iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 5 **Issue:** V **Month of publication:** May 2017

**DOI:**

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# A Peer to Peer Traffic Identification Method Using K-Means Clustering, SVM & Genetic Algorithm

Puja Raghav[1], Dr. Vivek Jaglan[2], Mr. Akshat Aggarwal[3]
[1,2,3,4] *Department Of Computer Science and Engineering*
*AMITY University Haryana, Gurgaon*

***Abstract:*** ***the utilization of shared (p2p) applications is developing significantly, which brings about a few major issues, for example, the system clog and traffic obstruction. Consequently p2p traffic identification is the most blazing theme of p2p traffic administration. Support vector machine (svm) has points of interest with settling little examples for p2p characterization issues. However, the execution of svm is basically reliant on its parameters. In this paper we propose genetic algorithm and k-means with svm to streamline the parameters of svm and have been connected to p2p traffic identification. The curiosity of the proposed strategy is that it uses just the extent of parcels traded between ips inside seconds. The recognized components of the proposed technique lie in that quick calculation, high identification precision, and asset sparing capacity. At last, experiment results demonstrate the satisfactory performance of the proposed method.***
***Keywords- p2p, svm, genetic algorithm, k-means algorithm***

## I. INTRODUCTION

*A. Overview*

A dispersed (P2P) framework is social affair of PCs, each of which goes about as a center point for sharing records inside the get-together. As opposed to having a central server to go about as a typical drive, each PC goes about as the server for the records set away upon it. Right when a P2P framework is developed over the Web, a central server can be used to record reports, or a scattered framework can be set up where the sharing of records is part between each one of the customers in the framework that are securing a given record.

In the most major sense, a disseminated framework is a clear framework where each PC fills in as a center point and a server for the records it just holds. These are the same as a home framework or office sorts out. Regardless, when P2P frameworks are developed over the web, the traverse of the framework and the archives open empower huge measures of data to be shared. Early P2P frameworks like Napster used client programming and a central server, while later frameworks like Kazaa and BitTorrent disposed of the central server and part up sharing commitments between various center points to free up information transmission. Appropriated frameworks are by and large associated with Web burglary and unlawful record sharing.

The fundamental use of P2P frameworks in business took after the association in the mid-1980s of unsupported PCs. Instead of the little unified PCs of the day, for instance, the Versus structure from Wang Research focuses Inc., which served up word planning and diverse applications to numbskull terminals from a central PC and set away records on a central hard drive, the then-new PCs had autonomous hard drives and certain CPUs.The insightful boxes in like manner had on board applications, which suggested they could be passed on to desktops and be useful without an umbilical rope associating them to a brought together server.

In its minimum troublesome casing, a disseminated (P2P) framework is made when no less than two PCs are related and share resources without encountering an alternate server PC. A P2P framework can be an unrehearsed affiliation—a couple of PCs related by methods for a Widespread Serial Transport to trade records. A P2P organize in like manner can be an interminable system that associations around six PCs in a little office over copper wires. Or, then again a P2P framework can be a framework on an extensively all the more stunning scale in which uncommon traditions and applications set up direct associations among customers over the Web.

In a P2P orchestrate, the "mates" are PC systems which are related with each other by methods for the Web. Records can be shared particularly between structures on the framework without the need of a central server. In a manner of speaking, each PC on a P2P composes transforms into a record server and furthermore a client.

The principle necessities for a PC to join a circulated framework are a Web affiliation and P2P programming. Customary P2P programming programs join Kazaa, Limewire, BearShare, Morpheus, and Procurement. These tasks connect with a P2P framework, for instance, "Gnutella," which empowers the PC to get to an expansive number of various structures on the framework.



©IJRASET: All Rights are Reserved

At the point when related with the framework, P2P programming empowers you to check for records on other people's PCs. At that point, diverse customers on the framework can filter for archives on your PC, however frequently simply inside a singular envelope that you have allocated to share. While P2P sorting out makes report sharing basic and beneficial, is also has provoked a huge amount of programming theft and illegal music downloads. Henceforth, it is best to fail in favor of alert and simply download programming and music from honest to goodness locales.

Dispersed (P2P) is a decentralized trades show in which each social occasion has comparable capacities and either get-together can begin a correspondence session. Not in the least like the client show, in which the client makes an organization request and the server fulfills the request, the P2P orchestrate exhibit empowers each center to act as both a client and server.

P2P structures can be used to give anonymized directing of framework development, huge parallel figuring circumstances, scattered limit and distinctive limits. Most P2P activities are fixated on media sharing and P2P is thusly as often as possible associated with programming and copyright encroachment.

