JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Optimization of Threats in P2P Networks by Intelligent Data Mining

[1]Chinu Mog Choudhari, [2]Sudeshna Das

[1,2]Assistant Professor, Department of Computer Science and Engineering, TIT, Narsingarh, Tripura, India

***Abstract*— A number of recent structured Peer-to-Peer (P2P) systems are built on top of distributed hash table (DHT) based P2P overlay networks. Almost all DHT-based schemes employ a tight-deterministic data placement and ID mapping schemes. This feature on one hand provides assurance on location of data if it exists, within a bounded number of hops, and on the other hand, opens doors for malicious nodes to lodge attacks that can potentially thwart the functionality of the P2P system. This paper studies several serious security threats in DHT-based P2P systems through three targeted attacks at the P2P protocol layer. The first attack explores the routing anomalies that can be caused by malicious nodes returning incorrect lookup routes. The second attack targets the tight data placement scheme. We show that replication of data items, by itself, is insufficient to secure the data items. The third attack targets the ID mapping scheme. We disclose that the malicious nodes can target any specific data item in the system; and corrupt/modify the data item to its favor. For each of these attacks, we provide quantitative analysis to estimate the ex-tent of damage that can be caused by the attack; followed by an experimental validation and defenses to guard the DHT-based P2P systems and counteract such attacks.**

***Index Terms*— Mining; P2P network; DHT; Threat.**

I. INTRODUCTION

The advent of peer-to-peer (P2P) file sharing systems heralds a new era in the field of Internet technology. While these systems alleviate the scalability problem that has dogged the client-server model, they present new data management problems. It is widely believed that the success of P2P file sharing systems depends upon the quality of service offered by such systems. Accordingly most of the present research in P2P systems has been concentrated on issues such as efficient data placement, fast file lookup, datareplication etc.

We argue that, in addition to the quality of service, there is another key aspect that impacts the success and continued sustenance of P2P systems. It is the quality of the datapresent in the system. For a file sharing system, no matter how excellent the lookup capabilities of a system are, or what file download speeds it offers, if the system does not have a large and growing number of interesting files, it will eventually fail to attract or retain users. Unfortunately, research on developing mechanisms to maintain or enhance the quality of data is yet to receive much attention from the P2P research community. This problem is exemplified by the phenomenon of free riding in many P2P file sharing systems.

The free riding problem affects the system in two significant ways. First, the number of files in the system becomes limited or grows very slowly. The number of popular files may become even smaller as the time goes by. This adversely affects user's interest in the system and they eventually pull out of the system. When users who share popular files pull out of the system, the system becomes poorer in terms of the amount of files shared. This is a unproductive cycle and it may eventually lead to the collapse of the system. Second, if only a few peers share popular files, all the downloading requests are directed towards those peers. This causes those peers to become hot spots, overloading their machines and causing congestion on their network. Peers frequently experiencing CPU overloads or network congestion due to the P2P system may exit the system if it affects their other routine activities.

In order to maintain the productivity and ensure the healthiness of a P2P file sharing system, there is a need for mechanisms that can help in securing cooperation from its users by encouraging them to share popular files. Surprisingly, none of the existing P2P files sharing systems, to our knowledge, offer or incorporate mechanisms that effectively encourage their users to share files of interest with other users in the system.

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical communitysometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new

treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together. Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor. In the public sector, data mining applications were initially used as a means to detect fraud and waste, but they have grown also to be used for purposes such as measuring and improving program performance. It has been reported that data mining has helped the federal government recover millions of dollars in fraudulent Medicare payments. The Justice Department has been able to use data mining to assess crime patterns and adjust resource allotments accordingly. Similarly, the Department of Veterans Affairs has used data mining to help predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs.

