شده در سال‌های اخیر مقایسه شده است [۲، ۳، ۵]. این سه روش به همراه ویژگی‌های مورد استفاده در نرم افزار متلب پیاده‌سازی گردید و نتایج اجرای این روش‌ها بر روی دو مجموعه داده با نتایج بدست آمده از ویژگی‌های استخراج شده مقایسه گردید. شکل (۷.۴) میانگین دقت شناسایی هر سه روش را در مقایسه با روش ارائه شده نشان می‌دهد. همانطور که مشاهده می‌شود، نتایج ارزیابی نشان دهنده آن است که ویژگی‌های ارائه شده با قطعیت میزان دقت شناسایی ترافیک برنامه‌های کاربردی شبکه را افزایش می‌دهد.

در نتیجه با استفاده از ویژگی‌های آماری استخراج شده از تحلیل رفتاری برنامه‌های کاربردی، مقادیر Recall، Precision و F-measure افزایش یافته و نتایج شناسایی برنامه‌های کاربردی در مقایسه با روش‌های مقالات اخیر افزایش یافته و به طور میانگین دقت ۹۷/۵٪ حاصل شده است. جدول (۲.۴) نتایج حاصل از ویژگی‌های استخراج شده را در مقایسه با کارهای اخیر نشان می‌دهد.



شکل ۶.۴: مقادیر Precision، Recall و F-Measure بدست آمده از آزمایش دو مجموعه داده با استفاده از تکنیک‌های یادگیری ماشین



شکل ۷.۴: مقایسه میانگین دقت شناسایی برنامه‌های کاربردی با سه تکنیک [۵] Chen، [۳] Peng و [۲] Qin

جدول ۲.۴: مقایسه نتایج مابین روش ارائه شده و کارهای اخیر

مرجع	روش شناسایی	ویژگی‌ها	دقت	پایگاه داده
[۲۴]	Bayesian Technique	flow duration, port number, inter-arrival time, packet size, bandwidth used , inter-arrival time FFT	96.26%	Hand Classified Traces
[۵۱]	Naive Bayes and Chi-Square test	Packet size in transport layer	97%	Two self collected datasets
[۱۳]	Protocol fingerprints	packet size, inter-arrival time, packets order	90%	Proprietary traces
[۵۹]	C4.5 Decision tree	packet size, byte rate, packet size (MIN, MAX, AVG, STD)	90%	UTM campus network
[۵۰]	C4.5 Decision Tree	Fast Correlation-Based Filter Algorithm	95.21%	Captured at Ohio University
[۴۳]	Feed-Froward	Fast Correlation-Based Filter Algorithm	95%	Hand Classified Traces
[۸۸]	Naïve Bayes, MLP, C4.5	AdaBoost	90%	Chinese educational site
ارائه شده	Bagging, BayesNet	Packet order, packet size, inter-arrival time, flow time, transmission rate	97.5%	Collected Dataset, UNIBS

۵.۴ ارزیابی ویژگی‌های مرتبه سوم

به منظور ارزیابی قابلیت ویژگی‌های استخراج شده در شناسایی ترافیک شبکه در شرایط اتلاف بسته، آنها را در یک شبکه عصبی چند لایه پرسپترونیک بکار میگیریم. مزیت اصلی استفاده از شبکه عصبی MLP قابلیت بالای این شبکه در یادگیری و پایداری آن در مقابل تغییرات ورودی می‌باشد. مشخصات و همچنین تعداد لایه‌ها و ورودی‌های شبکه عصبی مورد استفاده در جدول (۳.۴) آمده است.

جدول ۳.۴: مشخصات شبکه عصبی مورد استفاده

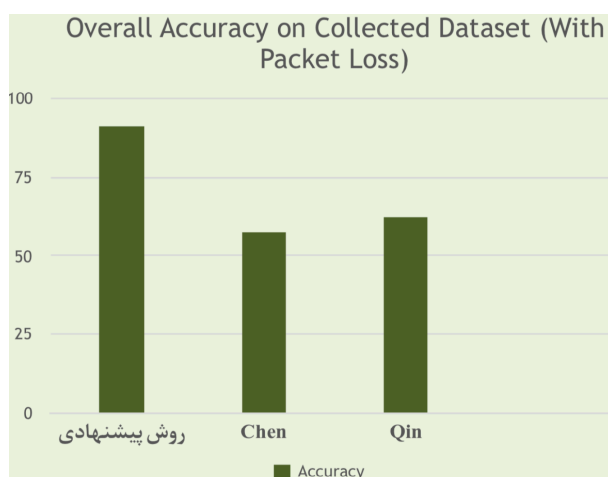
تعداد نرون‌های ورودی	۱۰ = تعداد ویژگی‌ها
تعداد نمونه‌های آموزش برای هر کلاس	۲۸۰۰
تعداد نمونه‌های آزمایش برای هر کلاس	۱۲۰۰
تعداد لایه‌های پنهان	۱
تعداد نرون در لایه پنهان	۷
نرخ یادگیری	۰.۵۰
حداکثر تعداد دفعات تکرار آموزش	۱۰۰
حد آستانه خطا برای توقف یادگیری	۱.۰

شبکه عصبی فوق پس از آموزش بر روی مجموعه داده جدول (۲.۳) (با احتساب نرخ اتلاف برای برنامه‌های P2P تا ۳۰٪ و برای برنامه‌های Client-Server تا ۵۰٪) اجرا شده و نتایج آن در جدول (۴.۴) نشان داده شده است. معیار ROC میزان حساسیت یا پیش بینی درست در مقابل پیش بینی غلط در یک سیستم طبقه‌بندی را نشان می‌دهد و منظور از TP میزان تشخیص‌های درست و FP میزان تشخیص‌های نادرست می‌باشد.

همانطور که مشاهده می‌شود با استفاده از ویژگی‌های استخراج شده دقت شناسایی در مجموعه داده حاوی اتلاف بسته به طور میانگین تا ۹۰/۵۳٪ افزایش یافته و به دقت شناسایی در محیط ایده‌آل (۹۷/۵٪) نزدیک شده است. اگر بخواهیم به طور دقیق‌تر نمونه‌های شناسایی شده را مورد بررسی قرار دهیم ماتریس درهم ریختگی نتایج بدست آمده در جدول (۵.۴) نشان داده شده است.

در جدول (۵.۴) برای هر برنامه کاربردی تعداد نمونه تشخیص داده شده از هر برنامه کاربردی نشان داده شده است. همانطور که مشاهده می‌شود با استفاده از ویژگی‌های استخراج شده دقت شناسایی نسبت به نتایج اولیه (شکل (۱۳.۳)) به میزان قابل توجهی افزایش یافته است و در مواردی که سیستم شناسایی نمونه مربوط به یک برنامه کاربردی را به اشتباه یک برنامه کاربردی دیگر تشخیص داده است، آن برنامه از خانواده همان برنامه کاربردی (خانواده برنامه‌های Client-Server یا خانواده برنامه‌های P2P)

می باشد. به عنوان نمونه هنگام تشخیص برنامه کاربردی Chrome اکثر نمونه‌های اشتباه تشخیص داده شده مربوط به برنامه‌های کاربردی Firefox و IE می باشد که هر دو از خانواده برنامه‌های Client-Server بوده و رفتاری مشابه دارند، و به همین صورت در شناسایی برنامه کاربردی Skype اکثر نمونه‌های اشتباه تشخیص داده شده مربوط به برنامه‌های Telegram و IDM می باشند که از خانواده برنامه‌های P2P بوده و رفتاری مشابه دارند. در ادامه کار جهت مقایسه روش ارائه شده، میزان دقت شناسایی با استفاده از این روش با دو روش ارائه شده توسط Qin [۲] و Chen [۴۴] مورد ارزیابی قرار گرفت. شکل (۸.۴) نتایج حاصل از دقت شناسایی روش ارائه شده در مقایسه با دو روش دیگر را نشان می دهد.



شکل ۸.۴: مقایسه دقت شناسایی روش ارائه شده با روش‌های Qin و Chen

جدول ۴.۴: نتایج شناسایی هفت برنامه کاربردی با استفاده از ویژگی‌های مرتبه سوم

برنامه کاربردی	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Chrome	0.958	0.772	0.750	0.796	0.032	0.750	74.96%
Firefox	0.972	0.798	0.765	0.835	0.025	0.765	76.5%
Internet Explorer	0.997	0.936	0.946	0.926	0.013	0.946	94.63%
Skype	0.983	0.906	0.899	0.912	0.014	0.899	89.93%
Telegram	1.00	0.972	0.994	0.952	0.008	0.994	99.39%
IDM	0.999	0.952	0.983	0.923	0.014	0.994	98.29%
Twitter	0.999	0.989	1.000	0.978	0.004	1.000	100%
Average	0.986	0.903	0.886	0.907	0.015	0.906	90.53%

جدول ۵.۴: ماتریس درهمریختگی شناسایی هفت برنامه کاربردی به کمک روش ارائه شده

برنامه کاربردی	Chrome	Firefox	Internet Explorer	Skype	Telegram	IDM	Twitter
Chrome	2856	386	172	188	79	127	2
Firefox	634	2889	91	46	40	76	0
Internet Explorer	39	111	3621	49	6	0	0
Skype	38	55	27	3414	66	111	85
Telegram	0	5	1	17	3768	0	0
IDM	22	15	0	28	0	3742	0
Twitter	0	0	0	0	0	0	3794
Average	0.986	0.903	0.886	0.907	0.015	0.906	0.905

۶.۴ جمع‌بندی

در این فصل به ارزیابی روش‌های ارائه شده در این رساله پرداخته شد. به همین منظور ابتدا با استفاده از اطلاعات متقابل، ویژگی‌هایی که از تحلیل رفتاری برنامه‌های کاربردی استخراج شده بود را مورد بررسی قرار داده و در ادامه با استفاده از الگوریتم‌های Bagging و Bayes Network عملیات شناسایی برنامه کاربردی انجام گرفت. در ادامه کار ویژگی‌های مرتبه سوم استخراج شده از ویژگی‌های رفتاری را بررسی کرده و پس از ارزیابی ویژگی‌های استخراج شده با استفاده از شبکه عصبی بر روی مجموعه داده تهیه شده، نشان داده شد که دقت شناسایی بهبود یافته و میزان آن به دقت شناسایی ترافیک شبکه در حالت ایده آل نزدیک شده است.

