GAUR: a method to detect Sybil groups in peer-to-peer overlays

K. Haribabu*

Department of Computer Science and Information Systems,
BITS Pilani,
Pilani 333031, Rajasthan, India
Email: khari@bits-pilani.ac.in
*Corresponding author

Chittaranjan Hota

Department of Computer Science and Information Systems,
BITS Pilani, Hyderabad Campus,
Hyderabad 500078, India
Email: hota@bits-hyderabad.ac.in

Arindam Paul

Department of Computer Science and Information Systems,
BITS Pilani,
Pilani 333031, Rajasthan, India
Email: arindampaul.bits@gmail.com

Abstract: Sybil attack is an important problem as the peer-to-peer networks grow in size and become prominent means for distributing multimedia. In order to validate the feasibility of using psychometric tests as an approach to detect Sybils in their entirety as Sybil groups, experiment is conducted by taking the tests on a population of considerable size. Selected people were given multiple questionnaires corresponding to multiple identities in the network. The survey data is analysed using DBSCAN clustering algorithm using several metrics. The results show that 75% of the Sybil groups were detected with 67% completeness.

Keywords: peer-to-peer; security; Sybil attack; psychometric tests; Luscher Colour test; MBTI; Clustering.

Reference to this paper should be made as follows: Haribabu, K., Hota, C. and Paul, A. (2012) 'GAUR: a method to detect Sybil groups in peer-to-peer overlays', *Int. J. Grid and Utility Computing*, Vol. 3, Nos. 2/3, pp.145–156.

Biographical notes: K. Haribabu is a Lecturer in the Department of Computer Science at Birla Institute of Technology & Science, Pilani, Rajasthan, India. He holds a Masters of Engineering degree in Software Systems. He is currently pursuing his PhD degree. His research interests are in the areas of search and security of peer-to-peer overlays.

Chittaranjan Hota is Associate Professor and Head of the Computer Science & Information Systems Department at Birla Institute of Technology & Science, Pilani, Hyderabad Campus, Hyderabad. He holds a Bachelors of Engineering degree in Computer Science, Masters of Engineering in Computer Science, and PhD in Computer Science & Engineering. He had worked as researcher and visiting professor at University of New South Wales, Sydney, Australia; Helsinki University of Technology, Helsinki; City University, London; and University of Cagliari, Italy. His research interests are in the areas of QoS over the internet, peer-to-peer overlays, MANETs and cloud computing.

Arindam Paul is working as a Project Assistant in the Information Processing Center at BITS Pilani. He is presently a graduate student in the Computer Science and Information Systems Department at BITS, Pilani. His research areas are map reduce, P2P systems, cluster computing, distributed systems, and software engineering.

1 Introduction

The peer-to-peer (P2P) paradigm had started becoming popular in the middle of 2000 among the music lovers. Since then, due to its inherent positive characteristics, the term 'P2P' has become very popular amongst the internet users, researchers and industries. The emergence of P2P overlay file sharing networks has increased the interest amongst the internet users to use the internet beyond web browsing and exchanging e-mails. There is large number of applications such as Napster (Napster, 1999), Gnutella (Gnutella, 2000), Kazaa (Kazaa, 2001), eDonkey (eDonkey, 2000) etc. developed for deploying P2P technologies in the internet. There is an exponential increase of users taking interest in these applications. Researchers have contributed large number of models based on P2P overlay networks, suiting to different requirements such as data sharing, distributed file systems, anonymity, media streaming etc. Owing to their ease of deployment and scalability, industries have formed the consortium called the 'P2P-Next Generation' (P2P-Next, 2008) to support research to effectively use the P2P technology for mass media distribution.

RFC 5694 (Camarillo, 2009) gives an official definition of a peer-to-peer system. A system is to be considered P2P if the elements that form the system share their resources in order to provide the designated services. The elements in the system provide both client and server services, i.e. providing services to other elements as well as request services from other elements. An overlay network is a logical (virtual) network at the application layer providing connectivity, routing and messaging amongst the addressable end points (Buford et al., 2009). They have their own topology different from underlying physical network. They have their way of routing messages with the help of the internet and addressing the end points. Overlay networks are frequently used as a substrate for deploying new network services, or for providing a routing topology not available from the underlying physical network. P2P overlay networks are categorised as unstructured and structured. An unstructured P2P system is composed of peers joining the network with some loose rules, without any prior knowledge of the topology. Gnutella and Kazaa are examples of unstructured P2P overlay networks. In structured P2P overlay networks, network topology is tightly controlled and content is placed not at random peers but at specified locations that will make subsequent queries more efficient. Most of the structured P2P overlays are Distributed Hash Table (DHT) based. Content Addressable Network (CAN) (Ratnasamy et al., 2001), Chord (Stoica et al., 2001), and Pastry (Rowstron and Druschel, 2000) are some examples of structured P2P overlay networks.

