Measurement and Analysis of Cloud User Interest: A Glance From BitTorrent

Lei Ding*, Yang Li*, Haiyang Wang*, and Ke Xu†

*University of Minnesota Duluth, Duluth, Minnesota, USA

†Tsinghua University, Beijing, China

{dingx426, yangli, haiyang}@d.umn.edu, xuke@tsinghua.edu.cn

ABSTRACT

Cloud computing has recently emerged as a compelling method for deploying and delivering services over the Internet. Its popularity greatly boosts the development of cloud-based systems while also generates a considerable amount of cloud traffic. To better manage this increasing traffic, the understanding of cloud user interests is therefore important for both content and Internet service providers.

In this paper, we aim to shed new light on the learning of cloud user interest. Our study for the first time shows the existence of cloud users in such real-world content distribution systems as BitTorrent. Based on this observation, we further explore the similarity of content preferences between cloud and non-cloud users. Surprisingly, our statistical model analysis indicates that the users in the cloud AS have significantly different interests from all the observed non-cloud ASes. More dedicated researches are therefore required to better manage this elevating yet unique cloud traffic in the future.

1. INTRODUCTION

Cloud computing has rapidly emerged as the driving trend in global Internet services. A wide spectrum of applications is now migrating to the cloud computing platforms, looking for better performance, reliability, and cost efficiency [13]. Moreover, the ability of cloud architectures to scale rapidly and efficiently support more workloads than traditional data centers also entices more organizations to move their workloads to the cloud. Such an elevating popularity, however, generates a considerable amount of cloud traffic and puts unprecedented pressure on the Internet. According to the latest *Cisco Global Cloud Index* [6], the annual global cloud IP traffic will reach 14.1 ZB (1.2 ZB per month) by the end of 2020, up from 3.9 ZB per year (321 EB per month) in 2015. To better manage this increasing traffic, the understanding of cloud user interest is therefore critical for both content and Internet service providers (ISPs). For example, with the knowledge of similarity/dissimilarity between cloud and non-cloud users, the ISPs will be able to better manage their caching contents based on cloud users' special preferences. Unfortunately, the enterprise cloud systems are mostly distributed, and the passive measurement of global cloud user interest is therefore nearly impossible. To make the matter worse, the service level agreement (SLA) [20] also prevents cloud service providers, such as Amazon [2], Google [10], and Rackspace [18], to release their detailed user interests to the general public. Therefore, the existing cloud-based measurement studies are generally limited to the resource usage behavior as well as system scaling issues [1, 19].

To mitigate such a challenge, our study for the first time explores the existence of cloud users in such real-world content distribution systems as BitTorrent. We successfully captured the existence of cloud peers in BitTorrent (BT) networks. This observation naturally bridges Internet P2P systems to cloud computing. In particular, our measurement indicates that 17 percent of BT torrents has over 10% peers from cloud. The ratio of cloud peers can even exceed 50% for some very popular torrents. Moreover, the existence of cloud user in BT also provides an initial yet important step to understand the similarity of cloud and non-cloud autonomous systems (ASes). We find that the cloud users are distributed inhomogeneously over the Internet torrents. In detail, the cloud users are more interested in torrents with movie and TV contents. Very few cloud users are willing to join torrents with music and software applications.

It is worth noting that we can hardly compare the cloud AS to all other ASes on the Internet. It is therefore hard to say if the cloud users are indeed having different interests from all the other users. To this end, we further developed a novel clustering approach to measure the similarity of user interests across all the observed ASes. This approach successfully distinguished Amazon cloud with other non-cloud ASes. This implies that the cloud users/ASes have significantly different interests from all other non-cloud ASes. Therefore, we will need to build special and more dedicated traffic management strategies to manage/optimize the cloud users in the future.

The rest of this paper is organized as follows. In Section 2, we present the related works. After that, Section 3 discusses the measurement configuration as well as the obser-

vations. Section 4 further explores the similarity of user interest based on AS clustering. Section 5 summarizes some further discussions and concludes the paper.

