Effective and Real-time In-App Activity Analysis in Encrypted Internet Traffic Streams

Junming Liu

Yanjie Fu, Jingci Ming, Yong Ren

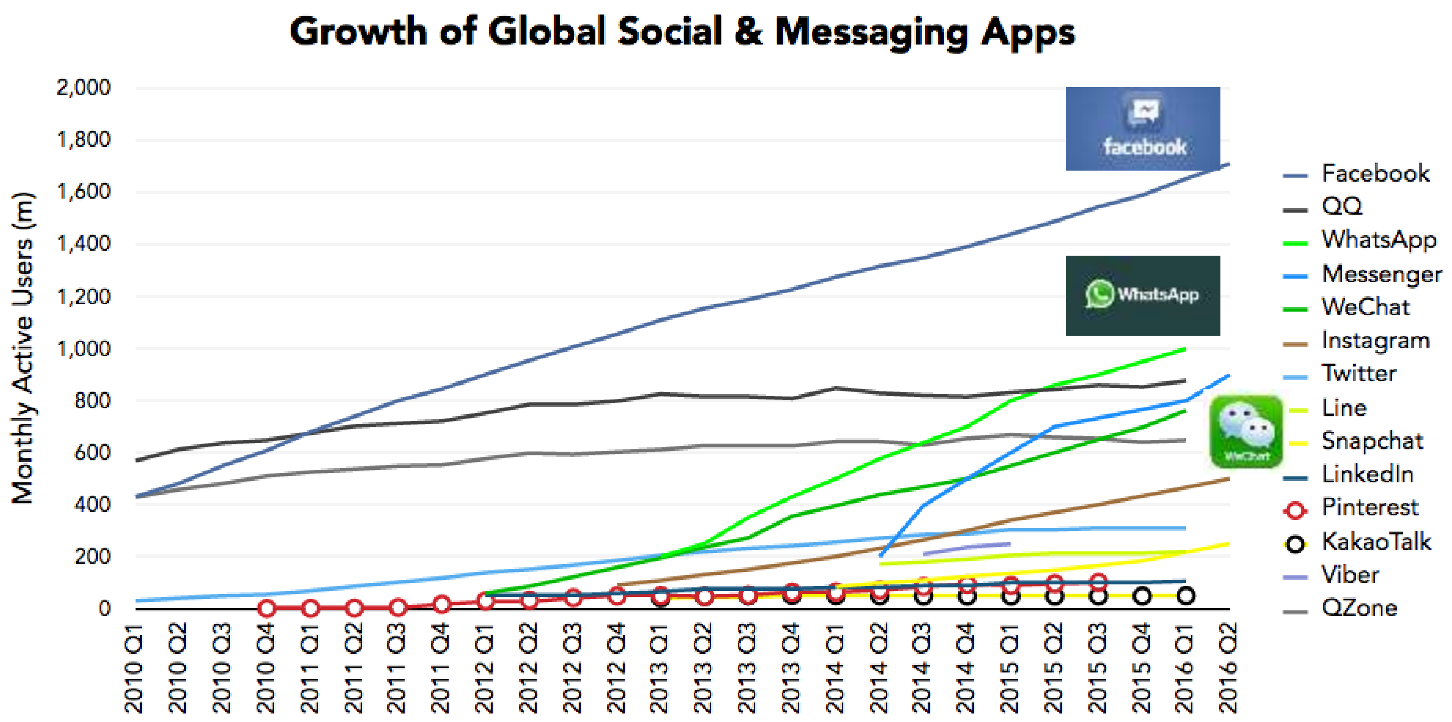
Leilei Sun, Hui Xiong

Rutgers University, USA

Futurewei Technology. Inc,. USA

Background

❑Explosive Growth in Mobile Apps



Source: Based on information contained in reports, press releases and other documents filed by these Social and Messaging App Companies with the U.S. Securities and Exchange Commission ("SEC") as well as materials disclosed on the websites of such Social and Messaging App Companies ("Reports"). ARK Investment Management LLC analyzed and internally ranked the social and messaging apps based on information in those Reports.

Ref: ARK INVEST. https://ark-invest.com/research/social-messaging-apps

Business in Mobile Apps

User's perspective:

- **Communicate** with each other in a social network, like multi-media messaging, moment post.
- **Engage** in commercial activities, like conference calls, paying bills, etc.

ISP's perspective:

- **Understand** users' preferences.
- **Provide** personized services or advertisements.
- **Improve** mobile users' satisfaction.

Challenges

THE STATE UNIVERSITY OF NEW JERSEY RUTGERS

- **Goal:** to discover mobile users' In-app activities
- **Problem:** Classify mobile Internet traffic into different usage categories in a **real-time manner**.
- **Challenges:**
  - **Encrypted Internet traffic** with very limited information from traffic packets (packet timestamp, packet length and packet protocol).
  - Need to handle **large traffic flows** from millions of users simultaneously as an **online analyzer.**

Preliminaries

Definition 1: Internet Traffic Flow

An internet traffic flow TF consists of a sequence of encrypted internet packets denoted by $\boldsymbol{TF} = \left\{(\boldsymbol{t_i}, \boldsymbol{P_i})_{\boldsymbol{i=1}}^{\boldsymbol{I}}\right\}$ where $\boldsymbol{I}$ is the total number of packets and $\boldsymbol{P_i}$ represents the packet received at time $\boldsymbol{t_i}$