1.1 Security in peer-to-peer overlays

Security is the fundamental issue to be addressed when the system involves multiple users and their shared resources. Large-scale peer-to-peer overlays involve millions of user identities and their devices contributing to the functioning of the network. Authentication, integrity, confidentiality and non-repudiation are some of the security properties expected to be supported by the system. These issues become significantly more challenging than in the case of traditional domains due to distributive ownership, lack of centralised control and lack of global knowledge in large-scale peer-to-peer overlays. Therefore peer-to-peer overlays face additional security risks when compared to security issues in network applications. Peer-to-peer overlays introduce an additional layer called 'overlay layer' which involves specific security risks which are not common in the internet applications. Attacks on the overlay can be divided into Message routing attacks, Sybil attacks, and Eclipse attacks. Message routing attacks work by modifying the node's routing tables. Routing tables are key resource for the stability of structured networks. Castro et al. (2003) noted that an attacker can obtain specific node IDs and strategically position itself in the overlay in such a way that it controls the access of specific peers or objects. Also poisoning of routing tables and message forwarding attacks are possible (Wallach, 2002; Castro et al., 2003). In Eclipse attacks, one first gains control over large number of nodes along strategic routing paths and then separate the network into different sub-networks. Traffic between the sub-networks has to go through one of the attacker's node. This way the attacker disrupts the network in a systematic way to propagate false files in a fast paced way. Eclipse attack is possible when the identities of the network are already in control which is achieved through Sybil attack.

In Sybil attack, an entity can represent itself as multiple identities in the overlay and thus gain control over disproportionate resources. This attack was first pointed out by Douceur (2002). In this attack, an attacker influences the reputation of the system and objects and also carries out malicious attacks like disrupting the overlay operations. Some entities forge themselves as multiple identities in the network. Because of their large fraction of identities, the entities can control the network. It is very difficult to differentiate between a real identity and a Sybil identity. It is stated by Douceur (2002) that without a centralised authority it is not possible to completely eliminate the Sybil identities from the network. Since peer-to-peer overlays fit decentralised mechanisms, in this paper, we have put effort to develop algorithm to limit Sybils in a distributed way.

1.2 Sybil attack

Sybil attack is an attack where an entity in a peer-to-peer network can masquerade itself as multiple simultaneous identities in the network (Douceur, 2002). A peer-to-peer overlay file sharing network consists of set E of infrastructural entities e. An identity is an abstract representation that persists across multiple communication events. Each entity e attempts to present an identity i to other entities in the system. Each correct entity e will attempt to present one legitimate identity. Each faulty entity may attempt to present a legitimate identity and one or more counterfeit identities. Ideally, the system should accept all legitimate identities but no counterfeit entities.

The problem with such duplicitous mapping of many virtual identities on to one entity is the collective influence a single user can exert on the decisions and working of the entire network if the multiple identities created by the user form a significant fraction of the peer-to-peer network. This problem is pervasive in all distributed systems. This attack is possible in any distributed network but peer-to-peer network is an attractive field for this attack due to its lack of central control, and large size. Peer-to-peer networks have huge resources like processing power, bandwidth and storage contributed by the participants. In mobile ad-hoc networks and sensor networks, the nodes are constrained by their physical characteristics (Dinger and Hartenstein, 2006). Since peer-to-peer network is built at application layer the physical constraints don't limit the Sybil attacks. It is not very difficult to set up a Sybil attack because creating an identity in the network is as simple as starting another instance of the peer-to-peer client. If a malicious user has a vast pool of resources at his disposal, he can create large number of identities.

There are several instances mentioned in the literature about the instances of using this attack for selfish purposes. Peer-to-peer networks due to their large size become difficult to assess trustworthiness of a peer whom we interact with. Therefore reputation systems are used to aggregate the collective experiences of peers about other peers (Resnick et al., 2000). When a peer needs to interact with another with whom it has not interacted so far, the reputation system helps in making opinion about that peer. In online systems like Amazon, eBay etc. reputation systems are used to aggregate the ratings of sellers and goods. Such ratings have impact on business transactions (Craig et al., 2010). In file-sharing overlays, reputation of files affect users opinion whether to view that file or not. Such benefits attract malicious attempts to manipulate the reputation systems.