2. RELATED WORKS

There have been numerous studies on the measurement and analysis of Internet user interests as well as their content preferences. Guo *et al* [11] proposed a cooperative caching and sharing scheme of user's smart devices by user's interest. Man *et al* [16] measured mobile user interests by investigating large log data from mobile networks. Following these pioneer studies, Xia *et al* [21] tried to understand user interests from their geographic information in the traffic. It is also known that the understanding of user interest plays an important role in many Internet applications such as BitTorrent and video streaming systems. For example, Lu Jie *et al* [15] built a user interest model based on their historical query logs. Chen *et al* [5] seek to enhance search performance in unstructured P2P networks through exploiting users' interest. Moreover, Qin *et al* [17] adopted the classic k-medoids clustering to construct user Interest model for P2P document sharing systems. For video streaming systems, Yu *et al* [22] used a modified Poisson distribution to understand the user behaviors in a large-scale video-on-demand system. Dernbach *et al* [8] applied the observed user interest on cache content-selection for streaming video services.

Differing from the existing studies, this paper for the first time bridges the cloud user interest to the understanding of existing P2P systems. Our measurement shows that the cloud users are now an emerging force in such Internet applications as BitTorrent. Moreover, we also apply statistic models to better understand the unique content preferences of cloud ASes as well as their similarity to the traditional non-cloud ASes. Such comparison is significantly different from the existing cloud-based measurement studies [12, 9, 3].

3. MEASUREMENT OF CLOUD USER INTEREST FROM BITTORRENT

3.1 Measurement Methodology

In this experiment, we have applied a PlanetLab-based experiment to obtain the peer information of Internet BitTorrent swarms. Our design uses the PlanetLab as a large collection of distributed probing nodes to interact with real-world trackers and peers. We extracted a large number of real torrents as advertised by `www.limetorrents.cc`, one of the most popular torrent sites, from Dec 2016 to May 2017. In particular, we developed a Python script to automatically detect torrent URLs in each given webpages and downloaded the meta-info files ending with ".torrent", which resulted in $149,578$ meta-info files. It is known that the most popular torrents generate the majority of BT traffic [14]. To simplify our later statistical analysis, we selected the top 100 most popular torrents for discussion. The selection was done in a completely random way so the selection bias could be neglected in this study. These files include seven major content types: Movies, TV shows, Music, Games, Applications, Anime, and Others (unknow file type).

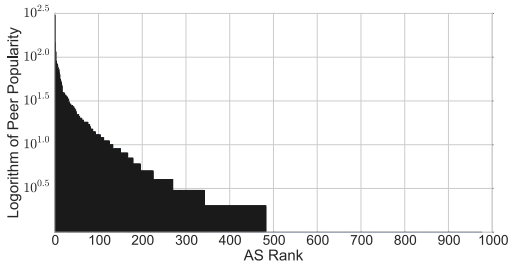

Figure 1: Peer popularity in torrents

To obtain the peer information in these torrents, we ran a modified version of CTorrent on the PlanetLab nodes. Ctorrent is an open-source and very classic BitTorrent client in FreeBSD. Different from the existing content-based experiments, our modified CTorrent clients actively joined existing torrents in the global Internet and recorded the observable peer information from the trackers and from other peers over time. As such, the small set of controlled PlanetLab nodes were able to capture the information of most peers in the BT swarms. Except for retrieving the peer existence and address information, our PlanetLab clients did not download or upload any real data of the shared contents. Hence, no copyrights were violated. The scanning efficiency of the experiment is also very high, with most of the torrent being finished scanning within a short timeframe (< 30 sec). To avoid biases, we have also filtered out all the PlanetLab nodes in the data of the following analysis. Our source code of the modified CTorrent client, as well as the raw data set (including the torrents information), can be found at `https://drive.google.com/drive/folders/0B3Mp-OEG5ahkN2I5eWhaaFl3b3c?usp=sharing`