Definition 2: Traffic Segment

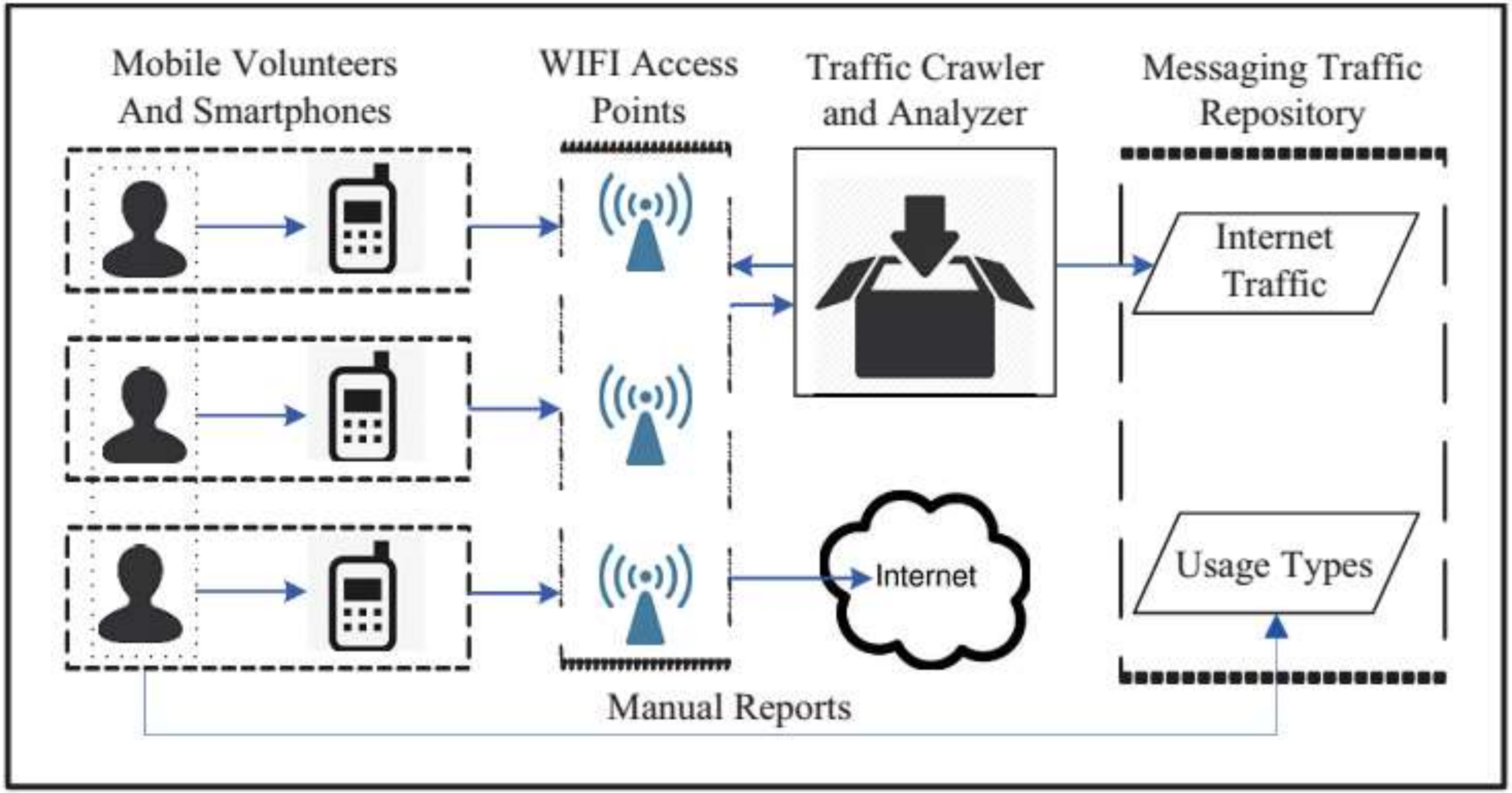
A traffic segment $S =< s_0, s_t >$ is a subsequence of an internet traffic flow from time s_0 to s_t.

Definition 3: Time Window Representation

A time window W_n records a small portion of traffic sequence starting from t_0^n to $t_{w_n}^n$. The size of a time window τ is fixed: $t_{w_n}^n - t_0^n \leq \tau$. There is a time gap Δ between adjacent time windows: $t_0^{n+1} - t_{w_n}^n \leq \Delta$.

Data Collection

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

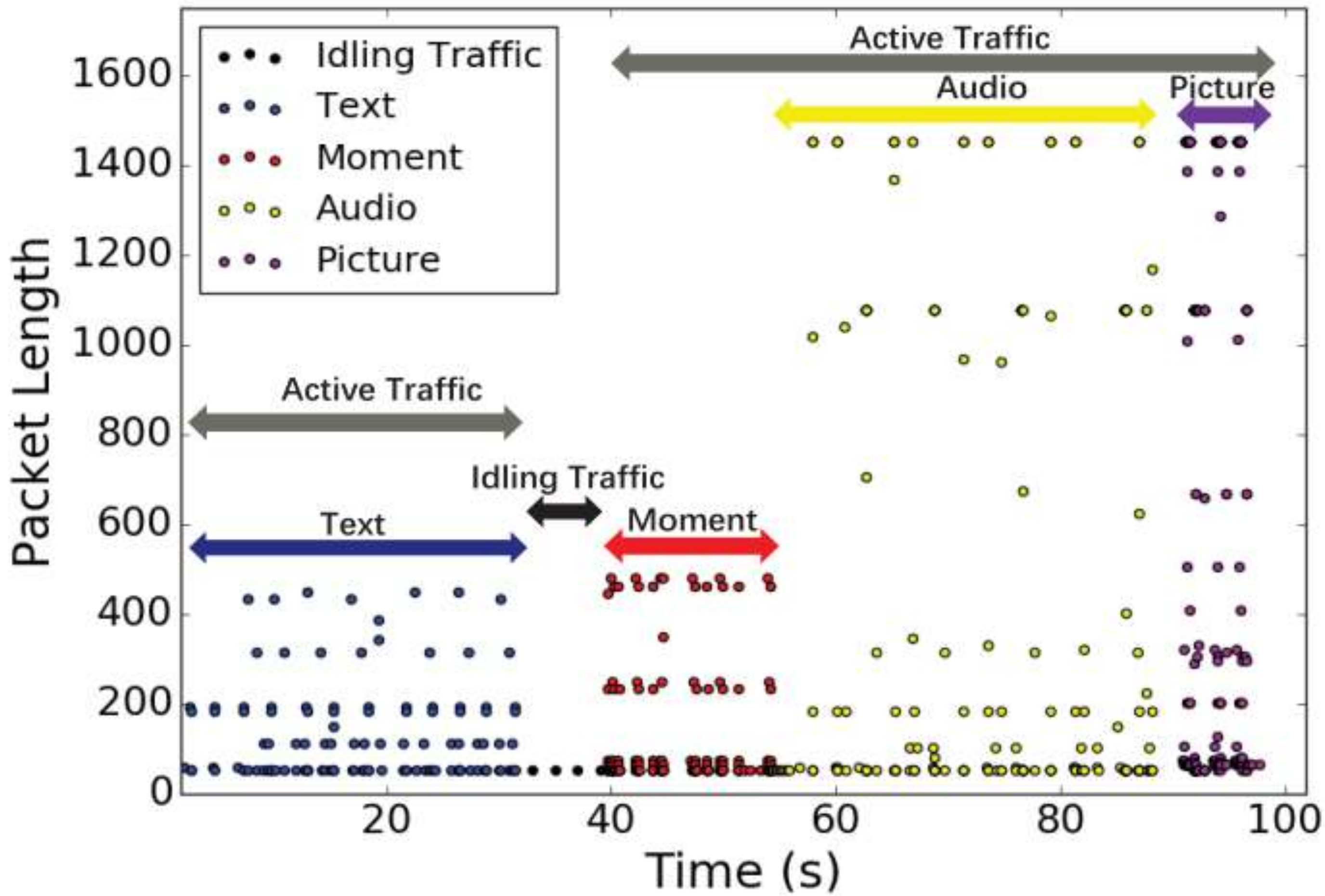


Data resources: daily usage of volunteers from Rutgers University and employees from major ISP

Traffic flow example

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Example of Collect Internet Traffic Flow



Problem Statement

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Given an incoming **traffic flow** $TF = \{(t_i, P_i)_{i=1}^{I}\}$, we need to classify **a sequence of in-App usage activities** denoted by $\{(b_n, e_n, u_n)\}_{n=1}^{N}$, where b_n, e_n, and u_n respectively represent the begin time, the end time, and the activity class.

