Streaming Media Traffic Characterizations Analysis in Mobile Internet

Hongyan Cui[1,2], Jia Wang[1,2], Fangfang Sun[1,2], Yunjie Liu[1,2]Kwang-Cheng Chen[3]
1.State Key Lab. of Networking and Switching Technology, 2.Key Lab. of network system architecture and convergence
Beijing University of Posts and Telecommunications, Email: cuihy@bupt.edu.cn
3.Department of Electrical Engineering National Taiwan University, Email:ckc@ntu.edu.cn

***Abstract*—In spite of great efforts to expand the network infrastructure, it is still hard to catch the rapid growing need to serve tremendous subscribers with significant traffic. To effectively allocate the resource among mobile telecommunication operators and Internet Service Providers (ISP), grasping the highly dynamic traffic patterns can enable effective network planning and optimization for state-of-the-art mobile Internet. We employ traffic behavior analysis method to analyze and model the Internet traffic. Due to the subscribers' number of streaming media accounting over 60% of the entire traffic, we conduct a complete study to analyze this kind of traffic. By contrast, we also give the analysis result of another application, Instant Message (IM). Applying our newly proposed clustering methodology, we describe the regional characteristic of applications. This research subsequently enables more subtle and efficient mobile Internet operating modes to benefit operators and ISPs.**

I. Introduction

The rapid system expansion, planning and optimization, green communication and other aspects of mobile Internet are facing tremendous challenges. Strategy Analytics [1] points out that Global wireless networks traffic will rise from current 5 EB to 21 EB in 2017. A few low-profit applications occupy a good portion of bandwidth resources, that leads to operators facing an operating dilemma with the network expansion.

As a matter of fact, a good portion of traffic commonly follows some regular rules. If these rules can be used to improve resource scheduling, the problem of unbalanced traffic and congestion areas caused by the inefficient, random, repeated traffic transmission could be resolved to a great extent. Recently researchers pay much attention to multi-scale features in mobile Internet [2]. [3] proposes a hybrid circle reservoir (HCR) ESN architecture applying to traffic prediction. [4] predicts users' browsing behavior by Markov chain model. [5] presents the flexible functionality of a user behavior based traffic emulation system which is capable of working on different platforms (Windows, Android), on different access technologies (wired, WiFi, 3G). [6] studies the communication network modeling and analysis of the problem in time and space with a burst of traffic by Markov chain. [7] proposes GPRS traffic analysis system based on cloud computing, and verifies users online time and other characteristics obey the power law distribution. [8] presents a traffic characterization called Digital Signature of Network Segment Using Flow Analysis (DSNSF) which can follow the trends and make predictions of network traffic effectively. [9] proposes a new hybrid video traffic model for MPEG-4, which produces results of high accuracy. In [10], the authors characterize subscriber mobility and temporal activty patterns and identify their relation to traffic volume. Meanwhile, they investigate how efficiently radio resources are used by different subscribers as well as by different applications. [11] provides the first fine-grained characterization of the geospatial dynamics of application usage in a 3G cellular data network.

However, with the new applications' springing up and the change of surfer behaviors, the existing researches can not describe the resent mobile traffic characteristic precisely. So we investigate the new rules and explore a set of traffic characteristics analysis methods based on the real data. Cisco points out that, the proportion of video traffic of the global data will surpass 66% [12]. Thus, among multitudinous types of traffic, we choose video streaming as the targets to analyze. The analytical result about traffic distribution, flow direction and regional traffic composition aims to prove useful information to operators and ISPs on network resource scheduling and planning.

The major technical contributions of this research include:

1) We analyze the main flow directions of video streaming in the city, and find the following special rules: the flow direction of video streaming flow direction is especially dispersed.
2) We apply our novel adaptive clustering algorithm to cluster all the SGSNs (Serving GPRS Support Nodes) in the city and portray the application characteristics of different geographic areas, such as volume of different applications,application composition in each cluster, the number of subscribers and traffic records, and so on.