فصل پنجم

جمع‌بندی و کارهای آینده

۱.۵ مقدمه

این فصل از دو بخش تشکیل شده است. در بخش اول، جمع‌بندی از روشهای پیشنهادی در رساله، و در بخش بعدی پیشنهاداتی برای کارهای آینده مطرح شده است.

۲.۵ خلاصه‌ای از رساله

در سال‌های اخیر همگام با توسعه اینترنت، برنامه‌های کاربردی در حال اجرا بر روی این بستر از قبیل شبکه‌های اجتماعی روز به روز عمومیت بیشتری یافته‌اند. لذا تقاضای رو به رشد کاربران برای استفاده از پهنای باند بیشتر و از طرفی دیگر محدودیت ظرفیت فیزیکی خطوط ارتباطی شبکه، سرویس‌دهندگان اینترنت را بر آن می‌دارد تا با اولویت‌دهی تخصیص منابع کیفیت بهره‌برداری کاربران از منابع شبکه را بهبود بخشند. لیکن با توجه به افزایش بدافزارها و تلاش آنها برای پنهان‌سازی ترافیک خود به منظور گریز از سیستم‌های تشخیص نفوذ و دیوارهای آتش، شناسایی و طبقه‌بندی ترافیک برنامه‌های کاربردی امری ضروری می‌باشد. براین اساس، می‌توان طبقه‌بندی ترافیک را لازمه انجام بسیاری از وظایف امنیتی، مدیریتی و کنترلی در شبکه دانست.

روش‌های موجود جهت شناسایی و طبقه‌بندی ترافیک برنامه‌های کاربردی را می‌توان در چهار گروه تقسیم‌بندی کرد: (آ) مبتنی بر شماره درگاه، (ب) مبتنی بر بررسی محتوای بسته‌های ترافیک، (ج) مبتنی بر رفتار کاربر، (د) مبتنی بر ویژگی‌های آماری بسته‌های ترافیک شبکه. در این میان روش‌های مبتنی بر تحلیل آماری بسته‌ها، دارای اهمیت بالایی می‌باشند. از جمله ویژگی‌های آماری مورد استفاده می‌توان به اندازه بسته‌ها و همچنین مدت زمان مابین ارسال و دریافت بسته‌ها اشاره نمود. در این روش‌ها، عملیات شناسایی ترافیک برنامه‌های کاربردی بدون نیاز به بررسی محتوای بسته‌های ترافیک و با هدف حفظ محرمانگی اطلاعات مبادله شده انجام می‌شود.

بررسی‌ها و آزمایش‌های انجام شده در زمینه روش‌های شناسایی ترافیک شبکه با استفاده از ویژگی‌های آماری نشان دهنده این موضوع می‌باشد که این روش‌ها در مقابل چالش‌هایی همچون بروزرسانی برنامه‌های کاربردی توسط شرکت ارائه دهنده برنامه و همچنین اتلاف بسته‌ها، کارایی خود را از دست داده و دقت شناسایی برنامه‌های کاربردی با کاهش چشمگیری مواجه می‌شود. به همین دلیل در این رساله هدف، فائق آمدن بر این چالش‌ها است. مهمترین بخش یک روش شناسایی ترافیک، یافتن ویژگی‌هایی با قابلیت ایجاد تمایز بین یک برنامه کاربردی و سایر برنامه‌ها می‌باشد. بدین منظور، هدف اصلی این رساله

ارائه ویژگی‌های آماری مقاوم در برابر ویرایش‌های مختلف یک برنامه کاربردی و تأثیرات رخداد اتلاف بسته در شبکه و همچنین موثر جهت شناسایی ترافیک برنامه‌های کاربردی توسط الگوریتم‌های یادگیری ماشین در داده‌های ترافیک شبکه و در لایه شبکه جهت حفظ محرمانگی داده‌ها است.

به جهت استخراج ویژگی برای فائق آمدن بر چالش‌های مطرح شده، ابتدا پایگاه داده‌ای حاوی بسته‌های ترافیکی هفت برنامه کاربردی فایرفاکس، کروم، اینترنت اکسپلورر، اسکایپ، تلگرام، دانلود منیجر و توییتر (با شرایطی که در فصل اول اشاره شد) جمع‌آوری گردید. در مرحله بعد رفتار این برنامه‌های کاربردی جهت ساخت، ارسال و دریافت جریانی از بسته‌ها در تعامل بین سرور و کاربران، مورد بررسی و تحلیل قرار گرفت و در نتیجه چهار ویژگی آماری اندازه بسته‌ها، مدت زمان مابین ارسال و دریافت بسته‌ها، مدت زمان جریان‌ها و نرخ ارسال و دریافت بسته‌ها در شبکه استخراج گردید. در ادامه کار برای اثبات میزان تمایزدهندگی برنامه‌های کاربردی توسط این ویژگی‌ها از اطلاعات متقابل در تئوری اطلاعات استفاده شد. اطلاعات متقابل ارزش و میزان تمایزدهندگی هر ویژگی نسبت به برنامه‌های کاربردی را نشان می‌دهد و نشان داده شد که مجموعه این چهار ویژگی به همراه ترتیب بسته‌ها بالاترین سطح تمایزدهندگی را در خصوص این هفت برنامه کاربردی دارا می‌باشد. در نهایت میزان دقت شناسایی برنامه‌های کاربردی دو پایگاه داده UNIBS (که ویرایش برنامه‌های کاربردی آن مربوط به سال ۲۰۰۹ می‌باشد) و پایگاه داده جمع‌آوری شده (برنامه‌های کاربردی سال ۲۰۱۶) با استفاده از الگوریتم‌های Bagging و Bayes Network مورد بررسی قرار گرفت و به طور میانگین دقت شناسایی به ۹۷/۵٪ رسید. در نتیجه ویژگی‌های ارائه شده در مقابل روزرسانی برنامه‌های کاربردی مقاوم و موثر می‌باشد.

در ادامه به چالش اصلی دیگری در خصوص شناسایی ترافیک شبکه با استفاده از ویژگی‌های آماری پرداخته شد. این چالش وابستگی ویژگی‌های آماری به عوامل و رخداد‌های شبکه موجود بین فرستنده و گیرنده به خصوص اتلاف بسته‌ها در مسیر یاب‌ها و در طول مسیر می‌باشد. در هنگام وقوع اتلاف بسته، دنباله‌ای از بسته‌ها از دست می‌روند، و در نتیجه ویژگی‌های آماری مورد استفاده جهت شناسایی برنامه کاربردی نیز تغییر می‌کنند. از جمله این تغییرات می‌توان به تغییر زمان ارسال و دریافت بسته‌ها، تغییر در ترتیب بسته‌ها، تغییر در نرخ ارسال و دریافت بسته‌ها و مدت زمان جریان‌ها اشاره نمود. به همین منظور، پایگاه داده‌ای از هفت برنامه کاربردی نامبرده شده، با حالت‌های مختلف اتلاف بسته فراهم شد تا تأثیر اتلاف بسته بر روی ویژگی‌ها و دقت شناسایی بررسی شود. نتایج این بررسی نشان می‌دهد که با افزایش میزان اتلاف بسته در ترافیک شبکه دقت شناسایی نیز به میزان قابل توجهی کاهش پیدا می‌کند.

در ادامه کار با تحلیل فراوانی توزیع داده‌ها در چهار ویژگی اندازه بسته‌ها، فاصله زمانی مابین بسته‌های ارسالی و دریافتی، مدت زمان جریان بسته‌ها و نرخ ارسال و دریافت بسته‌ها، ویژگی‌هایی مقاوم به رخداد اتلاف بسته جهت شناسایی برنامه‌های کاربردی استخراج گردید. میزان تمایز دهندگی این ویژگی‌ها به قدری دقیق بود که با اعمال این ویژگی‌ها به یک شبکه عصبی مرسوم، عمل شناسایی ترافیک شبکه صورت گرفته و نتایج شناسایی در داده‌های با نرخ متفاوت اتلاف بسته به طور میانگین به ۹۱/۵٪ رسید. نتایج حاصل از بررسی روش ارائه شده نشان می‌دهد که با استفاده از ویژگی‌های استخراج شده، دقت شناسایی برنامه‌های کاربردی هنگام رخداد اتلاف بسته در شبکه نزدیک به دقت شناسایی در حالت عدم رخداد اتلاف بسته می‌باشد.