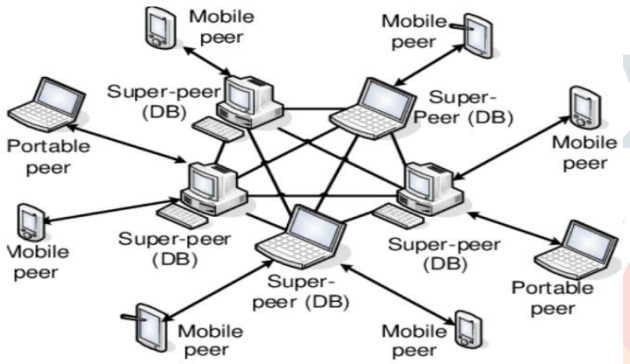

Fig.1 P2P Data Management Network

Another example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend precautionary measures. Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records.

In this paper, we explore a new approach that integrates Data Mining with P2Psystem which attempts to discover and extract new knowledge from the recorded data and information. This data is normally stored in databases, and can be of different nature such as peer id and reputation of the peer. The knowledge learned is represented in forms of rules, such as classification rules, prediction rules, association rules or clusters of rules. These results can be often used for identifying the peer behavior.

II. Related Work

[1] explains the details about various problems in P2Pnetworks and how they can be solved using reputation concepts. [4] Mainly concentrates on the various issues concerning data mining and gives us the formulas to identify the data mining based on the popularity, size and number of files shared by a peer. [2] and [3] give a broad outlook of a distributed way of identifying and isolating the data minings in the P2Psystem.

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing. A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the pre-computation and fast accessing of summarized data. Fig.1 gives a basic description of how a P2P Data Management Network looks like.

III. TARGETED ATTACKS AND DEFENSE MECHANISMS

Under the adversary model discussed above, a collection of malicious nodes can perform the following targeted attacks: Attack on the Routing Scheme (Routing Anomalies): The malicious nodes can lie about the next hop when asked about the node id of a node responsible for a particular data item and return incorrect lookup results, thereby, increasing the probability of lookup failures or dramatically increasing the cost of a lookup operation. We identify the key properties of the lookup protocol that determine the extent of damage caused by such an attack. Attack on the Data Placement Scheme (Information Decay): The malicious nodes corrupt the information stored in the system by repeatedly joining and leaving the network and every time corrupting the data they are assigned responsibility for. We show that even when replication with majority voting scheme are used, the information stored in the P2Pnetwork decays in course of time.

***Attack on the ID Mapping Scheme*:** The malicious nodes plant an attack on a specific data item stored in the P2P network. We show that such an attack is very powerful, though it is quite expensive for the malicious nodes to execute such an attack.

***Attack on the Routing Scheme*:** Routing Anomalies A typical DHT-based P2P system constructs an overlay topology in which every node plays the role of a client, a server, a router, and a domain name server. When nodes act as domain name servers - translating an identifier to the IP-address of a node that is responsible for it (see Property P4in Section 2), the malicious nodes can potentially exploit this feature to misguide legitimate nodes with incorrect lookups. This could result in denial of information - a legitimate node is denied access to a data item; or sub-optimal performance of the lookup algorithm. For example, a malicious node can lie about the next hop when asked about the node id of anode responsible for a particular data item and return incorrect lookup results. In the absence of no alternate paths any malicious node along the only route can block all requests (misguide), thereby, increasing the probability of lookup failures. In case there are alternative routes, this attack could delay the routing of requests to correct nodes, thereby dramatically increasing the cost of a lookup operation.

IV. DATA MINING IN P2P

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and

predictive data mining tasks that attempt to do predictions based on inference on available data. Data characterization is a summarization of general features of objects in a target class, and produces what is called *characteristic rules*. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, one may want to characterize the peers who play with him or the strategies that are used by a peer regularly. Data mining and malicious peer can be easily found using characterization. With concept hierarchies on the attributes describing the target class, the *attribute-oriented induction* method can be used, for example, to carry out data summarization. Note that with a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization.