Gyongyi and Garcia-Molina (2005) have reported that the web page rankings can be manipulated by setting up a link farm. Bhattacharjee and Goel (2005) have reported similar instances of manipulations using Sybil attacks. Sybil attack is commonly used to fool Google's PageRank algorithm (Bianchini et al., 2005). PageRank algorithm is one of the most commonly used algorithms to compute the reputation of peers in reputation systems (Kamvar et al., 2003). The major problem in peer-to-peer computational systems such as SETI@home is that, servers should ensure that the clients are not cheating by submitting deceptive results without fully performing all the computations specified. One way to detect this cheating is to allocate the same task to multiple clients. But this redundancy can be subverted if there is an agreement among the clients that they would return the same manipulated-result. In the internet it is possible that all these clients can be instances of the same devious entity who can synchronise the outputs of all the clients and thus mislead the server (Yurkewych et al., 2005). Sybil attacks create false routes in mobile ad hoc networks (Hu et al., 2002). Sybil attacks can disturb anonymous systems such as Tor by revealing user identities of anonymous routing protocols (Dingledine et al., 2004). Pastiche is a file storage system built on Pastry overlay. Sybil attacks can subvert the distributed quotas by free-riding cooperative file storage systems (Cox et al., 2002).

1.3 Purpose

In this paper we develop a novel approach for detecting Sybil groups. We have used psychometric techniques to assess the characteristics of participating identities in the peer-to-peer network. The first purpose of this work is to study the feasibility of using psychometric tests to assess the characteristics of the participants. The second purpose is to devise methods to overcome some of the limitations of this method. The third purpose of this work is to measure the extent of the effectiveness with which we can use this technique.

The rest of the paper is organised into sections as Literature survey, Background, GAUR, Experiment, Results analysis, Conclusion and References.

2 Literature survey

Douceur (2002) proved that it is not possible to completely eliminate the Sybils in a peer-to-peer network without a centralised authority which can verify the one-to-one correspondence between identities and entities. He described puzzle methods that exploit communication, storage or computational resource constraints. He proved that computational puzzle methods are not viable. In these puzzles, the verifier sends a large random value to every other identity it wants to verify. These identities must then compute the solution within a constrained amount of time. If an entity has more than one identity it will fail to compute the solution within this time. The paper says that this can be circumvented by taking help of other powerful nodes. Thus he advocates the existence of a central authority to prevent Sybil attacks. Solutions to Sybil attack can be categorised as challenge-response imposing constraints on resources, binding the identity to physical characteristics, central authority certification, characteristics of social networks based on trusted connections, based on Sybil behavioural characteristics and incentives.

2.1 Challenge-response approaches

The goal of resource testing is to attempt to determine if a number of identities possess fewer resources than would be expected if they were independent. Challenge-response utilises puzzle methods that exploit communication, storage or computational resource constraints of the participating nodes. In these puzzles, the verifier sends a large random value to every other identity it wants to verify. These identities must then compute the solution within a constrained amount of time. If an entity has more than one identity it will fail to compute the solution within this time. These tests include checks for computing ability, storage ability, and network bandwidth, as well as limited IP addresses (Levine et al., 2006). Douceur (2002) says that this can be circumvented by taking help of other powerful nodes and therefore, advocates the existence of a central authority to prevent Sybil attacks.

Borisov (2006) proposes to use computational puzzles to defend Chord from Sybil attacks. In Chord, every node sends periodic ping messages to its neighbours. This scheme proposes that along with every ping message, a sequence number and a challenge will be sent to neighbour. The puzzles are formed out of these challenges and sequence numbers. Even honest nodes are expected to solve these puzzles. Here the heterogeneity of the nodes in their computation capacity is not addressed. Rowaihy et al. (2007) present a hierarchical admission control system where at every level computation puzzle are used to validate the identity. It requires a joining node to solve the puzzles from leaf to root in tree of trusted nodes. This scheme only slows down the identity generation but not prevent the Sybil attack. Here the honest nodes are also subjected to the same tests.

Haribabu et al. (2009) proposed a challenge-response approach based on storage constraint. Sybil attack can completely subvert replication mechanism on file sharing systems. By knowing the mechanism of replication which is used in a particular P2P network, a malicious user (Sybil attacker) can create fake identities in the network so that the file replication of a particular file happens entirely or partially on the Sybil identities created by this particular user. Once the replica is in the hands of Sybil identity, it can corrupt, hide or destroy the copy especially if all copies are replicated on Sybil identities only. Sybil attack goes against maintaining quality and accessibility of content, and robustness of the network. The authors propose that the owner of the replica maintain a copy on successors. Since each identity of a Sybil has to store as many replicas as any normal entity, it will throttle the storage capacity of the Sybil entity. The Sybils are detected by regularly verifying the file.