Typically, conveyed applications empower customers to control various parameters of operation: what number of part relationship with search for or allow at one time; whose structures to interface with or avoid; what organizations to offer; and what number of structure advantages for devote to the framework. Some basically connect with some subset of dynamic centers in the framework with little customer control, in any case.

*B. K-Means Algorithm*

K-Means Algorithm is most ordinary and pervasive grouping gadget that is by and large used as a piece of various applications and it falls under the allotting calculations that points in building the different examples and assesses them by utilizing some model. With the given social occasion of n data, k different bundles are formed with each gathering having a stand-out centroid (mean) and accordingly the allotting is made. The letter k portrays the amount of groups ought to have been made. Right when number of n articles is to be amassed into k gatherings, K pack center is to be instated. Each question will be given to the closest gathering center and .the point of convergence of bundle is invigorated each time until state of no change occurs in the each gathering. The segments in each gathering will be in close contact with centroid of that particular group and will be unmistakable to the segments having a place with various bundles.

The total of the inconsistencies between the point and the centroid conveyed by specific detachment is used as the objective work. Indicate intra-amass contrast portrays the total of the squares of the slip-up between the point and separate centroids.

*C. Genetic algorithm*

Innate Calculations (GAs) are flexible heuristic request count in perspective of the transformative considerations of typical assurance and inherited qualities. Appropriately they address a canny abuse of a self-assertive interest used to handle upgrade issues. Though randomized, GAs are by no means, self-assertive, rather they abuse undeniable information to arrange the chase into the locale of better execution inside the request space. The basic systems of the GAs are planned to reproduce shapes in like manner structures vital for progression; especially those take after the models at first set around Charles Darwin of "survival of the fittest."

GAs relies on upon a closeness with the innate structure and direct of chromosomes inside a populace of people utilizing the accompanying establishments:

*1)* Individuals in a populace vie for assets and mates.
*2)* Those people best in every "opposition" will create more posterity than those people that perform ineffectively.
*3)* Genes from `good' people spread all through the populace so that two great guardians will now and again create
*4)* Suited to their condition. Genetic algorithm
*a)* randomly introduce population (t)
*b)* determine wellness of population(t)
*c)* repeat
*d)* select guardians from population(t)
*e)* perform hybrid on guardians making population(t+1)
*f)* perform change of population(t+1)
*g)* determine wellness of population(t+1)
*h)* until best individual is sufficient



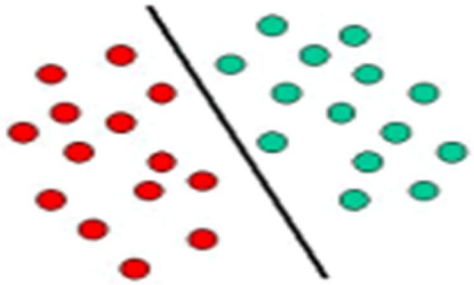
©IJRASET: All Rights are Reserved

*D. Support vector machine*

"Support Vector Machine" (SVM) is a controlled machine learning count which can be used for both request and backslide challenges. In any case, it is by and large used as a piece of plan issues. In this computation, we plot each data thing as a point in n-dimensional space (where n is number of parts you have) with the estimation of every component being the estimation of a particular mastermind. By then, we perform gathering by finding the hyper-plane that different the two classes outstandingly well (look at the underneath review). Support Vectors are fundamentally the co-ordinates of person perception. Support Vector Machine is a wilderness which best isolates the two classes (hyper-plane/line).

Support vector machines (SVMs) are an arrangement of directed learning strategies utilized for characterization, relapse and anomaly's identification.

Support Vector Machines depend on the idea of choice planes that characterize choice limits. A choice plane is one that isolates between arrangements of items having diverse class participations. A schematic case is appeared in the representation beneath. In this case, the items have a place either with class GREEN or RED. The isolating line characterizes a limit on the correct side of which all articles are GREEN and to one side of which all items are RED. Any new protest (white hover) tumbling to the privilege is marked, i.e., arranged, as GREEN (or named RED should it tumble to one side of the isolating line)



.