Our work can help operators and ISPs pertinently carrying some centralized planning and equipment management, optimize load balancing, as well as accurate and efficient resource scheduling.

II. Analysis of the main flow Direction

A. *Data Set Description*

The dataset of one city consists of 50,340,197 records of 19th Nov,2012, and the connecting subscribers is 3,270,860 which occupying the 11.1% of the city's population 29,450,000. The data has been collected by a telecom service provider. Each record consists of the users ID, the traffic type

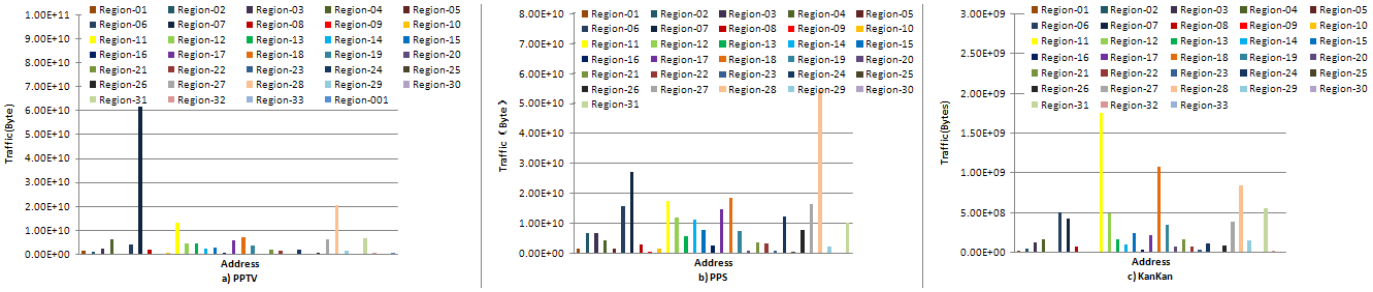

Fig. 1. Main Flow Direction of Video Streaming. Different color in figure indicates different target place, which is accurate to one city.

history, the web browsing history, the online duration, the date and time of the using history, as well as the phone numbers, and their LAC, CellID.

B. *The Flow Analysis of Streaming Media*

We analyze the main direction of streaming media flows. The *main flow direction* is signified by the geographical of the server corresponding with the flow's destination IP. PPTV, PPS and KanKan three applications in streaming media of their respective categories by a flow proportion share are shown in Table I:

From Table I we can see that the selected applications are the main parts of streaming media flows, occupying the 73.32% of the whole traffic volume, so we choose them as our analysis objects.

TABLE I
TRAFFIC PROPORTION OF MAIN APPLICATION IN VIDEO STREAMING

Video streaming		
PPTV	PPS	KanKan
27.07%	44.90%	1.35%

Here, we investigate three main streaming media applications in a city. As the flow directions of different days are similar, we just select one day to analyze. Figure 1 shows the flow directions of PPTV, PPS, KanKan, which are three main applications of video streaming. In addition to the highest traffic distributing in a few areas, the flow traffic of other regions is also great. The (highest traffic) here means the volume of the traffic to the direction whose traffic higher than other directions. Figure 2a) presents the PPTV flow conditions, in which the flow traffic to Region-07 is highest, 6.31E+10Bytes, accounting for 37.56% of the total flow traffic of the day. The flow traffic to Region-28 is 2.04E+10Bytes, accounting for 12.14% of the total flow traffic of the day, and the sum traffic of the other 32 regions accounts for 50.3% of total PPTV traffic that day. In Figure 2b), the PPS traffic flowing Region-28 is highest, 5.47E+10 Bytes, accounting for 19.89% of the total PPS flow of the day, and sum traffic in other 30 regions accounts 70.11% of the total PPS traffic flow. Figure 2c) shows that the KanKan traffic flowing to Region-11 is highest and the value is 1.84E+10Bytes, accounting for 22.01% of the total KanKan flow traffic. The sum of the other 32 regions accounts for 77.99% whole flow. Generally speaking, for video streaming, although the traffic flowing to some regions is significantly higher than other regions, but the traffic to other areas is still not negligible. The proportion of the highest traffic is 91.39% for PPTV, and 19.31% for PPS. The main reason is due to the higher traffic generated by video streaming service. In order to meet users' demand, a lot of ISPs deploy many servers working together to provide service.