2.2 *Binding identity to network metrics*

Bazzi and Konjevod (2005) proposed that an identity can be mapped to its physical location. There are two types of nodes: applicants and beacon nodes. Geometric certificate contains the distances measured between the node and the beacon and signed by both. This approach introduces a equivalence relation where all nodes in one relation can't be distinguished from others in the same relation. Here the defect is that if the Sybil is controlling entities in different relations then it is not possible to detect it. Also the algorithms to measure distance don't give stable values and requires considerable effort to achieve stable values. Bazzi et al. (2006) proposes a secure distance vector routing protocol that tolerates Sybil attack.

Wang et al. (2005) proposed a concept of net-print. The net-print of a node is built using node's default router IP address, its MAC address and a vector of RTT measurements from the node to designated land marks. This approach fails when the node changes its physical location. So this solution doesn't apply to mobile hosts.

Dinger and Hartenstein (2006) proposed a distributed registration mechanism for Chord. Each identity calculates its id as a hash of its IP address and port number and registers itself at *r* registration nodes in the Chord ring where *r* is system wide constant. This solution will work only if majority of the nodes are honest. Here the mapping is between identity and its IP address. The cardinality of this mapping is controlled through a distributed registration process. In this solution the influence of Sybil is limited to the number of IP addresses it can possess.

2.3 *Central authority certified node identities*

Castro et al. (2003) argue that the only practical solution to prevent Sybil identities in the peer-to-peer overlay network is to produce signed certificates that bind node identity to a public key and the IP address of the node. To prevent a malicious entity from obtaining a large number of certificates, the authors propose that each certificate can be issued against a charge. Certainly this solution prevents Sybils but it also slows down the propagation of the network services to new users. For IP based schemes special provisions have to be made for the nodes behind NAT based firewalls.

2.4 *Based on social network characteristics*

Danezis et al. (2005) presents a modified DHT routing model using a bootstrap tree for Chord to resist the impact of Sybil attacks. In the bootstrap tree, two nodes share an edge if one introduced the other into the DHT. These relationships are established outside the tree, off-line. With this logic Sybil nodes will attach to the tree at limited nodes. Also, a trust metric is used to minimise the probability of a malicious node being on the routing path of the Chord. It is not mentioned to what extent the logarithmic lookup times can be maintained with this approach. It increases the overhead in lookups. In Sybilguard (Yu et al., 2006), the authors have proposed a distributed algorithm to limit the entry of Sybil identities into a social network, exploiting the fact that there are very few trust edges between an honest and a Sybil group in a social network. They have designed a protocol in which the verification of a new entry into the network is done by intersection of random routes. The problem with these approaches is that they work only with networks that have evolved based on social trust relationships. This is not the case in a majority of the existing public peer-to-peer file sharing systems such as Gnutella, Freenet etc. Lesniewski-Laas and Kaashoek (2010) present Whanau, a one hop DHT based routing protocol which exploits social connections between users to construct routing tables which allow for Sybil resilient lookups. The file lookup algorithm suggested offers a significant speedup over the traditional flooding techniques seen in existing peer to peer networks. A major drawback of the approach is that it assumes that honest nodes have more social connections to other honest nodes rather than to Sybil nodes which may not always be the case. SybilInfer (Danezis and Mittal, 2009) offers a decentralised protocol to guard the network against Sybil attacks exploiting the fact that a Sybil attack would interfere with the fast mixing property of social networks. The approach entails a probabilistic model to help tag network nodes as either honest or Sybil wherein each such tag contains an assigned probability, referring to the degree of certainty of the result. The approach suffers from assuming that there is at least one honest node in the network which is known a priori whereas in reality there is always a remote possibility of an attacker mimicking the honest side of the social network as a consequence of which no detector would be able to distinguish the honest region from the corrupt one.

2.5 Based on Sybil behavioural aspects

Jyothi and Janakiram (2009) have proposed a Sybil monitor to observe and record the transactions of a node. Fake transactions are recorded by the monitor. The power of Sybil is in being a group. This solution looks at the individual nodes and their transactions. A transaction itself such as voting a file can't be named fake by looking at the individual node. Here the method serves its purpose only in the cases where the nodes behaviour can be termed bad by looking at its bad history. But in case of Sybils, the nodes need not necessarily be doing bad activities at the individual identity level.

Haribabu et al. (2010) have proposed that by collecting observations about the malicious activities of identities in the network, Sybil groups can be detected. Sybils as a group can collaborate to disconnect a part of the network from the rest, launch a denial of service attack over chosen nodes or a part of the network, or promote the propagation of an object. They attempt to identify Sybils based on the above characteristics. A neural network is trained to detect the malicious behaviour of a Sybil node and its detection improves with experience.