Fig: 1.1 Grouping of articles

Support Vector Machine (SVM) is basically a more tasteful strategy that performs order undertakings by developing hyperplanes in a multidimensional space that isolates instances of various class marks. SVM underpins both relapse and order errands and can deal with various persistent and straight out factors. For downright factors a spurious variable is made with case values as either 0 or 1. Therefore, an all-out ward variable comprising of three levels, say (A, B, C), is spoken to by an arrangement of three sham factors:

A: {1 0 0}, B: {0 1 0}, C: {0 0 1}

To build an ideal hyperplane, SVM utilizes an iterative preparing calculation, which is utilized to limit a blunder work. As indicated by the type of the blunder work, SVM models can be characterized into four unmistakable gatherings:

*1)* Classification SVM Sort 1 (otherwise called C-SVM characterization)
*2)* Classification SVM Sort 2 (otherwise called nu-SVM characterization)
*3)* Regression SVM Sort 1 (otherwise called epsilon-SVM relapse)

## II. PROPSED METHODLOLGY

For P2P development unmistakable confirmation issue we proposed a joined approach using unsupervised machine learning figuring K infers batching for classes data in perspective of segments. Support vector machines (SVM) are a champion among the most by and large used machine learning procedures for plan and backslide issues of little examples. Frankly, the execution of SVM is, as it were, liable to its parameters assurance. In the technique of collection by SVM, and high estimations, this has a broad assortment of employments, for instance, picture arrange, stand up to revelation, content course of action. SVM has a splendid ability to handle the request issues for 2 classes. The rule explanation behind P2P action conspicuous evidence is to decisively arrange two characterizations: P2P and non-P2P development. Therefore K-Means and SVM both will give more exact result. Inherited count is a kind of reference natural regular decision and ordinary genetic arrangement of the unpredictable chase computation; it is sensible for dealing with complex issues which are hard to tackle by customary inquiry calculations. GA begins from the underlying irregular arrangement of arbitrary era; it creates new arrangements by a specific determination, hybrid and change operation well-ordered emphasis.



©IJRASET: All Rights are Reserved

Volume 5 Issue V, May 2017
ISSN: 2321-9653

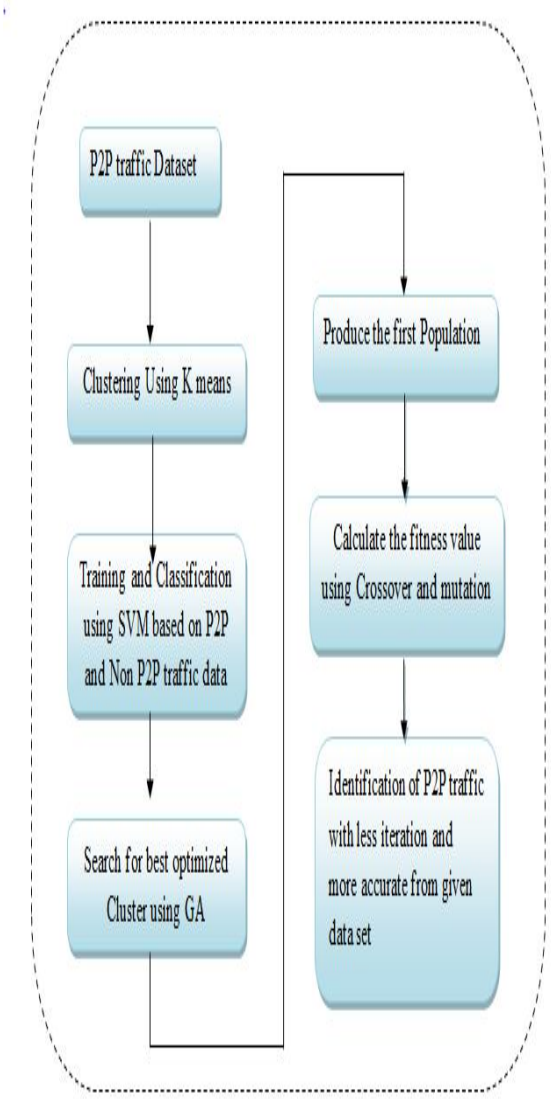




Fig: 1.2 Flowchart of Proposed Work

## IV. RESULT ANALYSIS / IMPLEMENTATION

The following Figure shows the response of energy consumption vs. transmission power traffic scenarios,

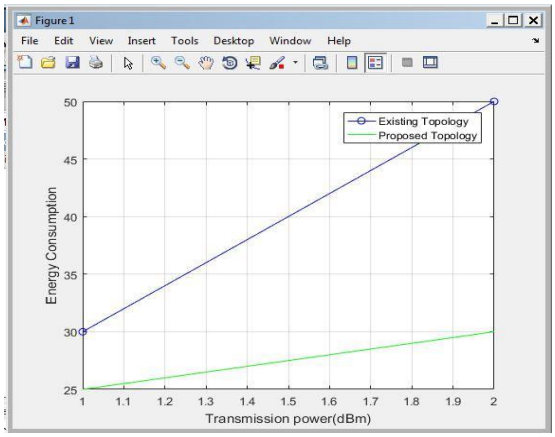



Fig: 1.3 Graph of energy consumption vs. transmission power



©IJRASET: All Rights are Reserved

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

*A. Network Traffic Information*

We will give the preparation and stacking of datasets by applying characterization utilizing support vector machine (SVM) method. The following are the perceptions for stacking the informational index and preparing the informational collection in MATLAB apparatus.