3.2 Observations

Given the IP addresses of the peers, we extracted their corresponding ASes as well as the location information through the "whois" command in Linux. This resulted in 5569 peers from 976 distinct ASes. Figure 1 shows the peer popularity across all torrents in these ASes. We can see that it can roughly be fitted by an exponential distribution ($y = e^{ax+b}$, where $a = -0.0112$, $b = 3.9295$). This means a majority of the ASes, however, do not host a noticeable number of BitTorrent peers, e.g., 65% of them have less than 100 peers across all torrents. Moreover, the peers in our dataset are located in 147 different countries. The top five ranked countries are US, Philippines, Brazil, UK, and Canada, and the portion of top five countries can be found in Figure 2. It is also easy to see that the top five countries take up over 40% of the total peers.

Different from the existing P2P measurement studies, our dataset indicates that a considerable number of BT peers are from cloud. In particular, Table 1 shows the peer population of the top-10 most popular ASes in our dataset. We

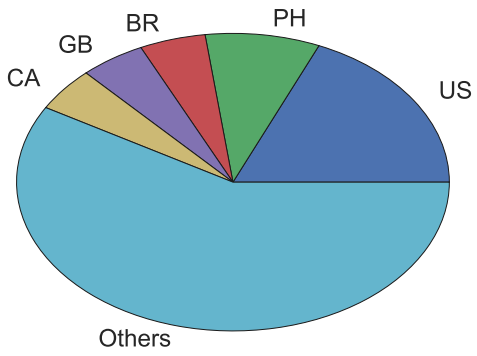

Figure 2: Geographic distribution of BT peers

ISP name	Peer#
Philippine Long Distance Telephone Company, PH	303
AMAZON-02 - Amazon.com, Inc., US	146
Comcast Cable Communications, LLC, US	115
CLARO S.A., BR	91
SOFTLAYER - SoftLayer Technologies Inc., US	85
NTL, GB	80
TELKOMNET-AS2-AP PT Telekomunikasi Indonesia, ID	76
ASN-TELSTRA Telstra Pty Ltd, AU	74
TMNET-AS-AP TM Net, Internet Service Provider, MY	72
GLOBE-MOBILE-5TH-GEN-AS Globe Telecom Inc., PH	66

Table 1: Peer population in top-10 ASes

can see that there are 146 BT peers from Amazon. Our investigation indicates that these IP addresses are assigned to Amazon's EC2 [2] virtual machines (VMs). As a popular cloud service provider and an autonomous system[1], Amazon is ranked at the second most popular AS in our dataset. More cloud peers will be observed if we further extend the scale of our analysis. In Figure 3, we can see the popularity of cloud peers in different torrents. In detail, many torrents have over 10% peers from cloud. The ratio of cloud peers can even exceed 50% for some very popular torrents. This means the existence of cloud peers is not a special case in BT torrents. Therefore, we can use this information to understand the cloud user interest in BitTorrent.

To explore cloud user's preference, Figure 4 compares Amazon to three very popular ASes in our dataset. We can see that the users in typical non-cloud ASes, such as *COMCAS* and *CLARO*, are more equally distributed in different torrents. The distribution of Amazon users, on the other hands, is clearly skewed. This means the cloud users are more likely to have clear preferences on certain types of contents/torrents which are movie and TV contents. Very few cloud users are willing to join torrents with music and software applications.

Based on the above measurements, it is easy to see that the cloud users are now an emerging force in such Internet applications as BitTorrent. These users also have a clear preference on movie and TV contents. However, we do not know if their detailed interests are similar/dissimilar to all other non-cloud ASes.

[1]It is known that Amazon consists of many autonomous systems. For the sake of simplicity, we use one AS (ASN:16509) to refer Amazon in this paper.