۳.۵ پیشنهادهایی برای ادامه کار

اگر چه نتایج نشان داده شده در این رساله کارایی روش ارائه شده را در شناسایی ترافیک شبکه در محیط‌های واقعی می‌رساند، اما بدلیل اهمیت بسیار زیاد موضوع و همچنین گستردگی کار در این حوزه لازم است تا موضوع شناسایی ترافیک شبکه از جنبه‌های دیگری مورد بررسی و تحلیل قرار گیرد. از اینرو پیشنهادهایی برای انجام ادامه کار در این زمینه قابل ارائه می‌باشد که این پیشنهادات عبارتند از:

- تحلیل و بررسی سایر برنامه‌های کاربردی همانطور که در فصل اول ذکر گردید، در این رساله تعداد هفت برنامه کاربردی مورد بررسی قرار گرفته است. این برنامه‌های کاربردی نماینده گروه‌های بسیار زیادی از برنامه‌های کاربردی هستند که امروزه در جامعه جایگاه بالایی در بین مردم پیدا کرده‌اند. در این رساله با تحلیل رفتار این هفت برنامه کاربردی ویژگی‌های آماری استخراج گردید که مقاوم در برابر چالش‌های موجود و موثر جهت شناسایی ترافیک شبکه هستند. با ویژگی‌های آماری ارائه شده، هفت برنامه کاربردی با دقت بالایی شناسایی شده و باقی برنامه‌های کاربردی در شبکه به عنوان کلاس هشتم در نظر گرفته می‌شوند، ممکن است این ویژگی‌ها برای برنامه‌های کاربردی دیگر این حد از تمایز را نداشته و لازم باشد تا ویژگی‌های دیگری استخراج گردد. لذا به عنوان پیشنهادی برای ادامه کار لازم است تا برنامه‌های کاربردی دیگری همچون برنامه‌های پیام رسان، برنامه‌های مدیریت ایمیل مورد تحلیل و بررسی جامعی قرار گرفته و عملکرد سامانه شناسایی ترافیک بر روی این برنامه‌های کاربردی نیز آزمایش گشته و در صورت لزوم ویژگی‌های دیگری نیز استخراج گردد.
- بررسی تاثیر رخداد های دیگر در شبکه بر روی دقت شناسایی

در این رساله تاثیرات رخداد اتلاف بسته‌ها بر روی ترافیک شبکه و سامانه تشخیص ترافیک شبکه مورد تحلیل و بررسی قرار گرفت و نشان داده شد که اتلاف بسته‌ها تاثیر چشمگیری بر دقت شناسایی دارد. در ساختار شبکه اینترنت، ترافیک شبکه با اجزاء و عوامل دیگری نیز مواجه می‌باشد، که این عوامل ممکن است بر ویژگی‌های آماری بسته‌ها تاثیراتی داشته باشند. از جمله این عوامل می‌توان به سازوکار دیوارهای آتش، سوئیچ‌های گوناگون، مسیرهای متفاوت و مسافت در شبکه اشاره نمود. به طور نمونه می‌توان به ساختار دیوارهای آتش اشاره نمود که این سیستم‌ها دارای صف‌هایی با اولویت‌های متفاوت می‌باشند که بسته‌ها در این صف‌ها قرار گرفته و پس از بررسی محتوای بسته اجازه عبور و یا عدم عبور صادر می‌شود. در اینصورت بایستی این مسئله مورد بررسی قرار گیرد که آیا عملیاتی که در این سیستم‌ها انجام می‌شود بر ویژگی‌های آماری بسته‌ها تاثیرگذار هستند یا خیر. به عنوان پیشنهادی برای کارهای آتی می‌توان بر روی تاثیرات احتمالی هر کدام از این عوامل بر روی ترافیک شبکه مطالعه و بررسی انجام داد.

- شناسایی برنامه‌های کاربردی در لایه‌های بالاتر

در این رساله از ویژگی‌های آماری لایه شبکه جهت شناسایی ترافیک شبکه استفاده گردید. دلیل این کار را عدم نیاز به پردازش بالا برای بازگشایی بسته‌ها، عدم رمزنگاری در این لایه و حفظ محرمانگی می‌توان اعلام کرد. روش‌های دیگری نیز در مقالات ارائه شده‌اند که از ویژگی‌های لایه‌های دیگر (مانند: اندازه پنجره ارسال در لایه چهارم) جهت شناسایی ترافیک شبکه استفاده کرده‌اند [۸۹]. این امر دارای چالش‌های دیگری همچون: میزان پردازش جهت بازگشایی لایه‌های بالاتر، استخراج ویژگی‌ها با وجود پروتکل‌های رمزنگاری همچون SSL, TLS در لایه‌های بالاتر و همچنین تاثیرات این الگوریتم‌ها بر ویژگی‌های آماری می‌باشد. به عنوان پیشنهادی برای ادامه کار می‌توان لایه‌های مختلف را مورد تحلیل قرار داد و در هر لایه ویژگی‌های آماری استخراج نمود که در برابر الگوریتم‌های رمزنگاری و همچنین چالش‌های مطرح در این رساله مقاوم باشند.

- شناسایی نوع محتوای

یکی از مباحث دیگر در حوزه شناسایی ترافیک شبکه، شناسایی نوع محتوای ارسال یا دریافت شده از بسته‌های ترافیک شبکه می‌باشد [۸۹]. منظور از شناسایی محتوای، تشخیص اینکه بسته‌های شبکه در حال انتقال متن، صوت، تصویر یا چندرسانه‌ای هستند. شناسایی محتوای را می‌توان بدون شناسایی برنامه کاربردی یا پس از شناسایی برنامه کاربردی انجام داد. به عنوان پیشنهادی

برای ادامه کار می‌توان با استفاده از تحلیل رفتاری، ویژگی‌های آماری استخراج نمود که بوسیله آنها بتوان نوع محتوای عبوری را با دقت بالایی شناسایی نمود.

- شناسایی تهدیدات و برنامه‌های مخرب

یکی دیگر از کاربردهای شناسایی ترافیک شبکه، شناسایی برنامه‌های کاربردی با رفتار مخرب می‌باشد. به عنوان نمونه می‌توان به حملات ¹DOS اشاره نمود [۹۰]. در این نوع از حملات حمله‌گر^۲ با استفاده از Botnetها در فواصل زمانی کم، بسته‌های متعددی را برای سرویس دهنده ارسال کرده و به موجب آن باعث از کار افتادن گیرنده یا سرویس دهنده می‌گردد. به عنوان پیشنهادی برای ادامه کار می‌توان رفتار این نوع حملات را مورد بررسی قرار داد و ویژگی‌هایی استخراج نمود که پیش از آنکه حمله‌گر بتواند رفتار مخربی داشته باشد ترافیک مربوط به این حملات را شناسایی نمود.

- در نظر گرفتن حالت برخط در شناسایی ترافیک

یکی دیگر از مسائل در شناسایی ترافیک شبکه بحث شناسایی برخط ترافیک در کمترین زمان و با کمترین تعداد بسته می‌باشد. کارهای زیادی در این حوزه انجام شده است ولی همچنان نمی‌توان تعیین نمود که برای انجام برخط شناسایی ترافیک شبکه چه معیارهایی را بایستی در نظر داشت و همچنین چه تعداد بسته از جریان لازم است تا بتوان ترافیک شبکه را شناسایی نمود [۸۶]. به عنوان پیشنهادی برای ادامه کار می‌توان در این زمینه که چه عواملی بر شناسایی برخط تاثیر گذار هستند و اینکه چه تعداد بسته برای شناسایی ترافیک با دقت بالا لازم است تحقیق و مطالعه انجام داد.

¹Denial Of Service

²Attacker

پیوست اول

مبانی شبکه

۱.آ ساختار شبکه

تعریف ۱.آ

شبکه‌های کامپیوتری مجموعه‌ای از کامپیوترهای مستقل و متصل به یکدیگرند، که با یکدیگر ارتباط داشته و تبادل داده می‌کنند.

♣ مستقل بودن کامپیوترها بدین معناست که، هر کدام دارای واحدهای کنترلی و پردازشی مجزا بوده و نبود یکی بر دیگری تاثیرگذار نیست.

♣ متصل بودن کامپیوترها یعنی از طریق یک رسانه فیزیکی مانند کابل، فیبر نوری، ماهواره‌ها و... به هم وصل می‌باشند.

توجه



دو شرط فوق شروط لازم برای ایجاد یک شبکه کامپیوتری می‌باشند، اما شرط کافی برای تشکیل یک شبکه کامپیوتری داشتن ارتباط و تبادل داده بین کامپیوترها است.

اینترنت شبکه ویژه‌ای از شبکه‌ها که حاصل اتصالات داخلی شبکه‌های دولتی، دانشگاهی، عمومی و خصوصی در سرتاسر دنیا است. این شبکه بر اساس شبکه اولیه‌ای کار می‌کند که ¹ARPANET نام داشت و به وسیله موسسه ²ARPA که وابسته به «وزارت دفاع ایالات متحده آمریکا» است ایجاد شد. همچنین منزلگاهی برای وب جهان گستر ³است. در لاتین واژه Internet برای نامیدن آن بکار می‌رود که برای اشتباه نشدن با معنی عام واژه «شبکه متصل» حرف اول را بزرگ می‌نویسند.

اعضای شبکه اینترنت یا شرکت‌های سرویس دهنده آنها از آدرس‌های IP استفاده می‌کنند. این آدرس‌ها از موسسات ثبت نام آدرس تهیه می‌شوند تا تخصیص آدرسها قابل کنترل باشد. همچنین «سرویس دهندگان اینترنت» و شرکت‌های بزرگ، اطلاعات مربوط به در دسترس بودن آدرس‌هایشان را بواسطه قرارداد ⁴BGP با دیگر اعضای اینترنت مبادله می‌کنند. همه شبکه‌ها از اجزای سخت‌افزاری پایه‌ای تشکیل شده‌اند تا گره‌های شبکه را به یکدیگر متصل کنند، مانند کارت‌های شبکه، تکرارکننده، هاب‌ها، پل‌ها، سوده‌ها و مسیریاب‌ها. علاوه بر این، بعضی روشها برای اتصال این اجزای سخت‌افزاری لازم است که معمولاً از کابل‌های الکتریکی استفاده می‌شود. مسیریاب‌ها تجهیزات شبکه‌ای هستند که

¹Advanced Research Projects Agency Network

²Advanced Research Projects Agency

³World Wide Web

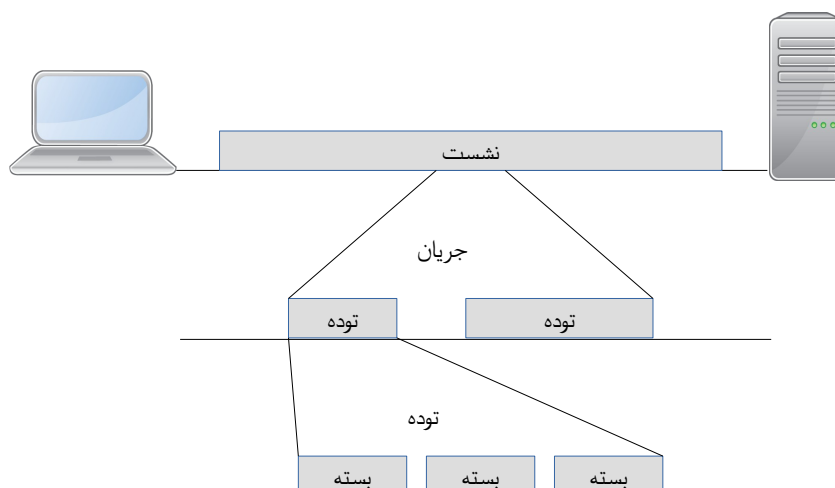
⁴Border Gateway Protocol

بسته‌های داده را با استفاده از سربسته^۵ و جدول ارسال تعیین مسیر کرده و ارسال می‌کنند. مسیریاب‌ها کار خود را در لایه شبکه انجام می‌دهند. همچنین اتصال بین بسترهای فیزیکی متفاوت را امکان‌پذیر می‌کنند. این کار با چک کردن سرایند یک بسته داده انجام می‌شود.