2.6 Incentives

Margolin and Levine (2007) analyse an economic approach to Sybil attack detection employing Dutch auction technique to determine the minimum possible reward to force the Sybil to reveal itself. This proposal requires implementation of digital currency. Also the model assumes that Sybil identities are rational.

The approach outlined in this paper is the expanded version of Haribabu et al. (2011). This approach using psychometric techniques to identify Sybil groups is different from above approaches in two ways. This approach is the first of its kind. And secondly it aims at finding the Sybil groups instead of aiming at detecting Sybil identities.

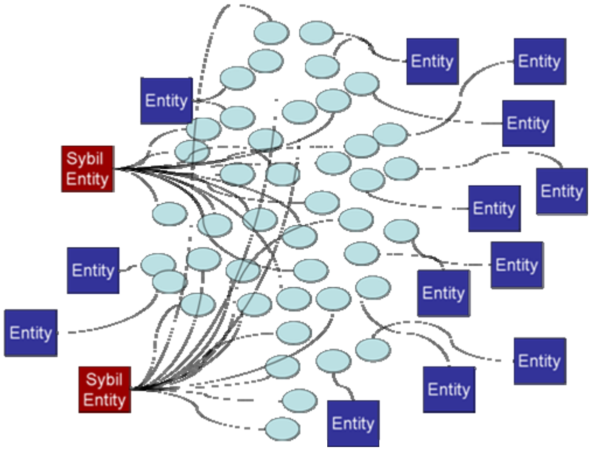
3 Background

In this section a brief background on the degree of difficulty of detecting Sybils in the network and applicability of psychometric techniques is presented.

3.1 Approaching Sybils

The strength of Sybil attack depends on the number of identities it creates and how much fraction of the network they occupy. The influence on the network is exerted as a group but not at the individual identity itself. Therefore in detecting a Sybil attack, it is very difficult to identify a Sybil at the identity level. In Figure 1, it can be observed that there are entities and identities (shaded circles). Some entities have represented themselves as multiple-identities forming large fraction of the network.

Figure 1 Sybil groups and Sybil identities (see online version for colours)



Having a centralised server issuing the logins or providing authentication itself is not sufficient to prevent Sybil attacks. Amazon, eBay etc. have centralised authentication systems but still they face Sybil attacks. The principle behind preventing Sybil attack is that one should be able to map the real infrastructure entity to the virtual identities and then put a limit on such number of mappings. Such a system requires one to authenticate the identity by a physical proof such as photo identity card, credit card etc. Such a restriction on enrolling new entities into the system severely limits the spread of systems among users.

Sybil entities need not necessarily be creating disturbance in the network such as launching distributed DoS attacks, or dropping the packets or poisoning the routing tables, partitioning the network etc. A Sybil entity may be doing the same thing like any other honest identity. For example, Sybil entities can position their identities strategically at different places in the network and make sure that every packet in the network will pass through at least one of its identity so that Sybil has control over the network. Looking at the identity, one can't say it is doing something malicious. It should be determined looking at the network level. Another example is that Sybil entity can increase the reputation of a particular file by making all its identities respond positively to the reputation metric. Looking at the individual identity it is difficult to say that what it is doing is wrong because every identity has freedom to respond positively or negatively to the file.

So the approach to detecting Sybils in a peer-to-peer network should be free from any assumptions about the behaviour at the identity level.

3.2 Psychometric techniques

There are many theoretical approaches to conceptualising and measuring personality. Some of them include the Minnesota Multiphasic Personality Inventory (MMPI), the Five-Factor

Model. There are some tools developed to measure personality. They include Personality and Preference Inventory (PAPI) and the Myers–Briggs Type Indicator (MBTI). PAPI measures personality in work environment. It is designed to determine behaviours and preferences which are related to workplace. The MBTI measurement is based on a psychometric questionnaire designed to measure psychological preferences in how people perceive the world and make decisions (Briggs and Myers, 1980). In this paper we use MBTI due to its applicability to normal population (Pearman and Albritton, 1997). MBTI test categorises the human personality into four pairs of cognitive functional types:

- Extraversion-Introversion
- Sensing-Intuition
- Thinking-Feeling
- Judgement-Perception

These terms have detailed meaning which are much different from what is indicated by the normal usage of that term. These meanings are not relevant in this study. MBTI states that an individual has preference to one of the functions in a pair. The questionnaires are designed to reveal these preferences.

Luscher (1969) proposed a colour test to assess the human personality. The test is based on selecting eight colours according to individual's liking for the colours. Each colour has a objective meaning and subjective meaning. The objective meaning remains the same for all individuals. But the subjective meaning of the colour is dependent on the position of the colour in the ranking of the colours by the individual. Liking for a colour has deep connection with the psychology of the person.