1. Traffic flow segmentation
2. Traffic segment in-app usage classification

Table 1: Usage Activities of three Different Mobile Apps (Class Label)

U#	Wechat	Whatsapp	Facebook
0	Audio	Audio	Moment
1	Location	Picture	Video upload
2	Picture	Voice Call	Video watch
3	Short Video	Text	Picture
4	Video Call	Short Video	New Video Upload
5	Moment	Location	
6	Text		
7	Voice Call		

Framework Overview

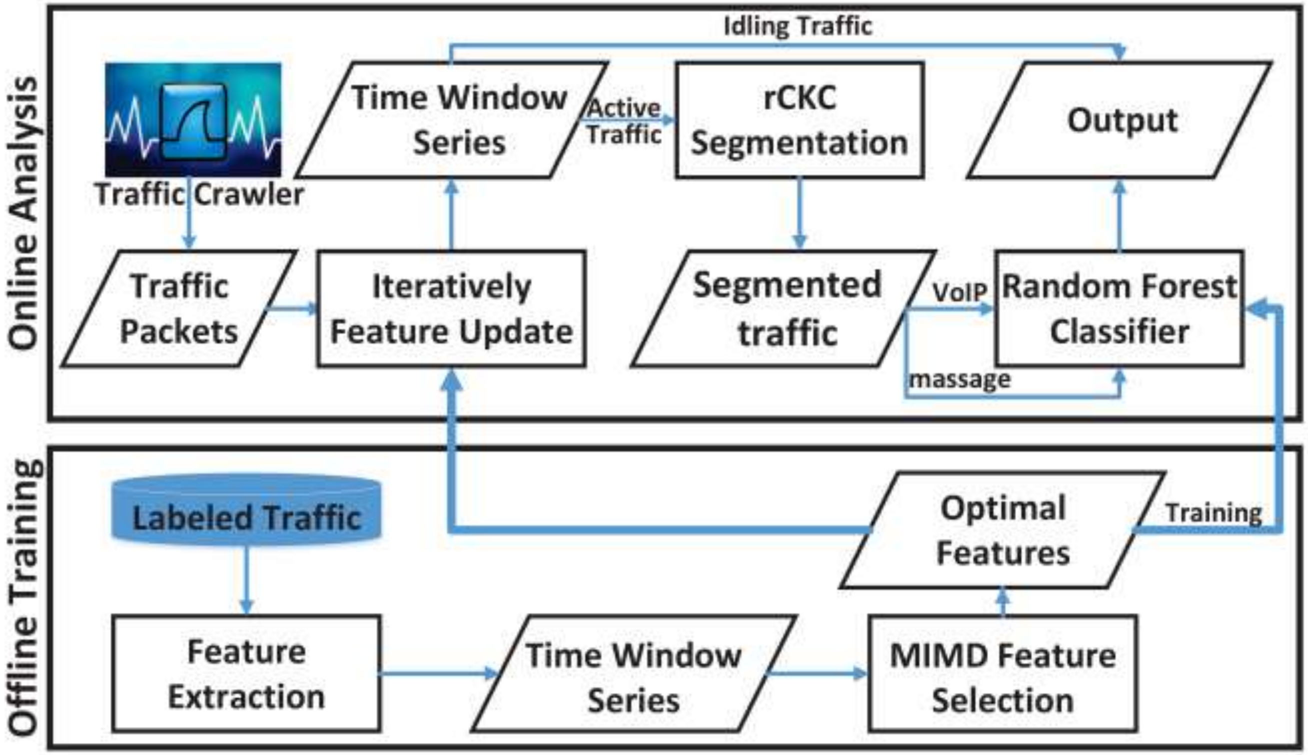


Figure 2: The Framework Overview.

Core algorithms

Offline Analysis: MIMD feature selection.

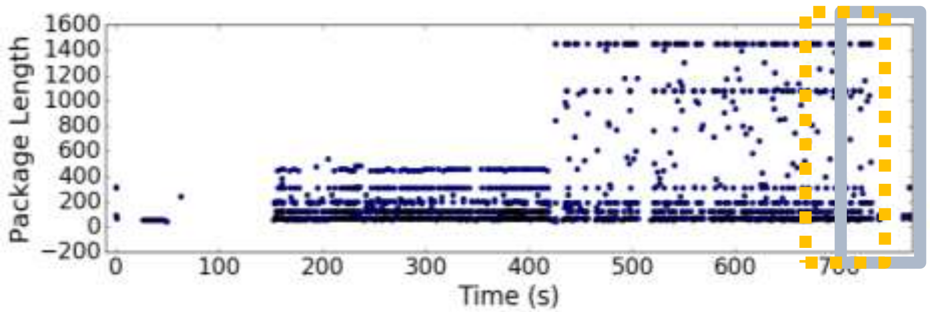
Online Analysis: rCKC traffic flow segmentation.