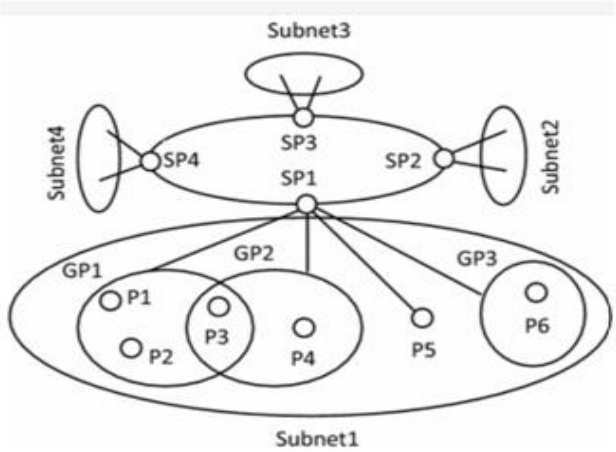

Fig. 2. Super warehouse in P2P

A. Classification

Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, we might want to classify our agents into 'Good' or 'Malicious' or 'Free riding' categories with regard to their reputation. The category or 'class' into which each peer is placed is the 'outcome' of classification process.

Case based reasoning is an apt classifier for P2P network, which uses the previous history for the process of classification. This method is more efficient because of the possibility of unsupervised classification. To solve a current classification problem (to find a data mining), the problem is matched against the cases in the case base, and similar cases are retrieved.

B. Prediction

Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values. For example if the peer is interested in predicting the chance of downloading without any interruption, he might simulate the whole transaction exactly with the history of the opponent. [5] Has effectively used the prediction technique to identify a winning strategy using data mining. We can use the same [6] Game theoretic technique to identify the malicious peer using data mining and Nash equilibrium concept.

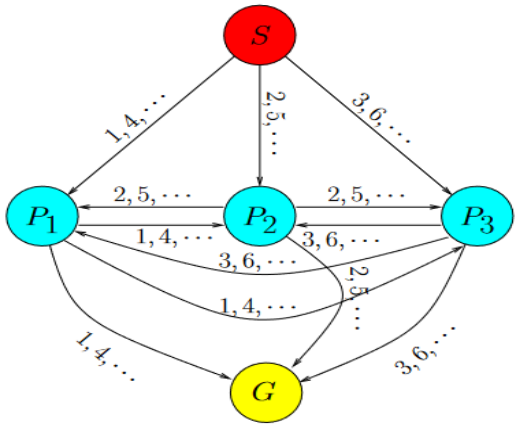

Fig.3 Peer Cluster

C. Clustering

Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called *unsupervised classification*, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (*intra-class similarity*) and minimizing the similarity between objects of different classes (*inter-class similarity*). For example if system is interested in finding the type of peer available into malicious and data mining, the system can use the history of each peer to form a cluster. Then any clustering time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. For example a peer can find the change in the strategy of each peer with respect to time and how the reputation gets changed. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

V. RESULT

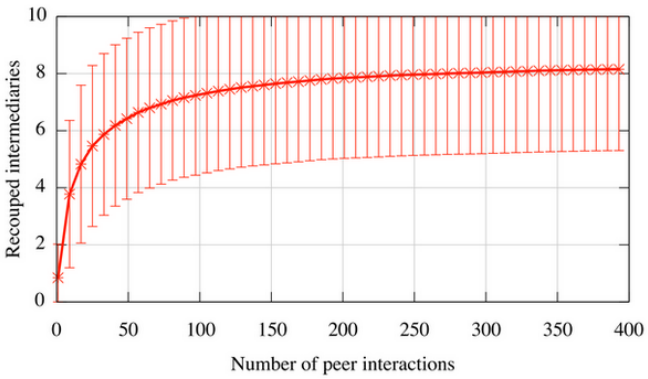

Fig.4 Peer hybrid interactions graph.

Outliers are data elements that cannot be grouped in a given class or cluster. Also known as *exceptions* or *surprises*, they are often very important to identify. While outliers can experiment on an

existing P2Pnetwork was analyzed. Java environment was used to develop the P2Pstructure. Initially the experiment was done with 20 nodes, and then the nodes were scaled up to 100 nodes.