مسیریاب‌ها از قراردادهای مسیریابی مانند کوتاه‌ترین مسیر استفاده می‌کنند تا با یکدیگر گفتگو کرده و بهترین مسیر بین هر دو ایستگاه را پیکربندی کنند. هر مسیریاب حداقل به دو شبکه (معمولاً شبکه‌های محلی، شبکه‌های گسترده و یا یک شبکه محلی و یک سرویس دهنده اینترنت) متصل است. بعضی انواع مودم‌های DSL و کابلی جهت مصارف خانگی درون خود از وجود یک مسیریاب نیز بهره می‌برند. [۹۱]

۲.۱ سطوح مختلف ترافیک

شکل ۱.۱ سطوح مختلف ترافیک را نشان می‌دهد. در ابتدا باید برخی واژه‌ها را بازتعریف کنیم، چرا که ممکن است بررسی ترافیک در سطوح مختلفی انجام گیرد.



شکل ۱.۱: سطوح مختلف ترافیک

۱.۲.۱ جلسه

تعریف ۲.۱ به مجموعه عملیاتی که پس از برقراری یک ارتباط بین دو فرآیند و با یک توافق اولیه آغاز و سپس یک سری تراکنش ادامه می‌یابد و در انتها در روالی هماهنگ و مورد توافق ختم می‌شود، یک جلسه می‌گویند.

⁵Header

با هر تماس یک میزبان^۶ با میزبان دیگر، یک جلسه بین آنها برقرار می‌شود. به عبارت دیگر هر جلسه متناظر با هر فعالیت کاربر^۷ می‌باشد [۹۲]. مثال‌های زیر نمونه‌هایی از این دست می‌باشند.

♠ یک کاربر با بازکردن یک اتصال^۸ TCP روی درگاه ۲۱ خدمت‌گزار یک جلسه^۹ FTP را آغاز می‌کند. در پایان جلسه FTP، اتصال TCP نیز بسته می‌شود.

♠ بازه‌های زمانی^{۱۰} ای که یک کاربر Dial-Up به یک ISP^{۱۱} متصل است.

♠ برای شبکه‌ها^{۱۲} ی اتصال گرا^{۱۳} نظیر ATM^{۱۴} یک جلسه توسط سیگنال‌دهی^{۱۵} ابتدا و انتها مشخص می‌شود.

توجه



مدل‌سازی ترافیک در سطح جلسه عبارت است از دریافت مشخصات شروع و طول هر جلسه.

۲.۲.۱ جریان

در طول هر جلسه هر میزبان ممکن است یک یا چندین جریان بسته ارسال نماید [۹۲]. واژه جریان به معانی گوناگونی در مقالات برداشت شده است، لیکن برداشت کلی‌ای که می‌توان از جریان داشت یک سری از بسته‌های نزدیک به هم است که در یک مسیر مشخص بین یک جفت میزبان مشخص در حال انتقالند. بسته‌هایی موجود در یک جریان معمولاً دارای سربسته‌های یکسانی می‌باشند (مثل ID^{۱۶} پروتکل^{۱۷}، شماره درگاه و نشانی^{۱۸} مبداء و مقصد) که وجه تمایز هر جریان است. برای مثال هر جلسه FTP شامل دو جریان بسته می‌باشد. یک جریان مربوط به جویبار کنترلی^{۱۹} از طریق TCP و شماره درگاه ۲۱ و دیگری جویبار داده^{۲۰} که از طریق یک شماره درگاه TCP توافق شده ارسال می‌گردد.

توجه



مدل‌سازی ترافیک در سطح جریان عبارت است از دریافت مشخصات شبه تصادفی شروع و طول هر جریان.

⁶Host

⁷User

⁸Connection

⁹File Transfer Protocol

¹⁰Time Interval

¹¹Internet Service Provider

¹²Network

¹³Connection-Oriented

¹⁴Asynchronous Transfer Mode

¹⁵Signaling

¹⁶Identification

¹⁷Protocol

¹⁸Address

¹⁹Control Stream

²⁰Data Stream

مثالی که برای جریان به کار می‌رود، فیل و موش است (با توجه به سائز آن‌ها). موش به فاز شروع آهسته^{۲۱} (افزایش نرخ^{۲۲} تا وقوع اولین اتلاف بسته) و فیل به پیش‌گیری از ازدحام^{۲۳} اشاره دارد. اصطلاح دیگری که کمتر به کار می‌رود عبارت است از آسیابک^{۲۴} که برای جریان‌های کمتر از ۲ ثانیه و لاک‌پشتی^{۲۵} که برای جریان‌های بیشتر از ۱۵ ثانیه به کار می‌رود [۹۳]. اندازه‌گیری‌ها نشان می‌دهد که بین ۴۰ تا ۷۰ درصد ترافیک اینترنت (تا سال ۲۰۰۷) مربوط به جریان‌های کوچک بوده که مشخصاً ترافیک وب^{۲۶} می‌باشند (ترافیک ناشی از مشاهده یک تارنما^{۲۷} توسط کاربران). جریان طولانی معمولاً شامل ترافیک غیر وب^{۲۸} بوده و قسمتی کوچکی از ترافیک کل را تشکیل می‌دهند (تا سال ۲۰۰۷) گرچه به علت طولانی بودن آن (از چند ساعت تا چند روز) اهمیت فراوانی دارند.

۳.۲.۱ توده

هر جریان می‌تواند چندین ترافیک توده متناوب را شامل شود که هر یک شامل بسته‌های ارسالی پشت‌سرهم می‌باشد. از این دست، می‌توان به مثال‌های زیر اشاره کرد:

- ✓ ترافیک توده معمولاً در پروتکل‌های مبتنی بر پنجره^{۲۹} رخ می‌دهد، جایی که میزبان مجاز به ارسال یک پنجره^{۳۰} از داده‌هاست و برای ارسال پنجره بعدی باید منتظر دریافت اعتبار جدید باشد.
- ✓ در جلسه FTP اگر فایل بزرگ باشد، به چندین بسته بخش‌بندی^{۳۱} می‌شود.
- ✓ صحبت جهشی^{۳۲} که بیانگر صحبت متوالی و پشت‌سرهم یک نفر در یک مکالمه صوتی است و منجر به تولید توده می‌شود. حال آنکه در حالت عادی، مکالمه شامل شنیدن و صحبت کردن رفت و برگشتی است.

۴.۲.۱ بسته

نهایتاً می‌توان ترافیک را از دیدگاه تک تک بسته‌ها بررسی کرد.

توجه



در سطح بسته تنها فرایند^{۳۳} ورودی^{۳۴} بسته‌ها بررسی شده و به ساختار بالاتر ترافیک یعنی توده، جریان و جلسه آن توجهی نمی‌شود.

²¹Slow-Start

²²Rate

²³Congestion Avoidance

²⁴Dragonfly

²⁵Tortoises

²⁶Web Traffic

²⁷Website

²⁸Non-Web Traffic

²⁹Window-Based

³⁰Window

³¹Segmentation

³²Talkspurt

بیشتر توجهات در مقالات و بالتبع در این پژوهش روی این سطح از ترافیک است. مطالعه و بررسی سطوح بالاتر نیازمند جمع‌آوری و تحلیل حجم بیشتری از داده است. چراکه جریان و جلسه در طول دقیق و ساعت‌های متمادی تولید شده است که نیازمند جمع‌آوری می‌باشند. از کارهایی که در زمینه مورد اخیر انجام گرفته می‌توان به تحلیل چندین ساعت ترافیک توسط [۹۴] و یا مطالعه روی داده جمع‌آوری شده طی ۸ روز در [۹۳] اشاره کرد.

۳.آ تبادل داده در شبکه

جهت تبادل داده در شبکه لازم است ابتدا بین دو ماشین ارتباط برقرار شود. برای ایجاد یک ارتباط بین دو ماشین در یک شبکه، سه روش مختلف وجود دارد:

♠ سوئیچینگ مداری^{۳۵}: در این روش برای انتقال اطلاعات بین دو ماشین ابتدا یک اتصال فیزیکی بین مبدأ و مقصد برقرار می‌شود.

♠ سوئیچینگ پیام^{۳۶}: در روش سوئیچ پیام که صرفاً مختص انتقال داده‌های دیجیتالی است، هر ایستگاه یک اتصال دائمی و همیشگی وصل با مرکز سوئیچ خود دارد. مرکز سوئیچ یک کامپیوتر با تعداد زیادی درگاه دیجیتال ورودی/خروجی است که دارای حافظه اصلی و حافظه جانبی می‌باشد. هر ایستگاه به محض آنکه تمایل به ارسال داشته باشد، با اضافه کردن اطلاعات لازم به ابتدای داده‌ها، آنرا در قالب یک پیام تحویل مرکز سوئیچی که به آن متصل است می‌دهد.