C. *The Flow Analysis of IM*

As a comparison scheme, the analysis result of IM is show in Figure 2. The mean applications of IM are QQ, MSN and WeChat. By the analysis of flow direction, our statistical results are consistent with the features of IM. Users primarily send short text, when sending a message to the server. Even for the newly arisen IM application WeChat, although voice communication is emphasized, which in principle is sending recording to servers and the communication partner downloading from the server to listen. Due to the geographical distribution of the company's server is more converging, thus the most traffic of IM flows a handful of areas. The proportion of the highest traffic is over 99.99% for WeChat, and 67.64% for QQ.

Generally speaking, the flow directions' characteristics of video streaming and IM have a clear distinction. For video streaming, although the traffic flowing to some regions is significantly higher than other regions, but the traffic to other areas is still not ignored. The proportion of the highest traffic is 91.39% for PPTV, and 19.31% for PPS. The main reason is due to the higher traffic generated by video streaming service. In order to meet users' demand, a lot of ISPs deploy many servers working together to provide service.

III. ANALYSIS OF THE APPLICATION ON SGSN

A. *Overall characteristics of each SGSN cluster*

We analyze the generated Internet records, total 1.07E+12Bytes by 3,270,860 users connected to 921 SGSNs of one day. We try to probe the characteristics of the application on SGSN, which represents the regional characteristic to a certain extent. In order to describe the characteristics across-the-board, we introduce another kind of application IM (instant messaging) except for streaming media. Further more, we choose the typical applications QQ, MSN and Wechat as research objects.

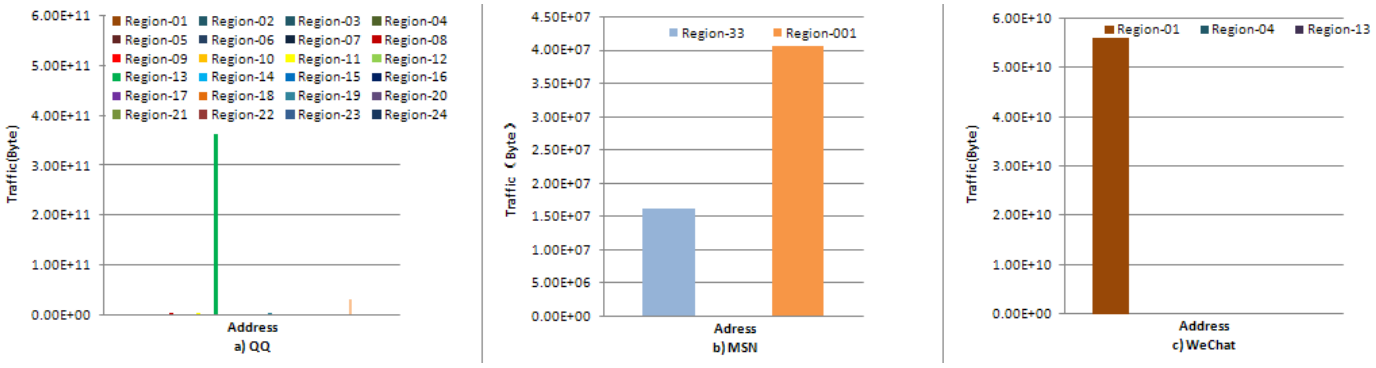

Fig. 2. Main Flow Direction of IM. Different color in figure indicates different target place, which is accurate to one city. The y axis is rendering logarithmic coordinate.