Framework Overview

Input: Raw traffic flow

Output: Activity class and its start-end time

1. Time window feature vector representation

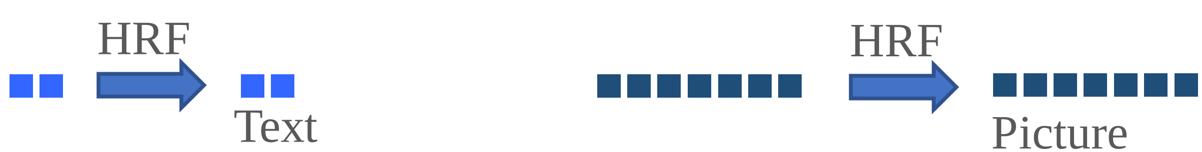


Time window sequence

■ : Feature of traffic window of feature vector

2. Recursive connectivity constrained clustering (rCKC) for segmentation

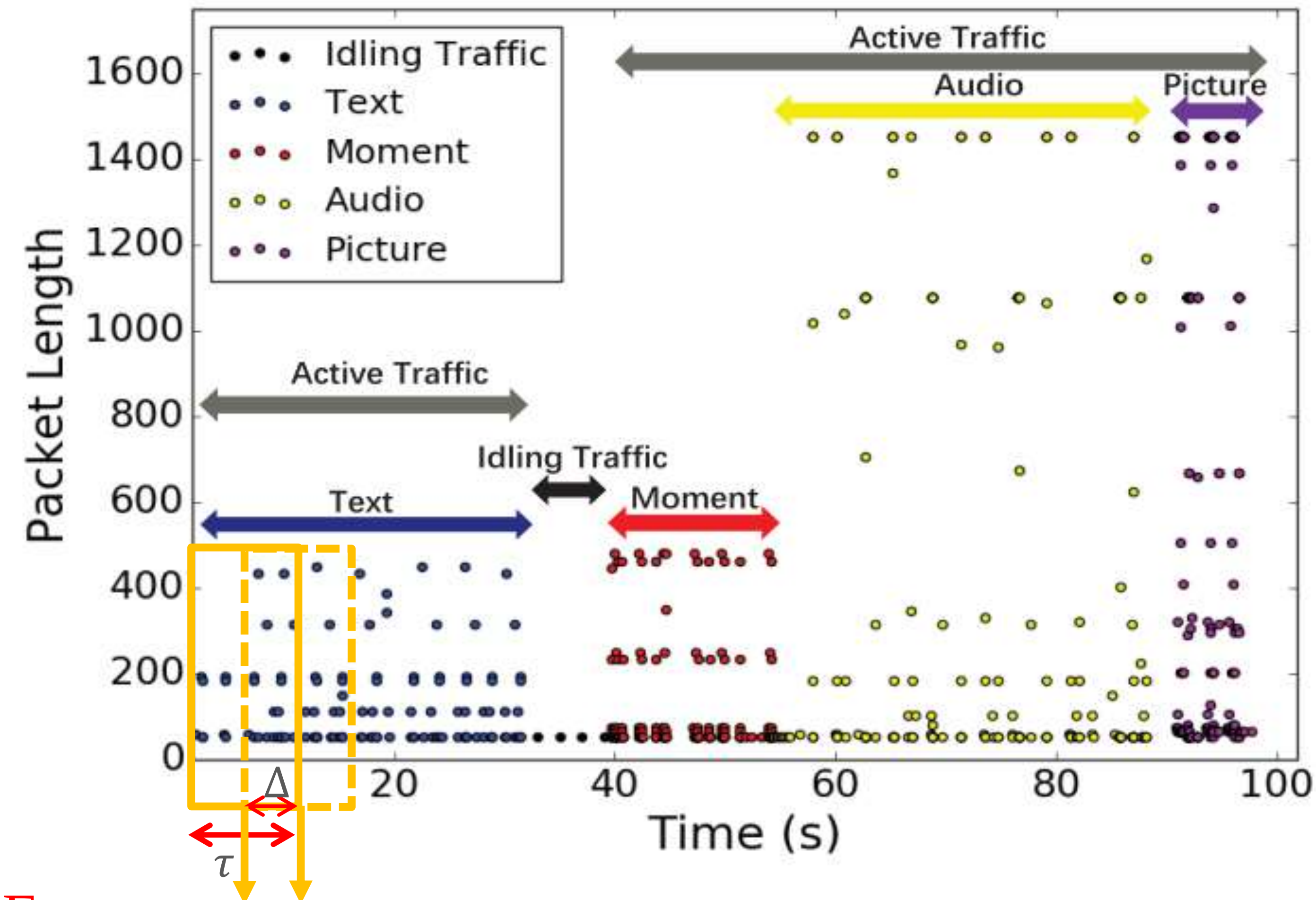
3. Segmented traffic usage activity classification



4. Output: labeled traffic

Offline Analysis

Time series feature extraction



Feature Vector $F_0\ F_1, \ldots$

Full feature set $\dim(V) = 30$

Offline Analysis

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Full feature set

- Packet length related features: basic statistics of packet lengths, hopping count, length of longest monotone subsequences, size percentiles, forward variances and backward variances.
- Packet time related features: basic statistics of adjacent packet time intervals, kurtosis, skewness.
- Traffic packet density (average number of packet second).
- Traffic speed (average packet size per second).

Advantages:

✓ High in-app usage activity classification accuracy.

Disadvantages:

- **Not completely independent** feature elements.
- **High latency** due to complex feature extraction.
- **Large memory** requirement for high dimension feature vectors.
- **Low impact** on segmentation.

Offline Analysis

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

***M**aximizing **I**nner activity similarity and **M**inimizing **D**ifferent activity similarity measurement (**MIMD** feature selection).

- Similarity of normalized feature vector of dimension N (Gaussian kernel)

$$SD(\mathbf{F}, \mathbf{F'}) = \frac{1}{N}\sum_{n=1}^{N} e^{-(F_n - F'_n)^2}$$

- Maximizing Inner activity similarity

$$\max IAS(A; \mathbf{F}), \quad IAS(a_i; \mathbf{F}) = \frac{1}{n_i^a}\sum_{k=1}^{n_i^a} SD(\mathbf{F}_{i,k}^a, \bar{\mathbf{F}}_i^a)$$

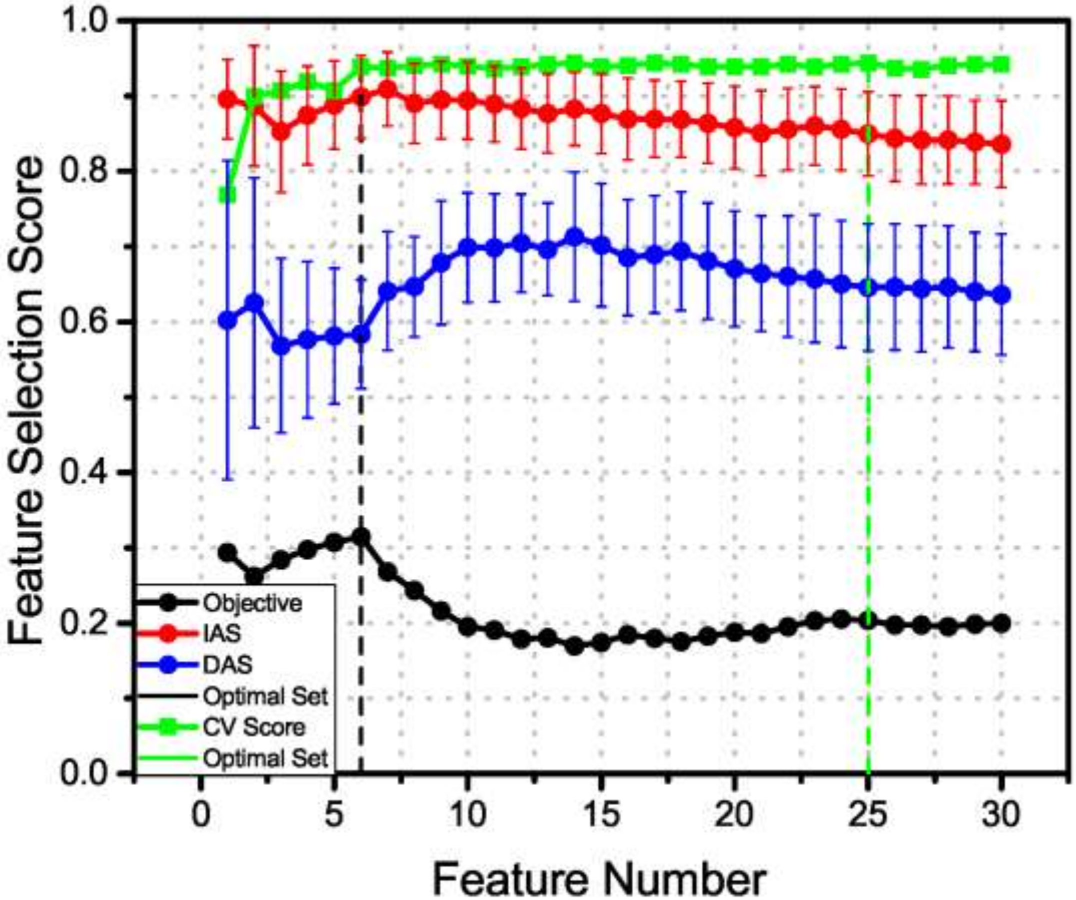
- Minimizing Different activity similarity

$$\min DAS(A \neq A'; \mathbf{F}), \quad DAS(a_i, a'_j; \mathbf{F}) = SD(\bar{\mathbf{F}}_i^a, \bar{\mathbf{F}}_j^{a'})$$