♠ سوئیچینگ بسته^{۳۷} و سوئیچینگ سلول^{۳۸}: مشکلات ناشی از عدم محدودیت طول پیام، باعث شد که در روشهای جدید بر روی اندازه پیام ارسالی محدودیت گذاشته شود و ایستگاه اجازه نداشته باشد در هر بار ارسال، کل یک پیام بزرگ را یکجا بفرستد، بلکه موظف است آنرا به قطعات کوچکتری به نام بسته تقسیم کرده و ضمن اضافه کردن اطلاعات لازم برای بازسازی اصل پیام به هر بسته، آنها را بطور جداگانه به مرکز سوئیچ ارسال کند [۹۱].

³³Process

³⁴Arrival

³⁵Circuit Switching

³⁶Message Switching

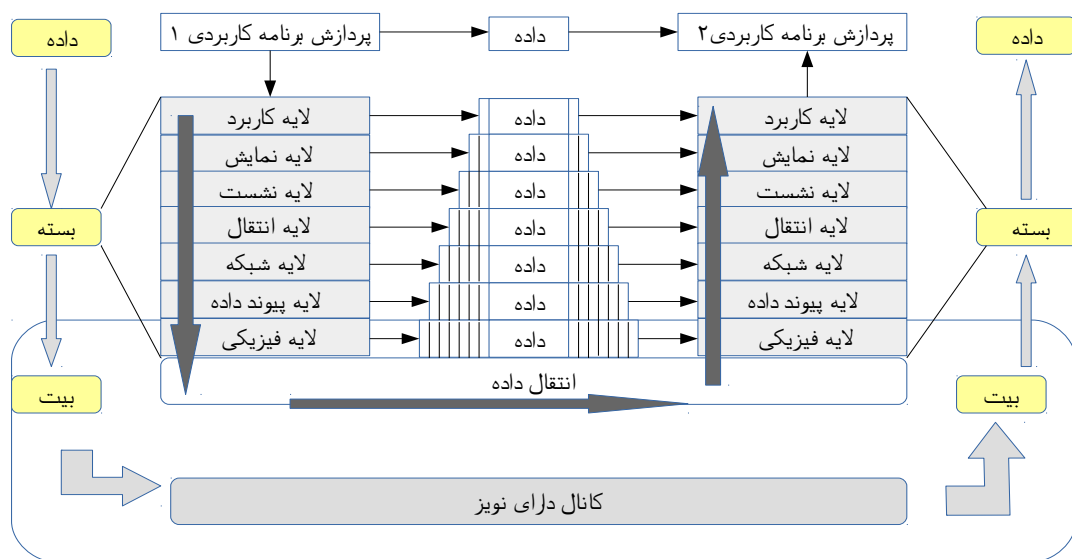
³⁷Packet Switching

³⁸Cell Switching



طراح یک شبکه باید تمام مسائل شبکه را تجزیه و تحلیل کرده و برای آنها راه حل ارائه کند ولی چون این مسائل دارای ماهیتی متفاوت از یکدیگر هستند، بنابراین طراحی یک شبکه باید بصورت لایه به لایه انجام شود.

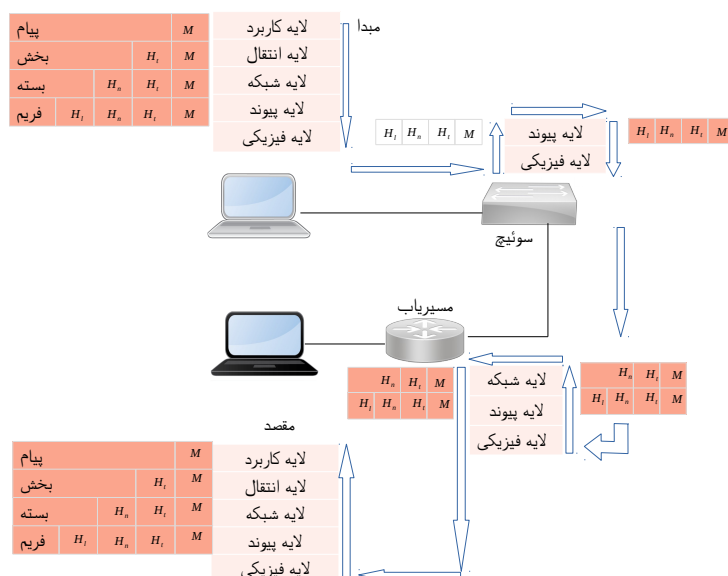
فرض کنید کاربر برنامه کاربردی X را بر روی رایانه خود اجرا می کند. بر طبق مدل هفت لایه ای شبکه های رایانه ای [۶]، هر لایه پیام های لایه بالاتر را دریافت می کند، و پس از اجرای برخی پردازش ها بر روی آن، مقداری داده به عنوان سربسته به آن اضافه کرده، و آن را به سوی لایه پایین تر رهسپار می کند. در نهایت در لایه فیزیکی پیام نهایی به صورت امواجی به سوی گیرنده رهسپار می گردد. شکل ۲.آ نمایش کلی از ارسال داده را در مدل هفت لایه ای شبکه نمایش می دهد. نکته حائز اهمیت این موضوع می باشد که



شکل ۲.آ: ساختار پیام در مدل هفت لایه شبکه های رایانه ای [۶]

تحلیل گر ترافیک می تواند به شرط عدم رمزنگاری، بسته ها را تا لایه کاربرد بازگشایی کند و از محتوای آن آگاه شود. وظیفه اصلی مسیر یاب تعیین مسیر بسته ها با استفاده از جدول ارسال می باشد. در شبکه نیز تعیین مسیر توسط آدرس IP صورت می گیرد. در نتیجه مسیر یاب های شبکه، کار خود را در لایه شبکه انجام داده و بسته ها تا لایه شبکه بازگشایی می شوند (شکل ۳.آ). این بدان معناست که، بررسی محتویات بسته تا لایه شبکه سربار چندانی نخواهد داشت. اما اگر تحلیل گر ترافیک، لایه های بالاتر را نیز بازگشایی کند، بایستی به جهت بازگشایی هر لایه و خواندن سرآیند و محتوی بسته ها، متحمل سربار محاسباتی و

زمانی زیادی شود.



شکل ۳.آ: مسیریاب‌های بازگشایی پیام‌ها در مسیریاب تا لایه شبکه (شکل 1.24 مرجع [۶])

۱.۳.آ پروتکل TCP

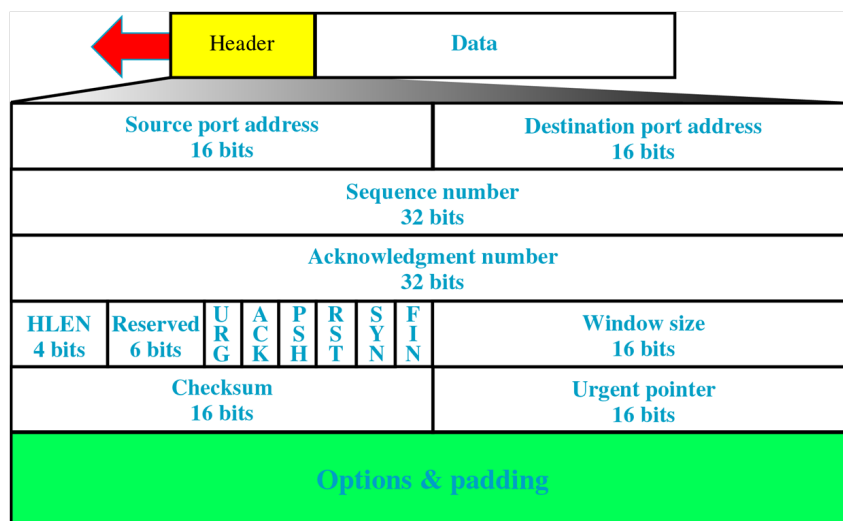
مهمترین وظیفه پروتکل TCP اطمینان از صحت ارسال اطلاعات است. پروتکل فوق اصطلاحاً "ارتباط اتصال گرا" نامیده می‌شود. علت این امر ایجاد یک ارتباط مجازی بین کامپیوترهای فرستنده و گیرنده بعد از ارسال اطلاعات است. پروتکل‌هایی از این نوع، امکانات بیشتری را به منظور کنترل خطاهای احتمالی در ارسال اطلاعات فراهم نموده ولی بدلیل افزایش بار عملیاتی سامانه کارائی آنان کاهش خواهد یافت. از پروتکل TCP به عنوان یک پروتکل قابل اطمینان نیز یاد می‌شود. به این علت که برای آگاهی از صحت اطلاعات ارسال شده، اطلاعات دیگری نیز به گیرنده فرستاده می‌شود. در صورتی که بسته‌های اطلاعاتی بدرستی در اختیار فرستنده قرار نگیرند، فرستنده مجدداً اقدام به ارسال اطلاعات می‌نماید. شکل ۴.آ

نمایی از سرآیند بسته‌هایی که توسط پروتکل TCP ساخته می‌شوند را نشان می‌دهد.

۱.۱.۳.آ دست‌دهی^{۳۹}

در هر اتصال رایانه‌ای مقداری بار اضافی وجود دارد که در اصطلاح دست‌دهی نامیده می‌شود و بدین معنی است که مودم از کامپیوتر سرور سوال می‌کند، آیا داده‌ها را دریافت کرده است و سرور پاسخ مثبت یا منفی می‌دهد. دست‌دهی سه مرحله‌ای برای برقراری اتصال در پروتکل TCP استفاده می‌شود. البته ابتدا طرفین باید آماده باشند و بعد بین طرفین مذاکرات سه مرحله‌ای زیر صورت می‌گیرد.