TABLE II
CHARACTERISTIC OF SGSN CLUSTERS

SGSN cluster	Cluster	Cluster II	Cluster III	Cluster IV	Cluster V	Cluster VI	Cluster VII	Cluster VIII
SGSN_Num	4	27	833	5	10	11	28	3
User_Num	4.92E+03	7.02E+04	1.36E+05	5.75E+03	7.36E+05	1.20E+06	1.11E+06	2.80E+03
Record_Num	6.00E+04	6.00E+04	2.78E+06	1.06E+05	9.43E+06	1.87E+07	1.73E+07	9.12E+04
Characteristic	higher traffic of cluster-center , most SGSN_IP adjacent	SGSN_IP rather dispersed	relatively lower traffic, less users, SGSN_IP rater dispersed	most SGSN_IP adjacent	huge traffic of cluster-center, large number of users, SGSN_IP continuous	huge traffic of cluster-center, large number of users, SGSN_IP continuous	more users, SGSN_IP address is divided into several contiguous IP addresses parts	SGSN_IP rather dispersed
Mean traffic of cluster-center	2.01E+09	2.99E+08	1.10E+07	7.59E+08	4.06E+09	7.93E+09	7.45E+08	1.57E+08
Total traffic of cluster-center	1.21E+10	1.80E+09	6.62E+07	4.55E+09	2.44E+10	4.76E+10	4.47E+09	9.43E+08
Use_num of cluster-center	1.23E+03	2.60E+03	1.63E+02	1.15E+03	7.36E+04	1.09E+05	3.98E+04	9.33E+02
Record_num of cluster-center	1.50E+04	6.89E+04	3.34E+03	2.12E+04	9.43E+05	1.70E+06	6.18E+05	3.04E+04

First, we classify all SGSNs to different clusters based on the characteristics of the traffic on the SGSNs. Due to the difficulty to determine in advance the number of all SGSNs clusters connected to by the users on this day, so in the paper, we use an adaptive clustering method proposed ourselves for clustering, which can adaptively determine the number of clusters according to the characteristics of different SGSNs. Specific process is as follows. We take each SGSN as a cluster object, the six applications traffic (QQ, MSN, WeChat, PPTV, PPS, KanKan), the number of records and the user access, a total of eight characteristics as the feature items of every SGSN, and the weights of each item are the same.

Our adaptive clustering is composed of two stages, first stage to determine the number of clusters, and then using the k-means clustering algorithm for clustering. We define a cluster validity index FPCVI to determine the number of clusters c [13]. The procedure of selecting the optimal value of c using FPCVI is as follows.

Algorithm 1 The calculating steps of FPCVI

1: For each value of c = 2, 3,... , c_{upper}, we carry out a clustering algorithm and compute F_c(c = 2, 3, ... , c_{upper});
2: Compute to get matrix Q;
3: Compute FPCVI(c) (c = c_{min}, ... , c_{max});
4: Compare FPCVI(c_{min}), ... , FPCVI(c_{max}) and then we get the final.

F_c is defined as the pattern matrix which indicates the degree that any two objects belong to a same cluster or different clusters, and Q is the final global pattern matrix which indicates the probability of belonging to a same cluster (or different clusters) for each pair of objects in data set. Applying our feature matrix to the formulation, calculate by the number of the cluster is eight, that is all the SGSNs of this day will be classified into eight cluster after the k-means clustering algorithm.

The input data D of k-means is 921*8 feature matrix, in which the row is on half of different SGSN and the colume stands for the feature of each SGSN. After adaptive clustering, we finally obtain 8 SGSN clusters, and the characteristics of each cluster are shown in Table II.

As the characteristics of each SGSN cluster have been listed in the above table, here we only emphasize three of the clusters, which have more outstanding characteristics.