- **MIMD Objective:**

$$\max \Phi(IAS, DAS)), \quad \Phi(IAS, DAS) = IAS - DAS$$

Offline Analysis



MIMD feature selection:

- Recursive feature addition
- A high dimension feature provide high CV accuracy but low MIMD score.
- Dimension of optimal feature set from MIMD measurement is 6.
- Optimal feature set keeps a high CV accuracy (0.55% lower than the highest value at dimension 25).

Offline Analysis

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Optimal feature set

Given a time window of *N* packets observation: $\{(\boldsymbol{t_1}, \boldsymbol{P_1}), \dots, (\boldsymbol{t_N}, \boldsymbol{P_N})\}$

- **Percentile 25%:** percentage of packets with length smaller than 25% maximum packet length L_{max}: $P_{25} = \frac{1}{N} \sum_{i=1}^{N} \delta(P_i.l < 25\% \, L_{max})$.
- **Percentile 75%:** percentage of packets with length greater than 75% maximum packet length L_{max}: $P_{75} = \frac{1}{N} \sum_{i=1}^{N} \delta(P_i.l > 75\% \, L_{max})$.
- **Top frequent continuous subsequence TCS:** the highest repeating frequency of packet subsequence of length 3.
- **Packet length variance var**: $var = \frac{1}{N} (\sum_{i=1}^{N} P_i.l^2) - (\frac{1}{N} \sum_{i=1}^{N} P_i.l)^2$
- **Traffic density:** number of packets per second: $TD = \frac{N}{t_N - t_1}$
- **Traffic speed:** average packet lengths per second: $TD = \frac{\sum_{i=1}^{N} P_i.l}{t_N - t_1}$

Traffic Flow Segmentation

Traffic flow segmentation algorithm (*rCKC*)

Recursive Connectivity Constrained KMeans Clustering

Challenges:

- Time series segmentation problem-time continuity constraint
- Optimal number of single activity segment is unknown (undecided K)

Algorithm 1 $rCKC(S = \{w_i, i = 1, 2, ..., N\}, K)$

Require: Input: $\{w_i(F_i; t_i), i = 1, 2, ..., N\}$

```
if IAS(S) > δ then
    output S = {w_i, i = 1, 2, ..., N; F(S)}
else
    Initial: C^0 = arg max_{c_j ∈ W} Σ_{j=1}^{K} DAS(c_j, c_{j+1})
    while C^p ≠ C^{p+1} do
        p → p + 1 %next iteration
        b^p = arg max_{b^p} IAS(S(w_{b^p} : w_N));
        C_1^p = (1/b^p) Σ_{i=1}^{b^p} (w_i)
    end while
    for j = 1 : K do
        rCKC(S_j, K)
    end for
end if
```

Objective:

Group a sequence of time windows $\{w_i\}_{i=1}^{N}$ into single-activity segments

Recursive strategy:

1. Check input segment IAS→split input segment or output as single-activity segment for in-app usage activity classification.
2. Initial K segments by maximizing the adjacent segment DAS.
3. Iteratively optimize $K - 1$ split point as sub-segment boundaries.
4. Each split sub-segment is fed into rCKC.

Online Implementation

Iterative feature vector update

Challenges:

- No enough cache space for large traffic flow from millions of users
- Fast packet processing with small and stable cache storage

Algorithm 2 Iteratively update feature and time window

Require: Two sets of temporary variable (initial 0):
tem=$(N, N_{25}, N_T, N_U, L, L^2, TCS)$
tem'=$(N', N'_{25}, N'_T, N'_U, L', L^2, TCS')$

while Receive packet P **do**
  if $P.t - T_0 \leq \tau$ **then**
    $Update(tem, P)$, $T_t = P.t$
    if $P.t - T_0 \geq \Delta$ **then**
      $Update(tem', P)$
    end if
  else
    $N_{25} = \frac{N_{25}}{N}$, $N_{75} = \frac{N_{75}}{N}$, $var = \frac{L^2}{N} - \mu^2$
    $TCS = \max_{L \in FCS} FCS(L)$
    $TS = \frac{L}{T.t - T_0}$, $TD = \frac{N}{T.t - T_0}$, $R_{pr} = \frac{N_U}{N_T}$
    Store feature: $N_{25}, N_{75}, var, TS, TD, R_{pr}$
    $tem = tem'$, $tem' = \mathbf{0}$, $Update(tem, P)$, T_0=$P.t$
  end if
end while

Objective:

Construct time window feature vectors online without the storage of raw packets.

Iterative strategy:

1. For each incoming Internet packet extract packet information $(t, P.l, P.\mathrm{Pr})$, update two sets of temporary variables tem, tem'.
2. tem variable is used for current time window feature vector construction and tem' for next time window.
3. The packet is released after tem, tem' update.

Experimental Data

Table 2: Statistics of the WeChat Training data.

#	Usage Type	Records	Packets	Traffic	Tra/min
1	audio	136	44K	23.53M	208K
2	location	112	119K	31.79M	348K
3	picture	100	132K	103.2M	986K
4	sight	63	163K	141.11M	1.33M
5	video call	100	1,170K	239.76M	2.17M
6	moment	67	7K	1.18M	50K
7	text	229	30K	4.5M	32K
8	voice call	105	265K	32.54	758K

Table 3: Statistics of the Whatsapp Training data.

#	Usage Type	Records	Packets	Traffic	Tra/min
1	audio	176	72K	26.62M	436K
2	picture	197	178K	141.5M	2.26M
3	call	194	143K	21.64M	287K
4	text	202	34K	3.22M	42K
5	video	173	483K	472M	11.06M
6	location	80	11.52K	8.03M	47.77K

Table 4: Statistics of the Facebook Training data.

#	Usage Type	Records	Packets	Traffic	Tra/min
1	moment	101	40K	21.65M	607K
2	videoup	75	21K	6.56M	238K
3	video watch	108	1,216K	1,326M	42M
4	picture	97	57K	51M	2.26M
5	new video	77	844K	825M	10M

Table 2, 3, 4 show the basic statistics of our collected single activity traffic data.