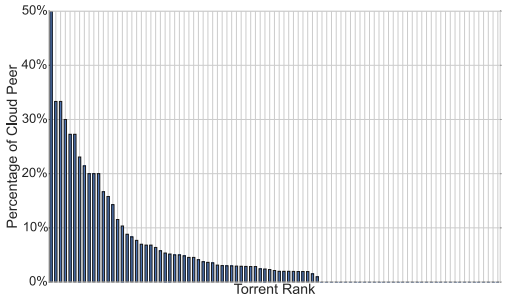

Figure 3: Percentage of cloud peers in torrents

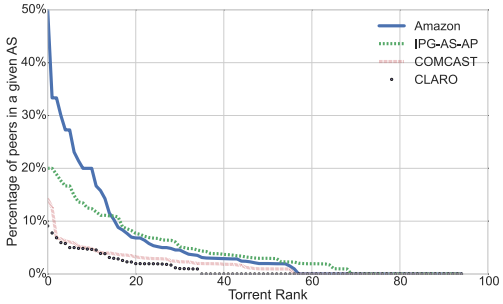

Figure 4: Peers ratio of different ASes

4. STATISTICAL ANALYSIS

The existence of the cloud-based AS (Amazon) brings up the question of learning its characteristics and its difference from the traditional ASes. In this section, we first preprocess the data set to a format suitable for carrying out the statistical analysis. By using a metric of pairwise correlation and clustering analysis, we show that Amazon cloud has some distinctive behaviors which can distinguish it from the traditional ASes.

4.1 Data Preprocessing

As mentioned in the previous sections, the data set contains the information of sampled torrent files and the IP addresses of peers which were downloading those files at the time of sampling. The data set can be described by an $m \times n$ data matrix $\mathbf{A}$, where m rows correspond to m different torrent files, including movies, TV shows, games, etc, while n columns correspond to n distinct ASes. In other words, element A_{ij} represents the count of peers in the jth AS which were downloading the ith torrent file. Each AS can therefore be represented by a m-dimensional vector called its profile. The profile of the kth AS is $I_k = (A_{1k}, A_{2k}, \ldots, A_{mk})$ where $k = 1, 2, \ldots, n$.

For an easy comparison, Amazon cloud is placed in the first column while all other traditional ASes are ranked in decreasing order by their total number of peers. Matrix $\mathbf{A}$ happens to be a sparse matrix since we sample for a short period of time and therefore some ASes might not have any downloading peers. It is observed that 10% of the ASes con-

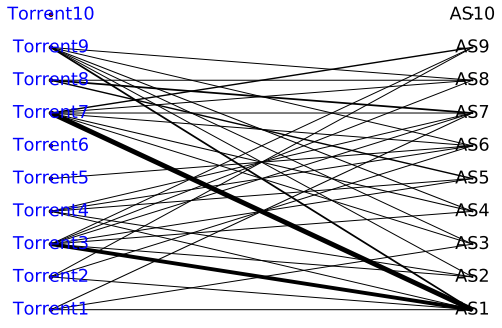

Figure 5: Bipartite graph representing the data set. Only the first ten ASes and ten torrent files are shown. The line width connecting an AS and a file represents the number of peers in the AS which were downloading the file.

tribute to about 90% of the total number of peers. We decide to analyze only the top 100 ASes since small ASes with many zeros do not give much information on the downloading behavior of their peers. In the future, if the experiment time is increased substantially, more ASes could be included in the analysis.

The structure of the data set can alternatively be illustrated using a weighted bipartite graph. A bipartite graph, also called a two-mode graph, is a graph whose vertices can be divided into two disjoint sets U and V such that edges only connect vertices between U and V. There are no edges between vertices in the same set. In this study, these two types of nodes are torrent files and ASes. Edges of the graph only exist between vertices of different types, that is, between a torrent file and an AS. The weight of the edge is the number of peers within the AS which were downloading that file at the time of the sampling. Figure 5 shows part of the bipartite graph generated in this study.