³⁹Handshaking



شکل آ.۴: سرآیند بسته‌های پروتکل TCP

📌 در مرحله اول طرف شروع کننده ارتباط، یک بسته TCP (خالی از داده) ارسال می‌کند که در آن بیت $SYN=1$ و بیت $ACK=0$ است و درون فیلد شماره ترتیب عدد x قرار داده شده که در آن x یک عدد تصادفی است. در حقیقت با این شماره به طرف مقابل اطلاع داده می‌شود که بطور قرار دادی ترتیب داده‌های ارسالی بجای 0 از شماره $x+1$ شروع می‌شود. در پروتکل TCP شماره ترتیب ۳۲ بیتی است لذا برای پیشگیری از مشکلات احتمالی ناشی از مساوی بودن شماره ترتیب بسته‌های ارسالی، داده‌ها از شماره ۰ شروع نمی‌شود، بلکه از یک عدد تصادفی (که به صورت خودکار تولید می‌شود)، شروع خواهد شد و در همان مرحله اول، این شماره ترتیب به طرف مقابل اعلام می‌گردد.

📌 در مرحله دوم طرف مقابل با دریافت تقاضای فوق اگر تمایل به برقراری ارتباط نداشته باشد با ارسال یک بسته خالی که در آن بیت RST به 1 تنظیم شده، این تقاضا را رد می‌کند ولی اگر تمایل به برقراری ارتباط بود یک بسته خالی از داده با مشخصات زیر تولید می‌کند:

بیت SYN را یک می‌کند. بیت ACK را یک می‌کند. مقدار فیلد Acknowledgement Number را $x+1$ قرار می‌دهد. مقدار فیلد Sequence Number را مقدار تصادفی y قرار می‌دهد.

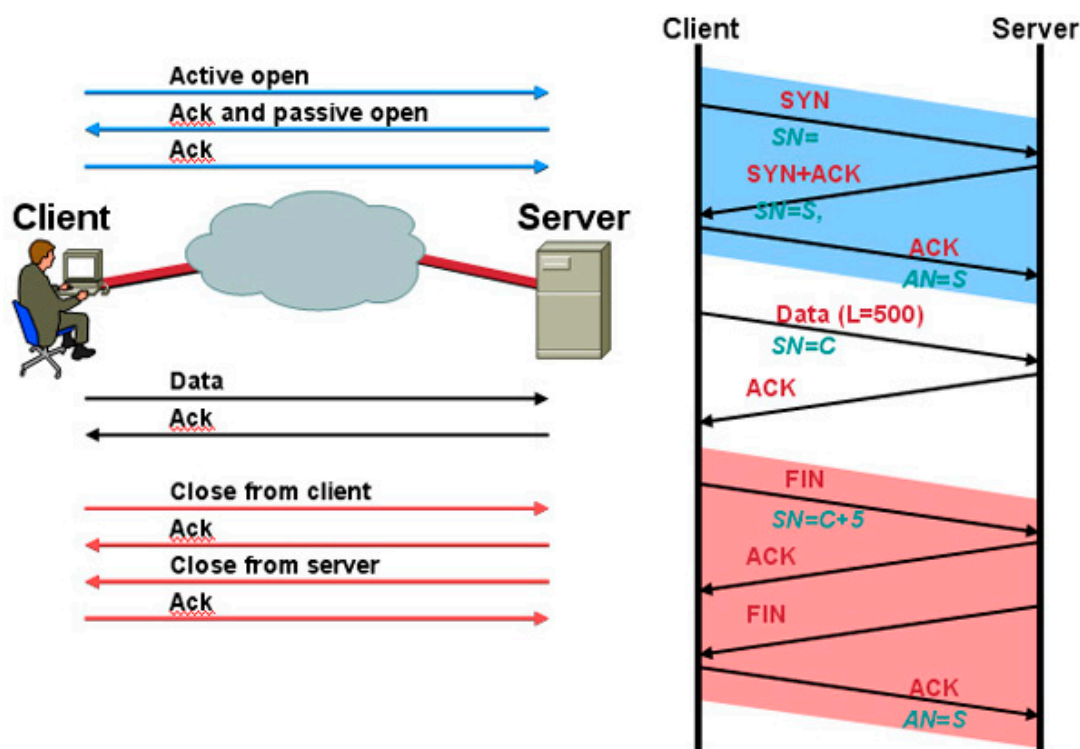
در این مرحله که به معنای پذیرش ارتباط است، طرف مقابل با قرار دادن مقدار فیلد $ACK=x+1$ نشان می‌دهد که شماره ترتیب x را پذیرفته و منتظر داده‌ها از شماره ترتیب $x+1$ به بعد است. در ضمن خودش عدد تصادفی x را در فیلد Seq. No قرار می‌دهد و به طرف مقابل اعلام می‌کند که شماره ترتیب داده‌های ارسالی از y خواهد بود.

📌 مرحله سوم شروع کننده ارتباط با قرار دادن مقادیر زیر شروع ارتباط را تصدیق می‌کند:

بیت SYN را صفر می‌کند. بیت ACK را یک می‌کند. فیلد Seq. No.=x+1 را قرار می‌دهد. فیلد ACK را برابر y+1 قرار می‌دهد.

در حقیقت با قرار دادن Seq. No.=x+1 و ACK=y+1 شروع کننده ارتباط اعلام می‌کند که بر روی پارامترهای شماره ترتیب توافق شده است و او پذیرفته که داده‌های طرف مقابل را از شماره ۱+y بپذیرد. پس از این مرحله، ارسال و دریافت داده‌ها توسط طرفین تا هنگامی که ارتباط با اطلاع طرفین خاتمه داده نشده است آزاد است.

برای خاتمه ارتباط روند زیر صورت می‌گیرد: طرفی که داده‌هایش برای ارسال تمام شده است یک بسته TCP ارسال می‌نماید که در سرآیند آن بیت FIN را یک قرار داده است. طرف مقابل این درخواست را دریافت و با ختم یک طرفه آن موافقت می‌کند. ولی چون ارتباط به صورت یک طرفه ختم می‌شود، طرف مقابل می‌تواند تا جایی که داده دارد آن‌ها را ارسال کند و نهایتاً در آخرین بسته، بیت FIN را یک بگذارد تا پس از تصدیق آن، ارتباط به صورت دوطرفه ختم شود (شکل ۵.آ).



شکل ۵.آ: شروع ارتباط در پروتکل TCP

۲.۳.آ پروتکل UDP

این پروتکل نظیر پروتکل TCP در لایه انتقال فعالیت می‌نماید. UDP برخلاف پروتکل TCP بصورت بدون اتصال است. بدیهی است که سرعت پروتکل فوق نسبت به TCP سریعتر بوده ولی از بعد کنترل خطا

تنظیمات لازم را ارائه نخواهد داد. بهترین جایگاه استفاده از پروتکل فوق در مواردی است که برای ارسال و دریافت اطلاعات به یک سطح بالا از اطمینان، نیاز نداشته باشیم.

مراجع

- [1] H. R. Loo, S. B. Joseph, and M. N. Marsono, "Online incremental learning for high bandwidth network traffic classification," *Applied Computational Intelligence and Soft Computing*, vol.2016, p.1, 2016.
- [2] T. Qin, L. Wang, Z. Liu, and X. Guan, "Robust application identification methods for p2p and voip traffic classification in backbone networks," *Knowledge-Based Systems*, vol.82, pp.152–162, 2015.
- [3] L. Peng, B. Yang, Y. Chen, and Z. Chen, "Effectiveness of statistical features for early stage internet traffic identification," *International Journal of Parallel Programming*, vol.44, no.1, pp.181–197, 2016.
- [4] B. Yamansavascular, M. A. Guvensan, A. G. Yavuz, and M. E. Karsligil, "Application identification via network traffic classification," in *Computing, Networking and Communications (ICNC), International Conference on*, pp.843–848, IEEE, 2017.
- [5] Z. Chen, Z. Liu, L. Peng, L. Wang, and L. Zhang, "A novel semi-supervised learning method for internet application identification," *Soft Computing*, vol.21, no.8, pp.1963–1975, 2017.
- [6] J. F. Kurose and K. W. Ross. *Computer Networking: A Top-down Approach*. Pearson Education, 6 ed. , 2012.
- [7] J. Muehlstein, Y. Zion, M. Bahumi, I. Kirshenboim, R. Dubin, A. Dvir, and O. Pele, "Analyzing https encrypted traffic to identify user's operating system, browser and application," pp.1–6, 2017.
- [8] M. S. Aliakbarian, A. Fanian, F. S. Saleh, and T. A. Gulliver, "Optimal supervised feature extraction in internet traffic classification," pp.102–107, 2013.
- [9] R. Alshammari and A. N. Zincir-Heywood, "Machine learning based encrypted traffic classification: Identifying ssh and skype," pp.1–8, 2009.
- [10] F. Ertam and E. Avci, "A new approach for internet traffic classification: Ga-wk-elm," *Measurement*, vol.95, pp.135–142, 2017.
- [11] G. Sun, T. Chen, Y. Su, and C. Li, "Internet traffic classification based on incremental support vector machines," *Mobile Networks and Applications*, vol.23, pp.1–8, 2018.
- [12] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol.160, pp.3–24, 2007.
- [13] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," *ACM SIGCOMM Computer Communication Review*, vol.37, no.1, pp.5–16, 2007.
- [14] D. J. Daley, "Revisiting queueing output processes: a point process viewpoint," *Queueing Systems*, vol.68, no.3-4, p.395, 2011.
- [15] D. J. Daley and D. N. Shanbhag, "Independent Inter-Departure Times in M/G/1/N Queues," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol.37, no.2, pp.259–263, 1975.
- [16] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, and Y. Guan, "Network traffic classification using correlation information," *IEEE Transactions on Parallel and Distributed Systems*, vol.24, no.1, pp.104–117, 2013.