In this study, the objective is to group the identities based on their common personality characteristics. For this purpose MBTI and Luscher colour test are employed. Since the purpose is to find the similarity amongst the personalities of identities in the network, the interest is only in the metrics and not on the actual meaning of each metric. So we limit our discussions to finding correlations rather than actual personality of the individuals.

4 Proposed method: GAUR

In this section we will discuss about the proposed method GAUR.

4.1 Outline

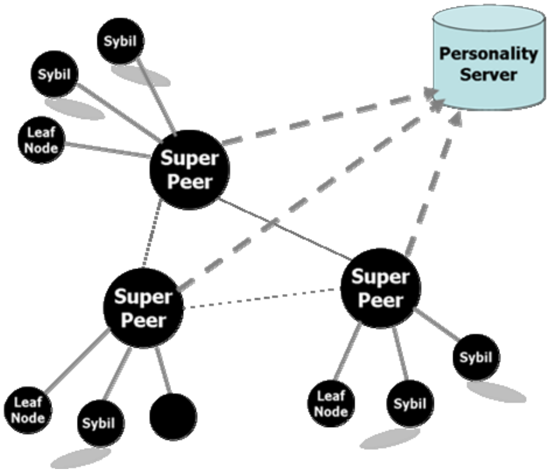
As discussed in Section 3.1, Sybils are present in groups. Extent of their influence is proportional to the group size. There can be several Sybil groups within a peer-to-peer network. In essence a Sybil group is defined as the set of identities created by a single individual. The identities act as per what the individual want them to act. If the identities occupy a large fraction of the network, then the functioning of the network can be easily affected for selfish purposes. The solution to this problem is to cluster identities which have similar personality characteristics. Such clusters are then tested using challenge-response protocols as described in Section 2. To identify personality characteristics, each identity in the network is given personality tests, namely Myers–Briggs Psychometric Test and Luscher Colour Test to answer. The response to the personality questionnaires are collected and analysed at one place to identify the clusters.

4.2 Architecture

We model our solution on a super-peer type unstructured network which employs a Gnutella like protocol between various client nodes. This allows for a questionnaire to be sent to the peers as a request and then, the solved questionnaire as a response.

The architecture is shown in Figure 2. The components involved are leaf nodes (some are Sybil identities), super-peers, and Personality Server. Personality Server is a central server that collects and stores the personality metrics of the leaf nodes. Having central server in a distributed network is not a flaw as long as that central server itself is not involved in routine network operations such as routing etc.

Figure 2 Architecture of GAUR (see online version for colours)



Some assumptions have been made in this model:

We assume that we can implement a strict protocol that all the peers in the network are compulsorily made to respond to the questionnaires, and also restricted from using the network if they do not respond to the questionnaire within a particular interval.

We also assume that the super-peers are not Sybil nodes, i.e. the Sybil peers exist only at the lowest level.

4.3 Protocol

The principle communications involved in the network apart from the routine communications are (1) Super peer → Leaf node (2) Leaf Node → Super peer (3) Super peer → Personality Server (4) Personality Server → Super peer. These communications and their contents are explained below.

- *Leaf node's Behaviour*: A leaf node receives a questionnaire from a super peer at regular times. The contents of the questionnaire are explained in the coming section. Super peer specifies the time within which it should receive the answers to the questionnaire. This is one way to differentiate between the honest and the Sybil entity. If the node doesn't answer within the expected time although it is still active in the network, it is given a warning that it might be labelled as a malicious node. In case, the node still does not answer even after some time interval after first warning, the node is then blacklisted or treated as a malicious node and taken out of the network. For the node to again come back, it has to join as a new node and follow the protocol.
- *Super-peer's behaviour*: Super-peer generates a questionnaire from few pools of questions each corresponding to one personality trait and sends it to the leaf nodes. It takes necessary action depending on whether a leaf node sends a response or not. It collects the responses from the leaf nodes and sends them to the Personality Server.
- *Personality Server*: The Personality Server receives the data from the super peers. It periodically runs a clustering algorithm to cluster the nodes based on their psychometric values. For every cluster it finds, it issues simultaneous computational puzzles to the nodes in that cluster. Generally the cluster contains at least some of the Sybil identities belonging to the same Sybil group. This fact is confirmed in our experiments as discussed in coming sections. In case of the presence of Sybil identities in a cluster, the Sybil entity will not be able to respond with answer to the resource-intensive-computational puzzle for each Sybil identity as an honest node can answer. By this differentiation between Sybil identities and honest nodes, the Personality Server can identify all these late-responding identities within the cluster as one Sybil group. These identities are communicated to super-peers so that they may restrict their cooperation or severe their connection.