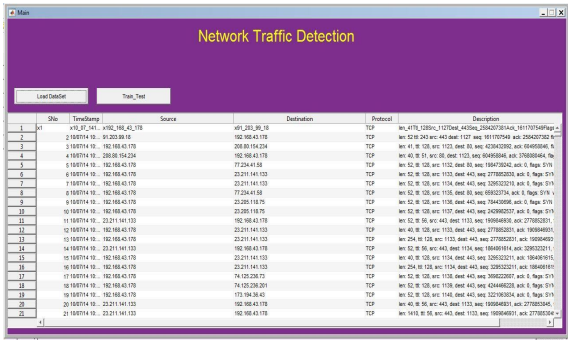



Fig: 1.4 shows the loading of dataset provided by SVM.

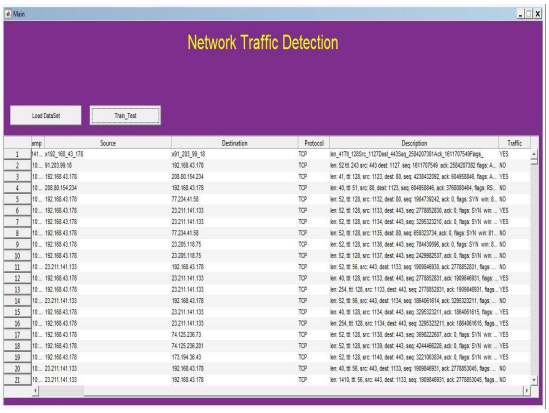



Fig: 1.5 Show the training of dataset provided by SVM.

The observations obtained by implementing simulation model for the traffic scenarios is provided in Table 1.1.The results are based on these observations.

Table 1.1: Observations for Varying Number of Node

| Methodology | Output |
|---|---|
| Existing Methodology (Genetic Algorithm) | 82.70 % |
| Proposed Methodology (K-Means, SVM and Genetic Algorithm) | 87.37 % |

## V. CONCLUSION

The conclusions introduced in this exposition of system activity distinguishing proof gives indispensable favourable circumstances to IP arrange building, organization and control and other key spaces. Current acclaimed strategies, for instance, port-based and payload-based, have exhibited a couple inconveniences, and the machine learning based procedure is a potential one. The activity is requested by the payload-self-governing truthful characters. This paper exhibits the differing levels in system activity examination and the huge data in machine learning space, looking at the issues of port-based and payload-based techniques in movement portrayal. Considering the need of the machine learning-based system, we attempt diverse things with K-means, SVM and GA to survey the productivity and execution. The trial happens on activity datasets pass on that the precision gained by our technique is


©IJRASET: All Rights are Reserved

progressed.

In this manner, all in all, the execution of P2P system activity is enhanced in productive way and with more precise outcomes.

## REFRENCES

[1] Jie Cao, Zhiyi Fang, Dan Zhang, and Guannan Qu, "Network Traffic Classification Using Feature Selection and Parameter Optimization", Journal of Communications Vol. 10, No. 10, October 2015, doi:10.12720/jcm.v.n.p-p doi:10.12720/jcm.10.10.828-835.

[2] Prof S. R. Patil, Suraj Sanjay Dangat, "Identifying Peer-to-Peer Traffic Based on Traffic Characteristics", Recent Advances in Computer Science, ISBN: 978-1-61804-320-7.