In addition, we collect two-activity traffic data with the time duration of each segment ranging from 5s to 120s.

Experiment

Study of Traffic Flow Classifier

Proposed Classifier:
Random Forest with VoIP-noVoIP traffic filtering. (HRF)

Baselines:
Random Forest; Support Vector Classifier; K-Nearest Neighbors Classifier; Gaussian Naïve Bayesian Classifier.

Evaluation Metrics:
Overall accuracy, Precision, Recall, F-Measure.

Experiment

Study of Traffic Flow Analyzer

Proposed Analyzer:

rCKC traffic flow segmentation + **HRF** segmented traffic classifier

Baselines:

AC + **RF:** Agglomerative Connectivity Constrained Clustering + RF

CUMMA: Adjacent packet merging strategy + RF

SW+RF: Sliding window based segmentation + RF.

Evaluation Metrics:

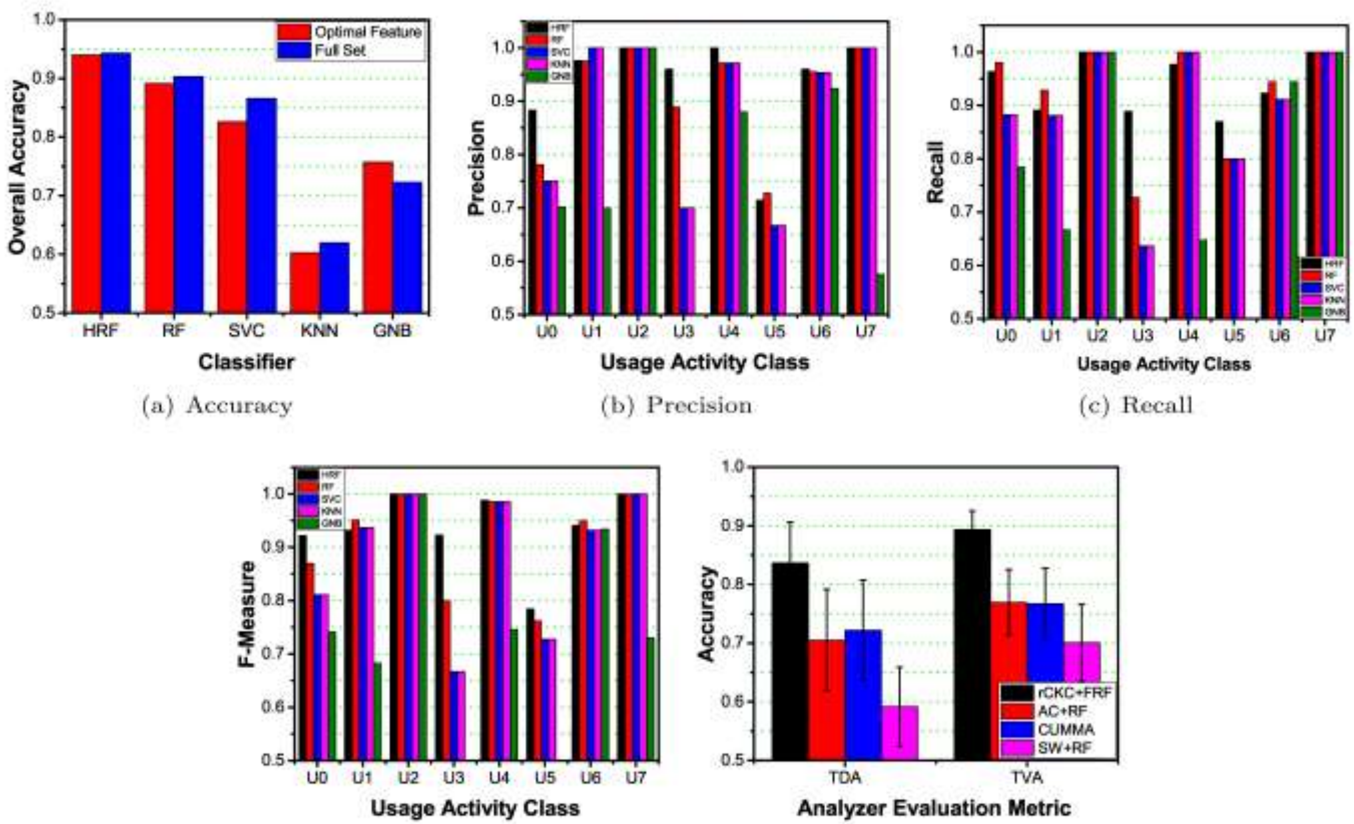
TDA: traffic duration accuracy.

TVA: traffic volume accuracy.

$$TDA = \frac{1}{T(TF)} \sum_{S} \sum_{\hat{S}} \delta(a_s - \hat{a}_s) T(S \cap \hat{S})$$

$$TVA = \frac{1}{V(TF)} \sum_{S} \sum_{\hat{S}} \delta(a_s - \hat{a}_s) V(S \cap \hat{S})$$