4.2 Projection

Since the goal of this study is to analyze the behavior of ASes while torrent files just play the role of linkage, we perform a projection onto the AS space by constructing an n-vertex simple graph where vertices represent ASes and two ASes are connected by an edge with a weight corresponding to their similarity. There are various ways to define similarity between two vertices. For example, several association indices were discussed in [4].

A common and natural choice is the Pearson correlation coefficient which is a measure of linear correlation between two sequences. In the current study, the sequences are the ASes profiles. Given two ASes profiles I_k and I_l, each being an m-vector, their Pearson correlation coefficient is computed as

$$r_{k,l} = \frac{\sum_{i=1}^{m}(A_{ik} - \bar{A}_k)(A_{il} - \bar{A}_l)}{\sqrt{\sum_{i=1}^{m}(A_{ik} - \bar{A}_k)^2}\sqrt{\sum_{i=1}^{m}(A_{il} - \bar{A}_l)^2}}, \quad (1)$$

where $\bar{A}_k$ is the average of the kth profile. An $n \times n$ correlation matrix $\mathbf{B}$ with elements $B_{kl} = r_{k,l}$ results from (1).

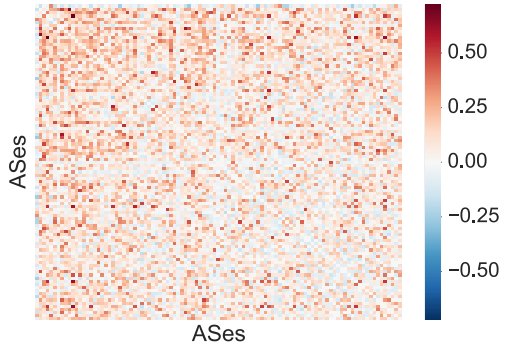

Figure 6: Heat map of matrix B. Warm colors indicate positive Pearson correlations between the pairs of ASes while cool colors indicate negative correlations. Note that Amazon cloud is in the first row and first column.

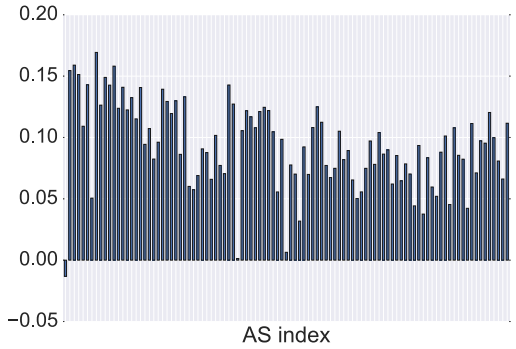

Figure 7: Mean of the Pearson correlation coefficients of an AS to all other ASes. Each bar represents one AS. Note that Amazon is the first bar which is the only one with negative mean correlation.

All entries in $\mathbf{B}$ range from -1 to 1, indicating whether the peers in two ASes share common downloading behaviors. The heat map representing $\mathbf{B}$ is shown in Figure 6.

4.3 Summary Statistics

Since we intend to find out whether Amazon cloud is similar or dissimilar to the traditional ASes, we first calculate some global summary statistics to search for an overall similarity patterns. We compute the column-wise mean values in the correlation matrix $\mathbf{B}$ which represents the overall similarity of a given AS to all other ASes in the study. If an AS has similar peer downloading behaviors as most other ASes, this means value tends to be positive, and vice versa. Figure 7 shows a bar plot of these mean values. In particular, the first bar, which represents Amazon cloud, is the only one with negative mean value. It shows that, on average, Amazon is dissimilar to other traditional ASes in terms of peers' downloading behaviors. All other bars in Figure 7 have positive means, indicating that all traditional Ases share some common downloading properties.

4.4 Hierarchical Clustering

Figure 7 gives us a general idea that there is indeed dissim-

ilarity between Amazon and the traditional ASes in terms of their peers' downloading behaviors. However, by taking the mean values, some detailed information is smeared out and lost. Remember that each bar in Figure 7 represents the average of $n-1$ Pearson correlation coefficients. For example, a positive mean value does not mean all correlation coefficients are positive.