- [17] P. Foremski, "On different ways to classify internet traffic: a short review of selected publications," *Theoretical and Applied Informatics*, vol.25, no.2, pp.119–136, 2013.
- [18] Y. Wang, Y. Xiang, and J. Zhang, "Network traffic clustering using Random Forest proximities," *IEEE International Conference on Communications (ICC)*, no.60970146, pp.2058–2062, 2013.
- [19] D. Adami, C. Callegari, S. Giordano, M. Pagano, and T. Pepe, "Skype-Hunter: A real-time system for the detection and classification of Skype traffic," *International Journal of Communication Systems*, vol.25, pp.386–403, Mar. 2012.
- [20] A. Finamore, M. Mellia, M. Meo, and D. Rossi, "KISS: Stochastic packet inspection Classifier for UDP traffic," *IEEE/ACM Transactions on Networking*, vol.18, no.5, pp.1505–1515, 2010.
- [21] S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, M. Mellia, and A. Pescapè, "Re-viewing Traffic Classification," in *Data Traffic Monitoring and Analysis SE - 6* (E. Biersack, C. Callegari, and M. Matijasevic, eds.), vol.7754 of *Lecture Notes in Computer Science*, pp.123–147, Springer Berlin Heidelberg, 2013.
- [22] M. Roughan and S. Sen, "Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification," *Proceedings of the 4th ...*, pp.135–148, 2004.
- [23] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in *Proceedings of the 13th international conference on World Wide Web*, pp.512–521, ACM, 2004.
- [24] A. W. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques Categories and Subject Descriptors," *Sigmetrics*, pp.50–60, 2005.
- [25] T. AbuHmed, A. Mohaisen, and D. Nyang, "A Survey on Deep Packet Inspection for Intrusion Detection Systems," *Information Security*, vol.24, no.1, p.10, 2008.
- [26] A. Finamore, M. Mellia, M. Meo, and D. Rossi, "KISS: Stochastic packet inspection Classifier for UDP traffic," *IEEE/ACM Transactions on Networking*, vol.18, no.5, pp.1505–1515, 2010.
- [27] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy, "Transport layer identification of P2P traffic," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement - IMC '04*, (New York, New York, USA), p.121, ACM Press, 2004.
- [28] T. Karagiannis, K. Papagiannaki, N. Taft, and M. Faloutsos, "Profiling the End Host," in *Proceedings of the 8th International Conference on Passive and Active Network Measurement, PAM'07*, (Berlin, Heidelberg), pp.186–196, Springer-Verlag, 2007.
- [29] V. Carela-Español, P. Barlet-Ros, M. Solé-Simó, A. Dainotti, W. de Donato, and A. Pescapé, "K-dimensional Trees for Continuous Traffic Classification," in *Proceedings of the Second International Conference on Traffic Monitoring and Analysis, TMA'10*, (Berlin, Heidelberg), pp.141–154, Springer-Verlag, 2010.
- [30] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification," *ACM SIGCOMM Computer Communication Review*, vol.36, no.5, pp.5–16, 2006.
- [31] A. Moore, D. Zuev, and M. Crogan, "Discriminators for use in flow-based classification," *Computer*, no.8, 2013.
- [32] G. Kesidis. *An introduction to communication network analysis*. John Wiley & Sons, 2007.
- [33] Z. Cao, G. Xiong, Y. Zhao, Z. Li, and L. Guo, "A survey on encrypted traffic classification," in *International Conference on Applications and Techniques in Information Security*, pp.73–81, 2014.
- [34] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Solé-Pareta, "Analysis of the impact of sampling on NetFlow traffic classification," *Computer Networks*, vol.55, no.5, pp.1083–1099, 2011.
- [35] A. Dainotti, A. Pescapé, and K. C. Claffy, "Issues and future directions in traffic classification," *Network, IEEE*, vol.26, no.1, pp.35–40, 2012.

- [36] S. Waichal and B. Meshram, "Router attacks detection and defense mechanisms," *Int. J. Sci. Technol. Res*, vol.2, pp.145–149, 2013.
- [37] T. W. Doeppner, P. N. Klein, and A. Koyfman, "Using router stamping to identify the source of ip packets," in *Proceedings of the 7th ACM conference on Computer and communications security*, pp.184–189, ACM, 2000.
- [38] D. Strom, "The packet filter: a basic network security tool," 2000.
- [39] G. Zeng, *Independent component analysis for data mining*. Ph.D. thesis, Rensselaer Polytechnic Institute, 2007.
- [40] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys and Tutorials*, vol.10, no.4, pp.56–76, 2008.
- [41] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning series, MIT Press, 2012.
- [42] J. Kim, J. Hwang, and K. Kim, "High-Performance Internet Traffic Classification Using a Markov Model and Kullback-Leibler Divergence," *Mobile Information Systems*, vol.1, pp.0–4, 2016.
- [43] W. Zhou, L. Dong, L. Bic, M. Zhou, and L. Chen, "Internet traffic classification using feed-forward neural network," in *Computational Problem-Solving (ICCP), 2011 International Conference on*, pp.641–646, IEEE, 2011.
- [44] Z. Chen, L. Peng, C. Gao, B. Yang, Y. Chen, and J. Li, "Flexible neural trees based early stage identification for IP traffic," *Soft Computing*, vol.21, no.8, pp.2035–2046, 2017.
- [45] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques," in *International Workshop on Passive and Active Network Measurement*, pp.205–214, Springer, 2004.
- [46] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *ACM SIGMETRICS Performance Evaluation Review*, vol.33, pp.50–60, ACM, 2005.
- [47] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," *ACM SIGCOMM Computer Communication Review*, vol.36, no.2, pp.23–26, 2006.
- [48] J. Park, H.-R. Tyan, and C.-C. J. Kuo, "Internet traffic classification for scalable qos provision," in *Multimedia and Expo, 2006 IEEE International Conference on*, pp.1221–1224, IEEE, 2006.
- [49] K. Singh, S. Agrawal, and B. Sohi, "A near real-time ip traffic classification using machine learning," *International Journal of Intelligent Systems and Applications*, vol.5, no.3, p.83, 2013.
- [50] L. Hu and L. Zhang, "Real-time internet traffic identification based on decision tree," in *World Automation Congress (WAC), 2012*, pp.1–3, IEEE, 2012.
- [51] D. Bonfiglio, M. Mellia, M. Meo, D. Rossi, and P. Tofanelli, "Revealing skype traffic: when randomness plays with you," *SIGCOMM Comput. Commun. Rev.*, vol.37, no.4, pp.37–48, 2007.
- [52] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," in *Proceedings of the 2006 SIGCOMM workshop on Mining network data*, pp.281–286, ACM, 2006.
- [53] T. T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks," in *Proceedings. 31st IEEE Conference on Local Computer Networks*, pp.369–376, IEEE, 2006.
- [54] T. Nguyen and G. Armitage, "Synthetic sub-flow pairs for timely and stable ip traffic identification," in *Australian Telecommunication Networks and Application Conference*, 2006.
- [55] J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, "Identifying and discriminating between web and peer-to-peer traffic in the network core," *Proceedings of the 16th international conference on World Wide Web - WWW '07*, p.883, 2007.

- [56] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," *Neural Networks, IEEE Transactions on*, vol.18, no.1, pp.223–239, 2007.
- [57] J. Eerman, A. Mahanti, and M. Arlitt, "Internet traffic identification using machine learning techniques," *Proc. of the 49th IEEE GLOBECOM. San Francisco*, 2006.
- [58] K. Singh and S. Agrawal, "Performance evaluation of five machine learning algorithms and three feature selection algorithms for ip traffic classification," *IJCA Special Issue on Evolution in Networks and Computer Communications* (1), pp.25–32, 2011.
- [59] J. Zhang, C. Chen, Y. Xiang, W. Zhou, and Y. Xiang, "Internet traffic classification by aggregating correlated naive bayes predictions," *IEEE Transactions on Information Forensics and Security*, vol.8, no.1, pp.5–15, 2013.
- [60] H. A. H. Ibrahim, S. M. Nor, A. Mohammed, A. B. Mohammed, *et al.*, "Taxonomy of machine learning algorithms to classify real time interactive applications," *Int. J. Comput. Networks Wireless Commun*, vol.2, pp.69–73, 2012.
- [61] D. Shi, Z. DingDing, and D. Wei, "The study of network traffic identification based on machine learning algorithm," in *Computational Intelligence and Communication Networks (CICN), Fourth International Conference on*, pp.205–208, IEEE, 2012.
- [62] A. B. Mohammed and S. Nor, "Near real time online flow-based internet traffic classification using machine learning (c4. 5)," *International Journal of Engineering (IJE)*, vol.3, no.4, p.370.
- [63] M. Soysal and E. G. Schmidt, "Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison," *Performance Evaluation*, vol.67, no.6, pp.451–467, 2010.
- [64] M. Hassan and M. N. Marsono, "A hybrid heuristics-statistical peer-to-peer traffic classifier," in *Computer Systems and Industrial Informatics (ICCSII), International Conference on*, pp.1–6, IEEE, 2012.
- [65] D. Binkley, "Source code analysis: A road map," in *Future of Software Engineering. FOSE'07*, pp.104–119, IEEE, 2007.
- [66] F. Nielson, H. R. Nielson, and C. Hankin. *Principles of program analysis*. Springer, 2015.
- [67] A. Ahmad and F. Farrukh, "Significance of social applications on a mobile phone for english task-based language learning," *Teaching English with Technology*, vol.15, no.2, pp.94–105, 2015.
- [68] L. Rosenfeld and P. Morville. *Information architecture for the world wide web*. O'Reilly Media, Inc., 2002.
- [69] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.27, no.8, pp.1226–1238, 2005.
- [70] D. E. McDysan. *QoS & traffic management in IP & ATM networks*, vol.18. McGraw-Hill New York, NY, USA, 2000.
- [71] M. Jain, D. S. Tomar, and S. K. Singh, "A survey on tcp congestion control schemes in guided media and unguided media communication," *International Journal of Computer Applications*, vol.118, no.3, 2015.
- [72] M. A. Kafi, D. Djenouri, J. Ben-Othman, and N. Badache, "Congestion control protocols in wireless sensor networks: A survey," *IEEE communications surveys & tutorials*, vol.16, no.3, pp.1369–1390, 2014.
- [73] N. A. Saqib, Y. SHAKEEL, M. A. Khan, H. MEHMOOD, and M. Zia, "An effective empirical approach to voip traffic classification," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol.25, no.2, pp.888–900, 2017.
- [74] J.-C. Bolot, "End-to-end packet delay and loss behavior in the internet," in *ACM SIGCOMM Computer Communication Review*, vol.23, pp.289–298, ACM, 1993.
- [75] M. Gandomi and H. Hassanpour, "Behavioral analysis of traffic flow for an effective network traffic identification," *International Journal of Engineering-Transactions B: Applications*, vol.30, no.11, p.1740, 2017.