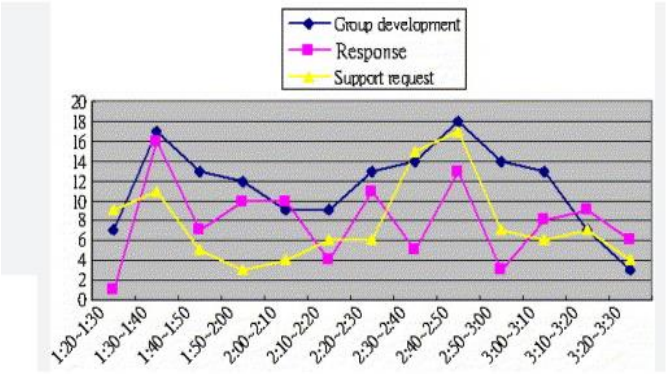

Fig.5. Response support interaction graph

Existence of behavior analyzer increased the interaction with good peer, indirectly isolating malicious and data mining in a P2Psystem issue. The query given by a good peer will be given higher preference, which indirectly suppresses the data mining action. And the query replies from a good peer will be given more weight, which indirectly reduces the malicious peer's action. So our concept of integrating data mining with P2Pnetwork greatly increases the lifetime of the p2p system by direct and indirectaction of eliminating data mining.

VI. CONCLUSION

While there are several ongoing research projects on improving the quality of service in P2P file sharing systems, there hasn't been much research to counter the problem of free riding effectively, which is essentially a data quality issue. To address the free riding problem in P2P systems, we have introduced this concept to measure the usefulness of every user to the system. We have proposed a free riding control scheme based on the general data mining functionalities. We expect that this paper to trigger further research in this area of P2P systems.

REFERENCES

[1] S.D. Kamvar et al. "Incentives for Combating Free riding on P2PNetworks". Euro-Par 2013 Parallel Processing.

[2] Lamport, R. Shostak, and M. Pease. The byzantine generals
problem. In IEEE Computer Society Press, 2012.

[3] Margaret H.Dunham, "Data Mining: Introductory and AdvancedTopics", Pearson Education 2014.

[4] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In Proceedings of SIGCOMM Annual Conference on Data Communication, August 2011

[5] M. Faloutsos, P. Faloutsos, C. Faloutsos, On Power-Law Relationships of the Internet Topology, SIGCOMM 1209.

[6] S.D. Kamvar et al. "Incentives for Combating Free riding on P2P Networks". Proceedings of EURO-PAR 2013

[7] Kubiatowics, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, W. Weimer, C. Wells, and B. Zhao. Oceanstore: An architecture for global-scale persistent storage. In Proceedings of the 9th International Conference on Architectural Support for Programming Languages and Operating Systems, November 2019

[8] Lakshmish Ramaswamy et al. "Free Riding: A New Challenge to Peer-to-Peer File Sharing Systems". Proceedings of Hawaii International Conference on Systems Science 36 2018

[9] Yi Wang. "Integration of Data Mining With Game Theory". International Federation for Information Processing (IFIP), Volume 207, 2020.

[10] Rohit Gupta and Arun K. Somani. "Game Theory As A Tool To Strategize As Well As Predict Nodes' Behavior In Peer-to-Peer Networks". Proceedings of the 11th International Conference on Parallel and Distributed Systems (ICPADS'05).

[11] H. Garcia Molina, S.D. Kamvar, A. Schossler "The EigenTrust Algorithm for Reputation Management in P2P networks", Technical report, Stanford University, 2020.

[12] J. Han, M. Kamber, "Data Mining: Concepts and Techniques",Harcourt India / Morgan Kauffman, 2019.

[13] Karakaya, M., Korpeoglu, I., Ulusoy, O. "A Distributed andMeasurement-based Framework Against FreeRiding in Peer-to-Peer Networks". proceedings of the 4th IEEE International Conference on Peer-to-Peer computing (P2P'03). Zurich, Switzerland, September, 2020.