To better understand the detailed patterns among all ASes, we perform a clustering analysis to group those ASes with similar downloading behaviors in the same clusters. Entry B_{kl} in the correlation matrix $\mathbf{B}$ is a similarity measure between two profiles I_k and I_l. The bigger the value B_{kl}, the more similar these two profiles are. Therefore, B_{kl} can be considered as a "distance", such that I_k and I_l have a small distance if B_{kl} is large. Traditionally, in correlation clustering, we work on the matrix $1-\mathbf{B}$ instead of $\mathbf{B}$ since entries in $1-\mathbf{B}$ are positively correlated with the distance. Specifically, if the distance between profiles I_k and I_j is large, $(1-\mathbf{B})_{kl}$ will be large and vice versa.

Since we only have the pairwise distances between AS profiles instead of their individual coordinates, some classical clustering algorithms such as K-means do not work here. Instead, we use the hierarchical clustering which only requires the pairwise distances between each pair of points. Hierarchical agglomerative clustering (HAC) is a general family of clustering algorithms that build nested clusters by merging the data successively in a bottom-up manner [7]. The hierarchy of the data is often represented by a tree called dendrogram. Initially, each single data point starts in its own cluster. Pairs of clusters are merged together to form a bigger cluster and so on until all data points belong to one single cluster. After forming the whole hierarchy tree, we can obtain small clusters by cutting the hierarchy tree into pieces.

The dendrogram in this study is shown in Figure 8, which is then cut into separate groups according to the desired number of clusters c. Figure 9 shows the heat map for $c = 2$ where the ASes are divided into two clusters of size 5 and 96. Amazon cloud is located in the first cluster which is in the upper left corner of Figure 9. It is separated from the majority of the data indicating a different user downloading interest from most of the traditional ASes.

We also try $c = 3$ and divide the ASes into three clusters. The result is shown in Figure 10. In this setting, the smaller cluster of size 5 is untouched. The bigger cluster of size 96 is further divided into two sub-clusters with sizes 54 and 42, respectively. For even larger values of c, the general picture is very similar. Amazon tends to be in a very small cluster, indicating its distinct downloading behaviors compared to the traditional ASes.

5. CONCLUSION

This paper takes an initial step towards the understanding of cloud user interest. Our measurement from BitTorrent for the first time showed the existence of cloud peers in BT. The follow-up comparison further revealed that the user interest of cloud users/ASes is significantly different from the classic non-cloud users/ASes. There are many possible future directions to explore. We are particularly interested in the detailed reasons of why cloud users/ASes are so unique. Moreover, we also aim to explore better traffic management approaches to handle the increasing cloud traffic.

6. REFERENCES

[1] O. A. Abdul-Rahman and K. Aida. Towards understanding the usage behavior of google cloud users: The mice and elephants phenomenon. In *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pages 272–277, December 2014.

[2] Amazon, Inc. Amazon Elastic Compute Cloud (Amazon EC2). `http://aws.amazon.com/ec2/`.

[3] J. Baliga, R. W. A. Ayre, K. Hinton, and R. S. Tucker. Green cloud computing: Balancing energy in processing, storage, and transport. *Proceedings of the IEEE*, 99(1):149–167, January 2011.

[4] J. I. F. Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. M. Walhout. Using networks to measure similarity between genes: association index selection. *Nature Methods*, 10:1169–1176, December 2013.

[5] G. Chen, C. P. Low, and Z. Yang. Enhancing search performance in unstructured p2p networks based on users' common interest. *IEEE Transactions on Parallel and Distributed Systems*, 19(6):821–836, June 2008.

[6] Cisco. Cisco global cloud index: Forecast and methodology, 2015-2020. `http://www.cisco.com`.

[7] W. H. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[8] S. Dernbach, N. Taft, J. Kurose, U. Weinsberg, C. Diot, and A. Ashkan. Cache content-selection policies for streaming video services. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, April 2016.