4.4 Apparent limitations

- *False Positives:* An honest user's psychology may match with the psychology of a malicious user, thus falling in the same cluster as the malicious user. Here the honest user also has to go through the challenge-response test. But this is not a limitation but an overhead on the honest user.
- *False Negatives:* The test is based on psychological metrics. So it might happen that the psychometric ratings of some Sybil identities coming from the same Sybil group may not fall within the same cluster boundaries. This way some of the Sybil identities may escape getting detected. This introduces false negatives in our proposed model. First of all such deviations of psychological pattern happen in few cases as found in our experiments. Secondly, since majority of the Sybil identities of a group are detected, the group origins can be traced and thus detect the missed out Sybil identities.
- *Random Answering:* Another apparent limitation is random answering by some automated means or even personally. One way to address the automated answers is by inserting CAPTCHAs in the questionnaire. To address this issue, super-peer by random selection stores a copy of the questionnaire and the answers received. Super-peers use this to cross verify the answers received later. If the answers are random, then it is very less likely that answers would match.

4.5 Questionnaire preparation

The questionnaire will evaluate the psychometric index of the peers based on MBTI or Luscher Colour Test. The questionnaire consists of three kinds of questions. Firstly, there are questions based on each of the four categories of personality traits according to MBTI model which gives us the information about the psychological orientation of a person. The personality traits are extroversion-introversion, sensing-intuition, thinking-feeling, judging-perceiving. For example, we can ask a person whether he likes to hear less and talk more or he likes to hear more and talk less. In Tables 1–3, some of the questions which help us to know about each of the traits in a person are presented. The options to these questions are designed in a relatable way to one's work or private life.

Table 1 Sample questions for testing introvert quality

Quality	*Questions*
Extroversion-Introversion	Do you talk more than listen or vice-versa? Do you have high energy or quiet energy? Do you want to stay behind scenes or want a public role?

Table 2 Some questions for testing intuition quality

Quality	*Questions*
Sensing-Intuition	Do you focus on details or do you see the big picture? Do you work at a steady pace or bursts of energy? Do you trust gut instincts or actual experience?

Table 3 Some questions for testing feeling quality

Quality	*Questions*
Thinking-Feeling	Do you appear cool and reserved or warm and friendly? Do you value honesty and fairness more or harmony and compassion? Do you tend more to see faults or you are quick to compliment others?

Secondly, we have two sets of colour test (each colour tests have eight colours) in each questionnaire. The leaf nodes are required to fill in their preferences of colours from most preferred to least preferred. Each colour is associated with some particular trait (Luscher, 1969). These associations are findings of the research in psychology field. For example the orange-red colour represents ‘force of will’ and corresponds to desire, domination, aggression, controlled passion, concern for others.

Also, the questionnaire includes CAPTCHAS or simple mathematical questions such as what is the total if one takes two apples out of 20 apples inserted at random places to prevent automated answers.

4.6 *Questionnaire evaluation*

The answers submitted by the leaf peers to the super-peer are communicated to the Personality Server. The evaluation is carried out at the Personality Server. There are two types of answers in a questionnaire. One is the set of options selected for the MBTI questionnaire. The other is list of rankings given to the 8-set colours.

To evaluate the similarity between the colour rankings of two different identities, rank-correlation coefficients are used. Kendall’s rank-correlation coefficient τ and Spearman’s rank-correlation coefficient ρ are used to cluster the identities based on colour test. The τ is computed by finding the concordant (N_c) and discordant pairs (N_d) for n number of items to be ranked.

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)}$$

The ρ is computed by finding the squared differences of rankings given to all n items.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

The selected options in the MBTI questionnaire are transformed into a vector of weights. Each vector consists of 8 weights corresponding 8 qualities in four dichotomic cognitive functional pairs namely extraversion-introversion etc. There is one vector corresponding to one questionnaire. The similarity between two such vectors is found by using two metrics Cosine similarity and Pearson correlation coefficient.

The clustering algorithm chosen is DBSCAN (Ester et al., 1996) for it supports arbitrary shaped clusters, and leaves out noise points. Also cluster formation doesn’t depend on the ordering of the data and doesn’t need number of clusters to be specified before-hand. Also the algorithm is designed for working with high dimensional data. The algorithm finds the clusters on the principle of neighbourhood of a point. Generally the neighbourhood is found with Euclidean distance. But Euclidean distance becomes an ineffective measure when there is high dimensional data. This problem is also known as curse of dimensionality (Bellman, 1957). Therefore in this work, we have used the above mentioned metrics to find neighbourhood of a given point. The time-complexity of the algorithm is $O(n^2)$ and space-complexity is $O(n^2)$. By using R*-tree indexing structures the complexity can be brought down to $O(n \log n)$.