[3] Satoshi Ohzahata, Yoichi Hagiwara, Matsuaki Terada, and Konosuke Kawashima, "A Traffic Identification Method and Evaluations for a Pure P2P Application"

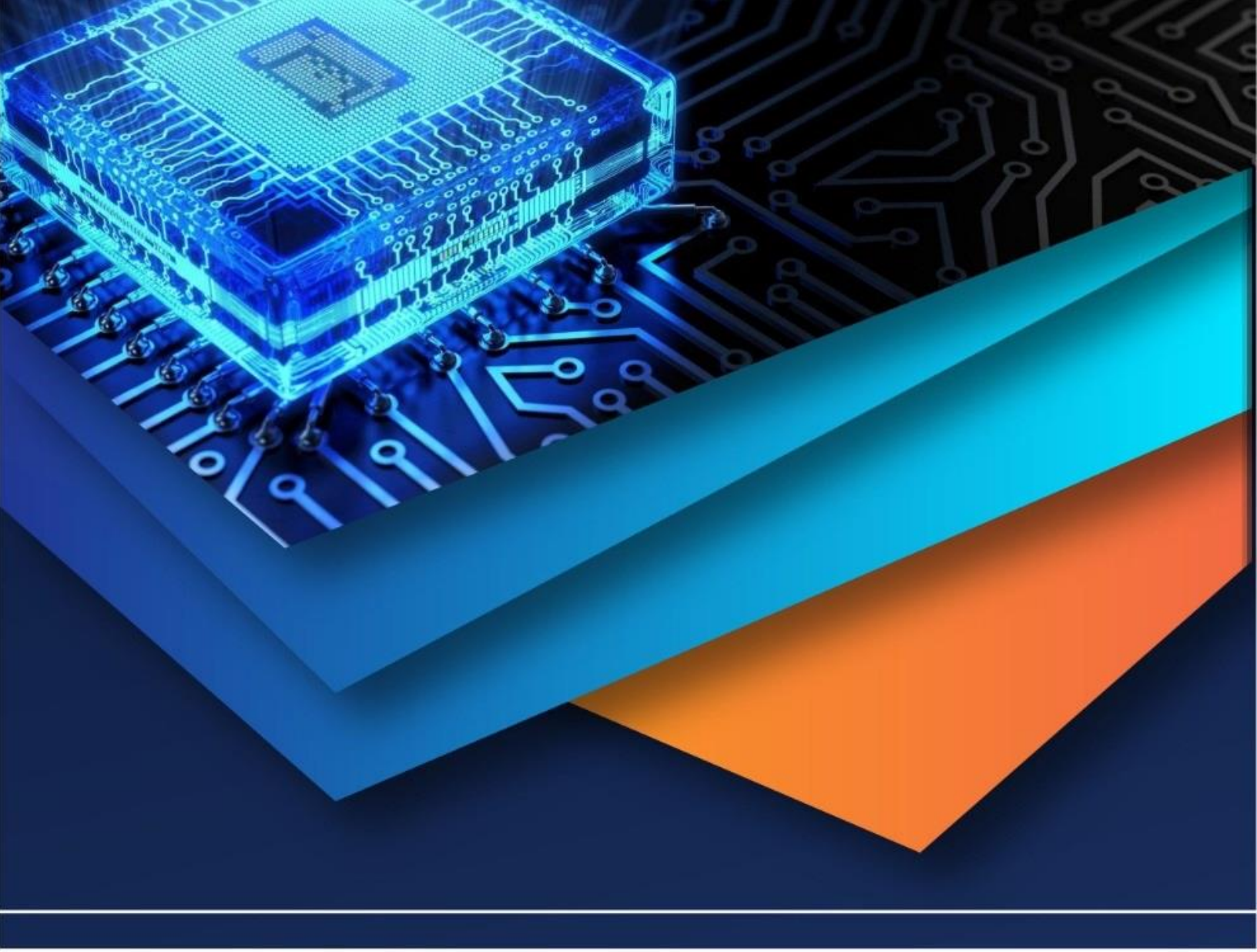
[4] Joseph Stephen Bassi, Loo Hui Ru, Khammas, Muhammad, Nadzir Marsono, "Online Peer-To-Peer Traffic Identification Based On Complex Events Processing Of Traffic Event Signatures", Jurnal Teknologi (Sciences & Engineering) 78:7 (2016) 9–16, eISSN 2180–3722.

[5] Jinghua Yan, Zhigang Wu, Hao Luo, Shuzhuang Zhang, "P2P Traffic Identification Based on Host and Flow Behaviour Characteristics", CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume 13, No 3, Sofia • 2013, ISSN: 1314-4081, DOI: 10.2478/cait-2013-0026.

[6] Marcell Perényi, Trang Dinh Dang, András Gefferth and Sándor Molnár, "Identification and Analysis of Peer-to-Peer Traffic", Proceedings of 5th International IFIP-TC6 Networking Conference, Coimbra, Portugal, May, 2006. © 2006 IFIP.



©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)