[9] I. Foster, Y. Zhao, I. Raicu, and S. Lu. Cloud computing and grid computing 360-degree compared. In *2008 Grid Computing Environments Workshop*, pages 1–10, November 2008.

[10] Google, Inc. Google Cloud Platform. `http://cloud.google.com`.

[11] Z. Guo, H. Jin, C. Zhao, and D. Liang. User interest based distributed cooperative caching and sharing in wireless networks. In *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 466–470, September 2016.

[12] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema. Performance analysis of cloud computing services for many-tasks scientific computing. *IEEE Transactions on Parallel and Distributed Systems*, 22(6):931–945, June 2011.

[13] A. Khajeh-Hosseini, D. Greenwood, and I. Sommerville. Cloud migration: A case study of migrating an enterprise it system to iaas. In *2010 IEEE 3rd International Conference on Cloud Computing*, pages 450–457, July 2010.

[14] S. Le Blond, A. Legout, and W. Dabbous. Pushing bittorrent locality to the limit. *Comput. Netw.*, 55(3):541–557, February 2011.

[15] J. Lu and J. Callan. User modeling for full-text federated search in peer-to-peer networks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 332–339, New York, NY, USA, 2006. ACM.

[16] N. Man, C. Xunxun, and W. Bo. Hierarchical user interest model based on large log data of mobile internet. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, pages 1–5, June 2016.

[17] C. Qin, Z. Yang, and H. Liu. User interest modeling for p2p document sharing systems based on k-medoids clustering algorithm. In *2014 Seventh International Joint Conference on Computational Sciences and Optimization*, pages 576–578, July 2014.

[18] Rackspace, Inc. Rackspace Cloud Platform. `https://www.rackspace.com/`.