First is the Cluster III. We find the number of SGSNs in this cluster is the biggest of all, which reaches 833 while the number of other cluster is less than 30. The concrete features are as follow:

(a) The traffic flow through the cluster-center is significantly less than the value of other cluster-centers. The magnitude of the Cluster III's mean traffic is at 10^7, while others' are at 10^8 or 10^9, and the magnitude of the Cluster III's total traffic is at 10^7, while others' are at 10^9 or 10^{10}.

(b) Observing the number of users and the records we can find that for Cluster III, the two value have 1 or 2 lower magnitude than other cluster-centers.

(c) In addition, Viewing SGSN IP of the 833 SGSNs in Cluster III, we can contain that the geographic distribution is more dispersed. The traffic flow through the SGSNs in Cluster III is relatively lower, as well as the number of the users and the records.

The other two clusters that attract our attention are Cluster V and Cluster VI. The numbers of SGSNs in the two clusters are equivalent, 10 and 11 respectively. Their common characteristics are as follow, and the distinction will be illustrated in the following.

(a) The mean traffic and the total traffic of Cluster V and Cluster VI's cluster-centers are the highest in all the clusters. The magnitude of the mean traffic achieves 10^9 and the magnitude of the total traffic reaches 10^{10};

(b) The number of users and records in Cluster V and Cluster VI are also larger than any other clusters. The number of users in Cluster V is 7.36E+04, and the number of records is 9.43E+05. The number of users in Cluster VI is 1.09E+05, and the number of records is 1.70E+06;

(c) By checking the SGSN IP addresses, we find the IP addresses are adjacent in Cluster V and Cluster VI. Particularly for the Cluster VI, nine of eleven SGSN IP addresses are completely continuous. The 24 bits ahead of the IP address are the same, and the last 8 bits range from 112 to 120.

It can thus be speculated that the SGSNs in the two clusters are carrying large traffic, and more accessed from users.

In order to explore the users' behavior behind the data, we inquire the congruent relationship between the SGSN IPs and the actual location. Further observing Table II and contrasting with the geographic position, we find that the cluster result just explain the actual behavior of the subscribers very well. It is well-known that the subscribers consume mobile traffic under diverse tariff. The expense is low at local, which means when the subscriber connect to the mobile Internet in the phone number registered place, the expense is low. With the corresponding, when the subscriber is at nonlocal, the tariff is high. Under the influence of expense, it is accessible that the users reduce the online behavior.They are inclined to use mobile Internet applications more often, which brings about the traffic decline. Turning to our analysis result, the clusters with higher traffic are the local SGSNs, such as Cluster VI and Cluster VII. These SGSN are all spread over the city we collect data from, that is the subscribers we study register the phone numbers. The clusters with lower traffic spread over all parts of the country, whose traffic is generated by wandering subscribers. To be specific, the SGSN IPs in Cluster III correspond to 25 different provinces, and the traffic flow through these SGSNs is much less than others.

B. *The law of the application in SGSN clusters*

We illustrate the characteristics of each SGSN by analyzing the application constitution of 8 cluster-centers. The six applications' proportions of each cluster-center are shown in Figure 1.