- [76] P. Santiago del Rio, J. Ramos, J. Garcia-Dorado, J. Aracil, A. Cuadra-Sanchez, and M. Cutanda-Rodriguez, "On the processing time for detection of Skype traffic," in *7th International Wireless Communications and Mobile Computing Conference*, pp.1784–1788, July 2011.
- [77] L. Ptáček, "Analysis and detection of Skype network traffic," *Masaryk University Faculty of Informatics, Master Thesis, Brno, Czech Republic, Spring*, 2011.
- [78] D. Adami, C. Callegari, S. Giordano, M. Pagano, and T. Pepe, "A Real-Time Algorithm for Skype Traffic Detection and Classification," *Smart Spaces and Next Generation Wired/Wireless Networking Lecture Notes in Computer Science*, vol.5764, pp.168–179, 2009.
- [79] D. Bonfiglio, M. Mellia, M. Meo, and D. Rossi, "Detailed Analysis of Skype Traffic," *IEEE Transactions on Multimedia*, vol.11, pp.117–127, Jan. 2009.
- [80] T. Silverston, O. Fourmaux, A. Botta, A. Dainotti, A. Pescapé, G. Ventre, and K. Salamatián, "Traffic analysis of peer-to-peer IPTV communities," *Computer Networks*, vol.53, pp.470–484, Mar. 2009.
- [81] F. I. Khan, "A Generic Technique for Voice over Internet Protocol (VoIP) Traffic Detection," *IJCSNS*, vol.8, no.2, p.52, 2008.
- [82] B. Clarke, E. Fokoue, and H. H. Zhang. *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics, Springer, 2009.
- [83] R. Peck, C. Olsen, and J. L. Devore. *Introduction to statistics and data analysis*. Cengage Learning, 2015.
- [84] P. S. Mann. *Introductory Statistics*. John Wiley & Sons, 2010.
- [85] A. Este, F. Gringoli, and L. Salgarelli, "On the Stability of the Information Carried by Traffic Flow Features at the Packet Level," *ACM SIGCOMM Computer Communication Review*, vol.39, no.3, p.13, 2009.
- [86] L. Peng, B. Yang, and Y. Chen, "Effective packet number for early stage internet traffic identification," *Neurocomputing*, vol.156, pp.252–267, 2015.
- [87] M. Hall, G. Holmes, B. Pfahringer, I. Witten, M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software : An update," *ACM SIGKDD explorations newsletter*, vol.11, no.1, pp.10–18, 2009.
- [88] Y. Wang, Y. Xiang, and S. Yu, "Internet Traffic Classification Using Machine Learning: A Token-based Approach," *14th IEEE International Conference on Computational Science and Engineering*, pp.285–289, 2011.
- [89] B. Li, M. Ma, and Z. Jin, "A voip traffic identification scheme based on host and flow behavior analysis," *Journal of Network and Systems Management*, vol.19, no.1, pp.111–129, 2011.
- [90] D. Zhao, I. Traore, B. Sayed, W. Lu, S. Saad, A. Ghorbani, and D. Garant, "Botnet detection based on traffic behavior analysis and flow intervals," *Computers & Security*, vol.39, pp.2–16, 2013.
- [91] C. M. Kozierok. *The TCP/IP guide: a comprehensive, illustrated Internet protocols reference*. No Starch Press, 2005.
- [92] J. W. Roberts, "Internet traffic, qos, and pricing," *Proceedings of the IEEE*, vol.92, no.9, pp.1389–1399, 2004.
- [93] N. Brownlee and K. C. Claffy, "Understanding internet traffic streams: Dragonflies and tortoises," *IEEE Communications Magazine*, vol.40, no.10, pp.110–117, 2002.
- [94] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski, "Modeling Internet backbone traffic at the flow level," *IEEE Transactions on Signal Processing*, vol.51, no.8, pp.2111–2124, 2003.

واژه‌نامه فارسی به انگلیسی

پویا Dynamic	آسیابک Dragonfly
پیش‌گیری از ازدحام . . Congestion Avoidance	آمار Statistical
پیوند Link	اترنت Ethernet
ت	
تاخیر هدایت پیشرو Forwarding delay	اتصال مداوم Persistent Connection
تارنما Website	اتصال‌گرا Connection-Oriented
تحلیل‌گر ترافیک Traffic Analyser	اتلاف بسته Packet Loss
تخصیص منابع Resource Allocation	ازدحام Congestion
ترافیک غیر وب Non-Web Traffic	اطلاعات Information
ترافیک وب Web Traffic	اطلاعات متقابل Mutual Information
تشخیص ناهنجاری Anomaly Detection	انترنوی Entropy
تصویر خاکستری Gray Scale Image	ایستان Static
ب	
تکرارکننده Repeater	بازخورد Feedback
توده Burst	بازه زمانی Time Interval
ج	
جریان Flow	بخش‌بندی Segmentation
جلسه Session	بسته Packet
پ	
جویبار داده Data Stream	پایداری Robust
جویبار کنترلی Control Stream	پروتکل Protocol
چ	
چندرسانه‌ای Multimedia	پل Bridge
	پنجره Window

ح

ش

Network شبکه Privacy حریم خصوصی
Slow-Start شروع آهسته Attacker حمله‌گر

خ

Port Number شماره درگاه
Dynamic Port Number شماره درگاه پویا
Private Port Number شماره درگاه محرمانه
Traffic Identification شناسایی ترافیک
Server خدمت‌گزار

د

ص

Talkspurt صحبت جهشی
Port درگاه
Gateway دروازه
Handshaking دست‌دهی
Firewall دیوار آتش

ط

Classification طبقه‌بندی

ر

ق

Flow Conservation Law قانون پایستگی جریان
Behavior رفتار
Encryption رمزنگاری
Lawful Interception رهگیری قانونی

ک

Efficiency کارایی
Performance کارایی
User کاربر
InterDeparture Time زمان مابین خروج

س

Congestion Control کنترل ازدحام
Flow Control کنترل جریان

گی

Throughput گذردهی

ل

Tortoises لاک‌پشتی
Signaling سیگنال‌دهی

Peer همتا
 Transport Layer لایه ترابری
 Physical Layer لایه فیزیکی
 Application Layer لایه کاربرد
 Edge لبه شبکه

م

Window-Based مبتنی بر پنجره
 Random Variable متغیر تصادفی
 Adjacent مجاور
 Payload محتوا
 Router مسیریاب
 Client مشتری
 Destination مقصد
 Source منبع
 Host میزبان

ن

Rate نرخ
 Address نشانی
 Queuing Theory نظریه صف
 Noise نویز

و

World Wide Web وب جهان گستر
 Feature ویژگی

ه

Hub هاب
 Charging هزینه
 Correlation همبستگی

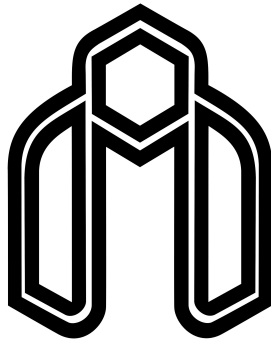
Abstract

There are huge petitions of network traffic coming from various applications in internet. In dealing with this volume of network traffic, network management plays a crucial rule. Traffic classification is a basic technique which is used by internet service providers to manage network resources and guarantee internet security. In addition, growing bandwidth usage, at one hand, and limited physical capacity of communication lines, at the other hand, lead providers to improve utilization quality of network resources. In fact, classification or identification of network is a critical task in network processing for traffic management, anomaly detection and also to improve network quality-of-service (QoS).

Port and payload based methods are two classical techniques which are applicable under traditional network conditions. However, many internet applications use dynamic port numbers for communications, which leads to difficulties in identifying traffic using port numbers. Also many applications encrypt the data before transmitting to avoid detection. Therefore, payloadbased techniques become ineffective for these traffics. In recent years, statistical feature-based traffic flow identification methods (STFIM) have attracted the attention of many researchers. The most important part of a STFIM is selection of efficient statistical features. Preliminary analysis shows that the changes of behaviors in different versions of an application and problem of packet loss in data transmission are the major challenges in employing STFIM for network traffic identification.

This affects the statistical characteristics of packets, such as the time interval between sending successive application packets, and in some cases significantly reduces the accuracy of traffic identification. The main objective of this paper is to examine the effects of packet loss on statistical features, and therefore the accuracy of identifying applications, as well as extracting appropriate features to overcome these effects. For this purpose, the behavior of four statistical features, including the packet size, the time interval between send and receive packets, the duration of the flows and the rate of sending packets, are investigated and by extracting the characteristics of their distribution, network traffic is identified. For this reason, a traffic database of seven applications with different packet loss rates has been analyzed and the accuracy of the identification of applications by the neural network has been analyzed. The results show that the extracted features are robust against the loss of packets and that it closely approaches the detection of network traffic in different modes of loss events depending on the ideal state (no loss of packet loss in the network).

Keywords: Network Traffic, Network traffic Identification, Machine Learning, Packet Loss, Behavioral Analysis, Data Mining



Shahrood University of Technology
Kharazmi International Campus

PhD Thesis

**Extraction of Statistical Features from Network
Traffic to Identify Applications on Routers**

By:

Mohammad Reza Gandomi

Supervisor:

Prof. Hamid Hassanpour

Advisor:

Associate Prof. Nematolah Shiri

May 2018