Wechat Performance Comparison



(a) Accuracy (b) Precision (c) Recall

(d) F-Measure (e) TDA-TVA

Figure 4: Performance Comparison of Wechat

Experimental Result

Whatsapp Performance Comparison

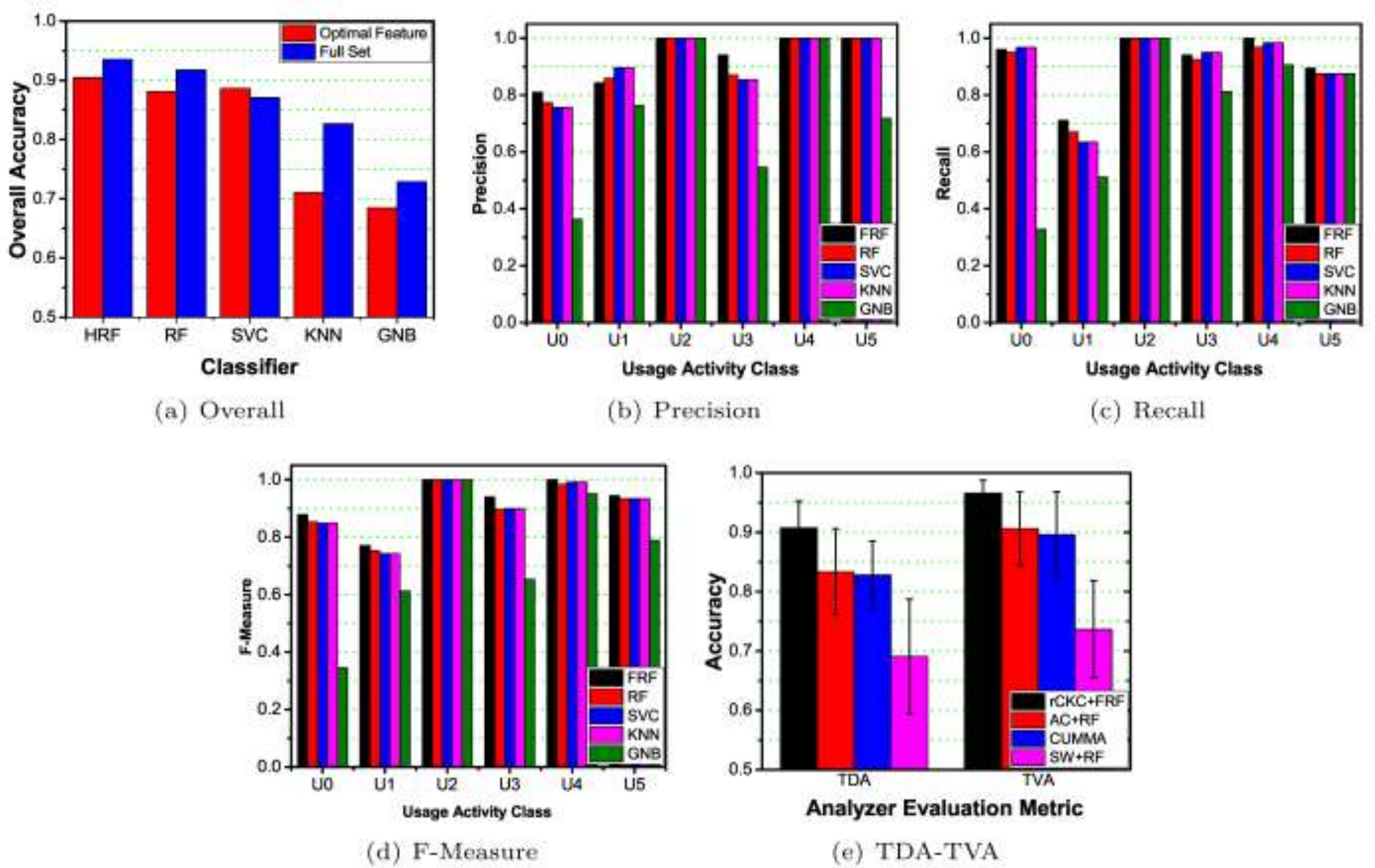


Figure 6: Performance Comparison of Whatsapp

Experimental Result

Facebook Performance Comparison

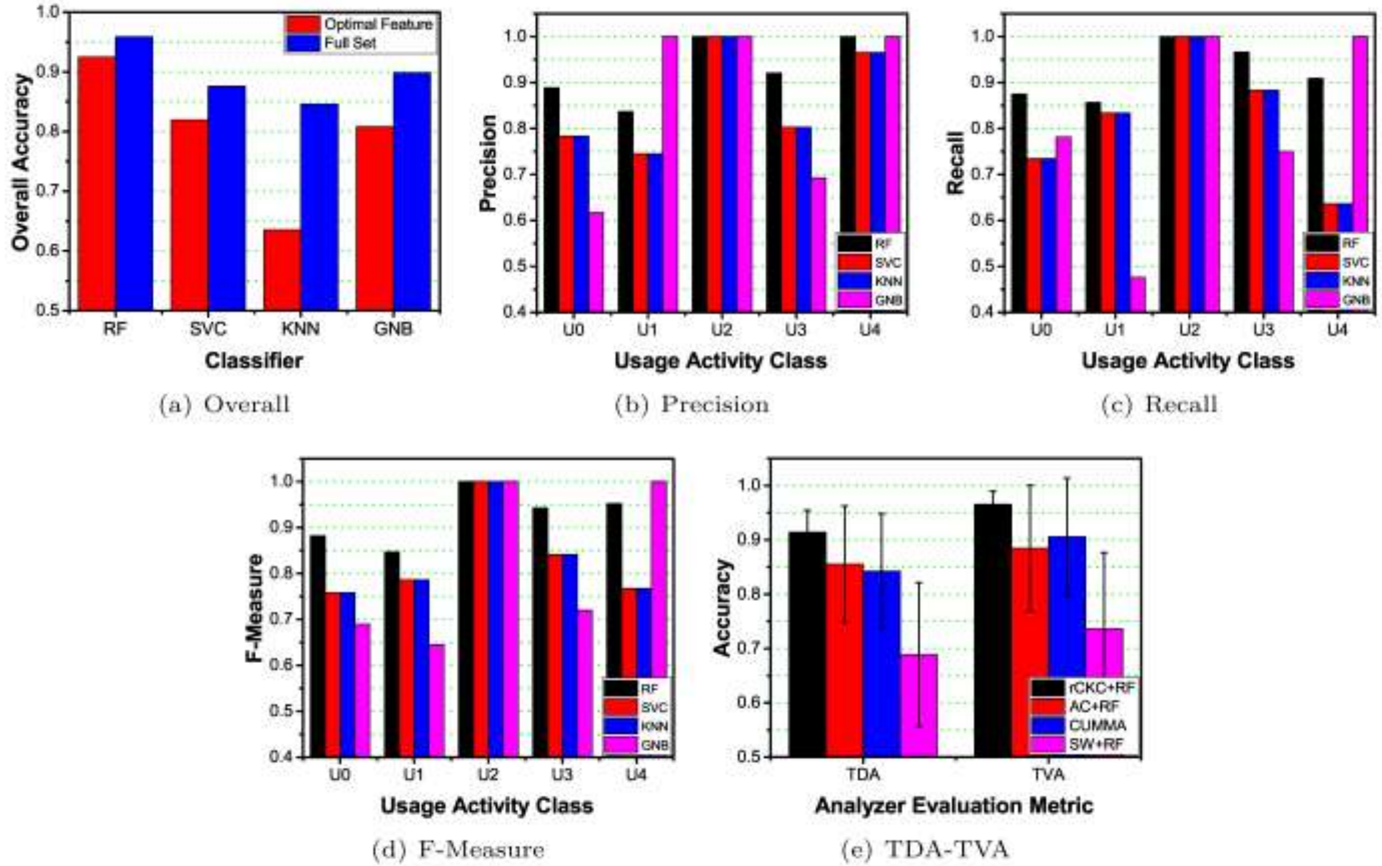
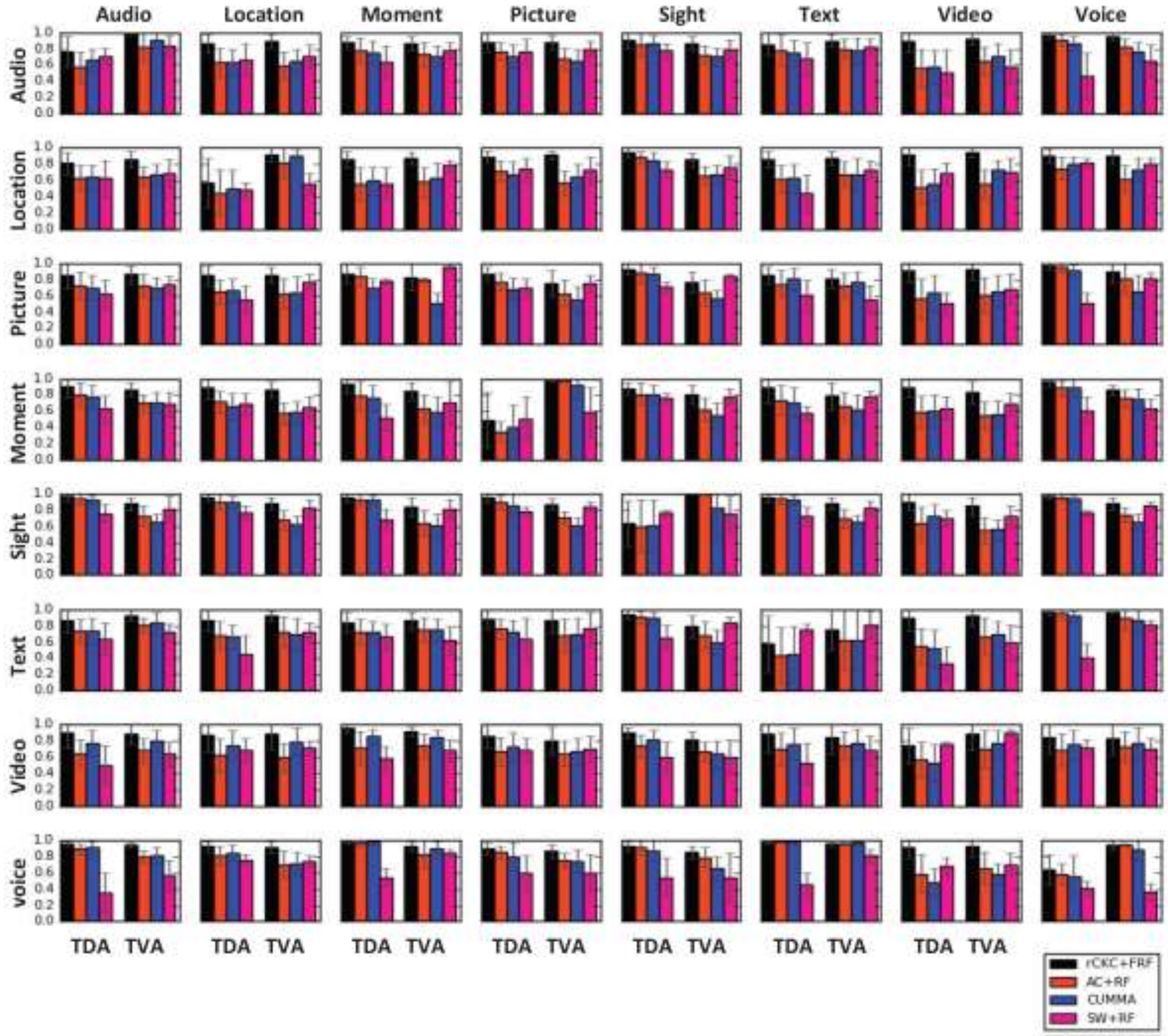