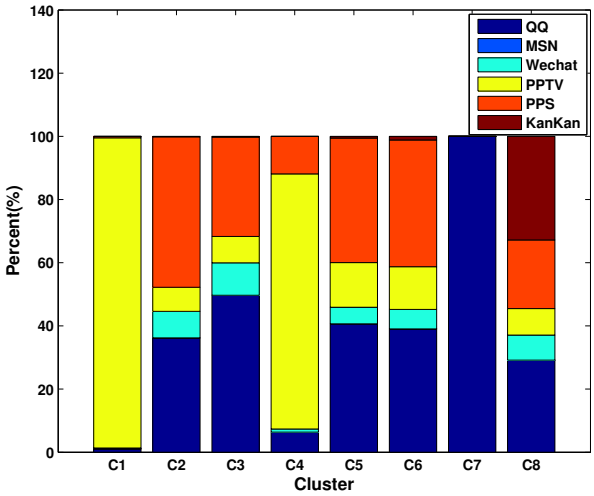

Fig. 3. Application percentage of each cluster-center

By observing Figure 1, we find that the center traffic composition of Cluster I and Cluster VII are obviously tendentiousness. In Cluster I whose SGSNs number is 5, PPTV application accounts for 98.2% of the total volume, and in Cluster IV whose SGSNs number is 4, the PPTV traffic is also relatively higher, accounting for 80.8% of the total flow. In Cluster VII whose SGSNs number is 11, the application proportion of QQ is almost 100%, while the rest of the five applications, besides KanKan having little traffic, MSN, WeChat, PPTV, PPS four types of application volumes are all zero. In addition, no matter for Cluster II and Cluster III who have the lower total traffic or Cluster V and Cluster VI who have larger total traffic, their application composition has similar rule, that is QQ accounting the biggest proportion of IM while the PPS volume being higher than other streaming media applications. The sum traffic of these two applications achieves about 80% of the total traffic.The percentage in other

clusters is around 1%. The statistical result just reflects their market share at present.

IV. Conclusion

This paper probe into a set of traffic characteristics analysis methods, and applies these methods on the massive real traffic data analysis from the mobile Internet. We obtain flow direction rules of media streaming.

Paying special attention to the data flow direction, we analyze the characteristics of the main flow traffic of streaming media. The analysis result shows that the distribution of streaming media traffic is especially dispersed, that means the destination IPs scattered around the country. Furthermore we cluster the SGSNs of a city through our adaptive clustering algorithm which describes the regional characteristic of applications. Results of the paper have important significance on the targeted bandwidth deployment, green resource scheduling, network planning and optimization of operators and ISPs.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61201153), the National 973 Program of China under Grant (2012CB315805), and Prospective Research Project on Future Networks in Jiangsu Future Networks Innovation Institute(BY2013095-2-16).

References

[1] "Mobile data traffic to grow 300 percent globally by 2017 led by video, web use, says strategy analytics," *[Online]*, http://techcrunch.com/2013/07/03/mobile-data-use-to-grow-300-globally-by-2017-led-by-video-web-traffic-says-strategy-analytics/.

[2] A. Vespignani, "Predicting the behavior of techno-social systems," *Science*, vol. 325, no. 5939, p. 425, 2009.

[3] H. Cui, C. Feng, Y. Chai, R. P. Liu, and Y. Liu, "Effect of hybrid circle reservoir injected with wavelet-neurons on performance of echo state network," *Neural Networks*, pp. 144–151, 2014.

[4] M. A. Awad and I. Khalil, "Prediction of user's web-browsing behavior: Application of markov model," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 4, pp. 1131–1142, 2012.

[5] S. Molnár, P. Megyesi, and G. Szabó, "Multi-functional emulator for traffic analysis," in *Proc. IEEE ICC*, 2013.

[6] Y. Wu, G. Min, K. Li, and B. Javadi, "Modeling and analysis of communication networks in multicluster systems under spatio-temporal bursty traffic," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 23, no. 5, pp. 902–912, 2012.

[7] C. Dong, S. Zhang, Z. Lei, J. Yang, and G. Cheng, "Analyzing gprs mobile network traffic with map reduce."

[8] M. V. de Assis, L. F. Carvalho, J. J. Rodrigues, and M. L. Proenca, "Holt-winters statistical forecasting and aco metaheuristic for traffic characterization," in *Communications (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2524–2528.

[9] I. Spanou, A. Lazaris, and P. Koutsakis, "Scene change detection-based discrete autoregressive modeling for mpeg-4 video traffic," in *Communications (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2386–2390.

[10] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 882–890.

[11] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3g cellular data network," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 1341–1349.

[12] C. V. N. Index, "Global mobile data traffic forecast update, 2010-2015," *Cisco white paper*, 2011.

[13] H. C. . K. Z. et al, "A novel clustering validity index based on frequent pairwise pattern," *Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on .IEEE*, pp. 1–5, 2013.