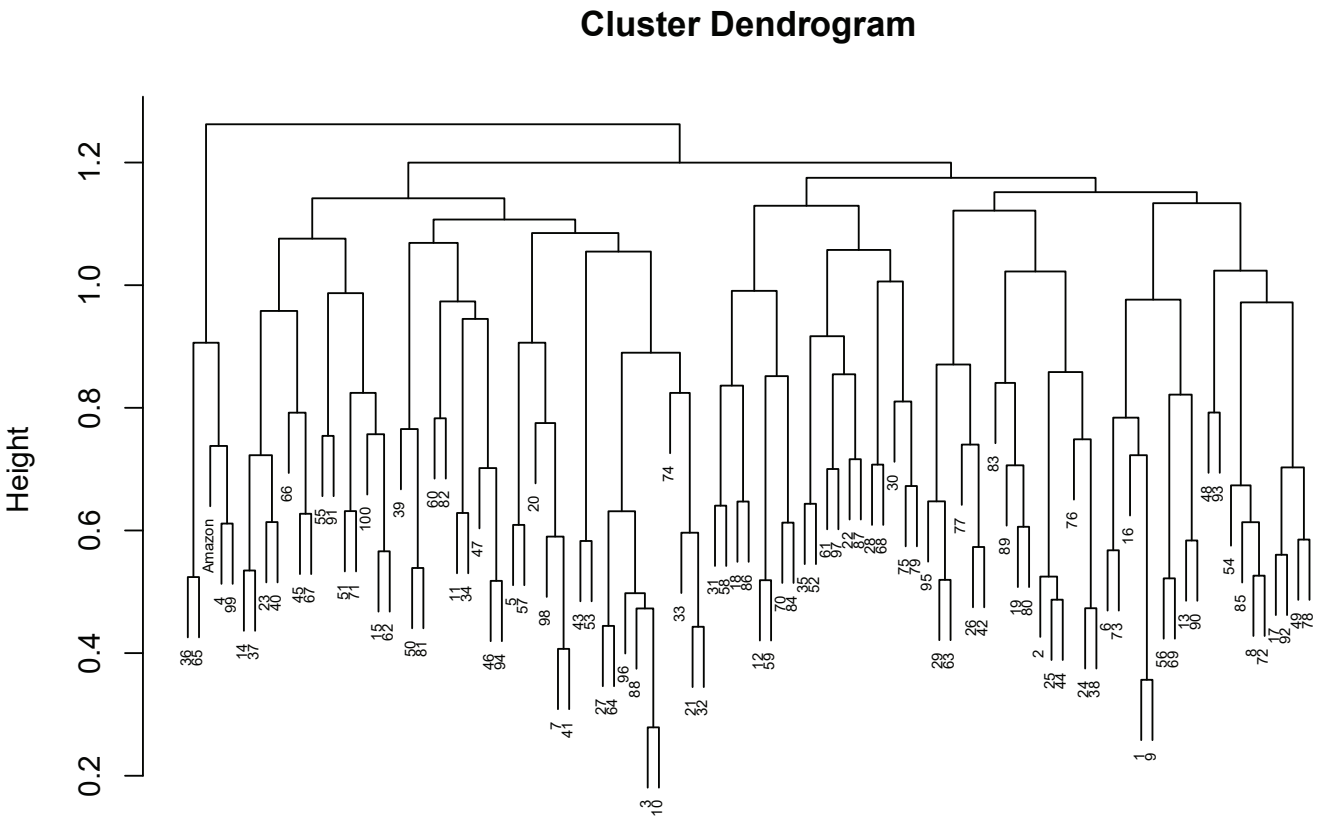

Figure 8: Cluster dendrogram using hierarchical clustering algorithm. Traditional ASes are labeled from 1 through 100. The names of the top 9 traditional ASes are listed in Table 1. This dendrogram is then cut to generate clusters.

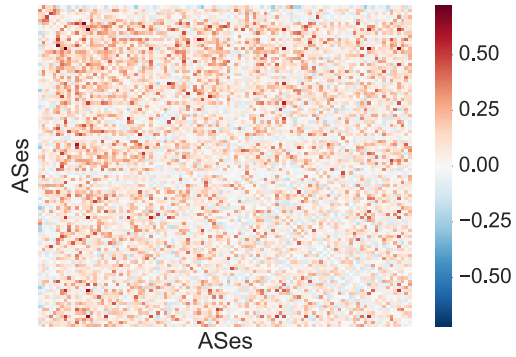

Figure 9: Heat map of the generated clusters $(c = 2)$. 101 ASes are divided into two clusters. The first cluster which is shown in the upper left corner has 5 ASes including Amazon cloud. The second cluster has 96 ASes.

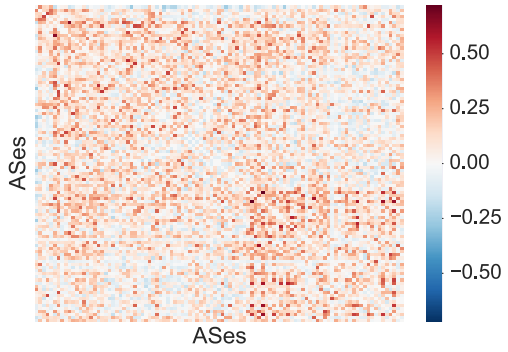

Figure 10: Heat map of the generated clusters $(c = 3)$. The first cluster of size 5 is the same as the one in Figure 9. The second and third clusters each have 54 and 42 ASes.

[19] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *Proceedings of the Third ACM Symposium on Cloud Computing*, SoCC '12, pages 7:1–7:13, New York, NY, USA, 2012. ACM.

[20] P. Wieder, J. Butler, W. Theilmann, and R. Yahyapour. *Service Level Agreements for Cloud Computing*. Springer, 2011.

[21] N. Xia, S. Miskovic, M. Baldi, A. Kuzmanovic, and A. Nucci. Geoecho: Inferring user interests from geotag reports in network traffic. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 1–8, August 2014.

[22] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. In *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*, EuroSys '06, pages 333–344, New York, NY, USA, 2006. ACM.