Figure 7: Performance Comparison of Facebook

Experimental Result

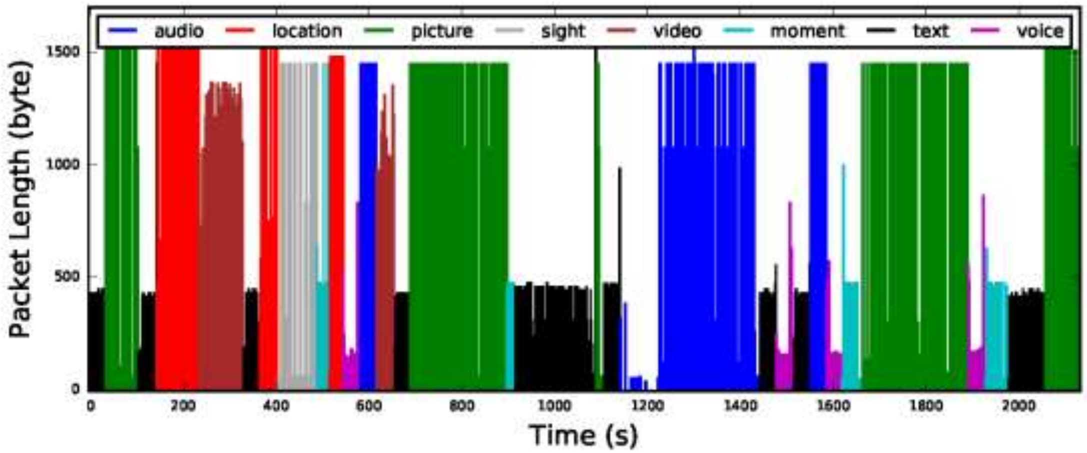
Wechat Two-activity Test



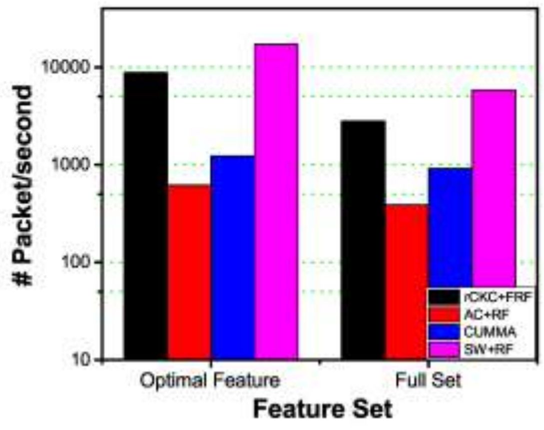
Experimental Result

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

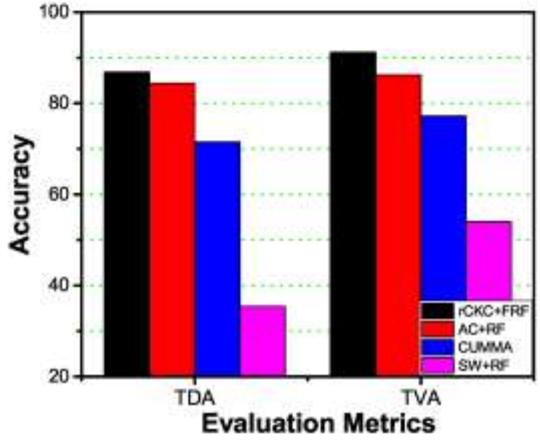
Online test



(a) Ground Truth Traffic Flow



(b) Efficiency



(c) Accuracy

Conclusion

An **online mobile app traffic analyzer** for classifying encrypted mobile app Internet traffic into different types of service usages.

- **MIMD** Internet packet time series feature selection criteria.
- **rCKC** Internet packet time series segmentation algorithm.
- **VoIP-noVoIP filtered RF classifier** for segmented traffic.
- **Online iterative feature vector update** strategy.
- Real world mobile Internet traffic of most popular Apps: **Wechat, Whatsapp and Facebook**