The algorithm takes two parameters ε and *minPoints*. ε is the neighbourhood radius. *minPoints* is the minimum number of points to form the cluster. The algorithm is very sensitive to these two parameters. It groups all the points that fall within the radius ε of the current point and its neighbours.

4.7 *Cluster validation*

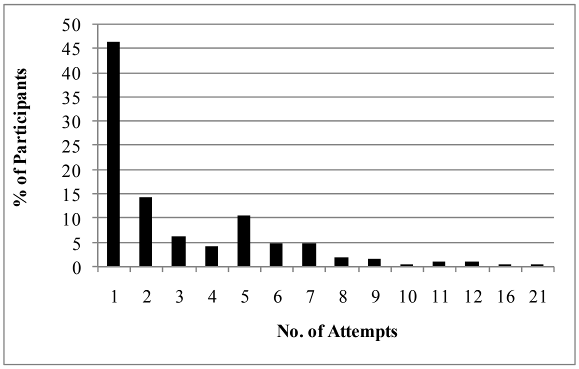
One of the most important issues in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data (Halkidi et al., 2001). What values for parameters ε and *minPoints*, the algorithm would give the best fit? To find out this, Rand Statistic (Rand, 1971), Jaccard Coefficient (Paul, 1901) and Fawlkes and Mallows index (Fowlkes and Mallows, 1983) are used to measure the cluster goodness. Rand Statistic gives ratio of true positive pairs (TP) and true negative pairs (TN) to total pairs in the data set. Jaccard coefficient gives ratio of true positive pairs to total of true positive, false positive (FP) and false negative pairs (FN). These two indices take values in the range [0,1]. Nearer the index to 1, closer are they.

5 Experiment set-up

The experiment for validating our solution is carried out by conducting a survey. Survey is chosen instead of simulation because the validation involves responses of humans depending on their psychological personality characteristics. These characteristics are difficult to simulate.

The survey is conducted among the faculty and students of our university for duration of 15 days through the medium of web with 185 people taking part in the survey. Each of them was requested to participate more than one time. The frequency of attempts is depicted in Figure 3. The participants are known to authors and are requested to attentively attempt the survey at different times so that they are not remembering the answers. Still there can be some degree of randomisation or overlaps in their responses. Overall, the survey mimics the real peer-to-peer network scenario where each participating identity is given a questionnaire to solve.

Figure 3 Survey participation statistics



There are 46.49% participants who attempted the survey only once. They are taken as honest entities in the network. There are 28% participants who attempted the survey more than four times. There are 25% participants who attempted the survey more than once and less than five times. The participants who attempted the survey more than once are taken as Sybil entities with each entity representing Sybil identities equal to the number of times they have attempted the survey. The survey is mapped to a network of 577 identities; out of which there are 86 honest identities and rest are Sybil identities grouped in 99 groups. The maximum Sybil identities are 21 in one group, the minimum and the average being 2 and 7.6 respectively. The Sybil groups can be classified into two classes: weak Sybil groups and strong Sybil groups. The weak Sybil groups are those whose number of identities is less than 5 and strong are those whose number of identities is equal or greater than 5. The weak groups are 47 and strong groups are 52. The Sybil groups in the mapped network are depicted in Figure 5.

Figure 4 shows the standard deviation of responses by individuals along with the number of times each have attempted. 65% of individuals or Sybil groups have standard deviation more than 0.2 indicating that some of their responses are outliers, i.e. they are too different from other responses. Those Sybil groups with standard deviation less than 0.2 mostly have two identities indicating high probability of being same. But when a person attempted larger number of times, responses are considerably differing. Considering this pattern of standard deviation, the input can be safely said to be a suitable sample for investigation.

Figure 4 Standard deviation of responses by individuals (see online version for colours)

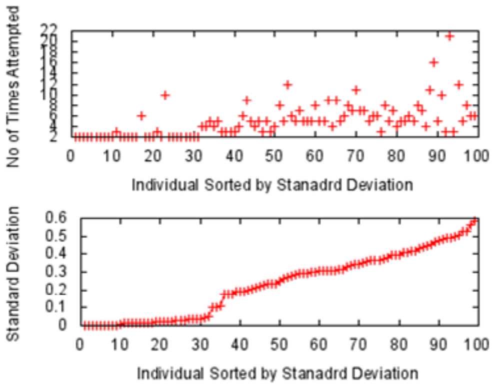
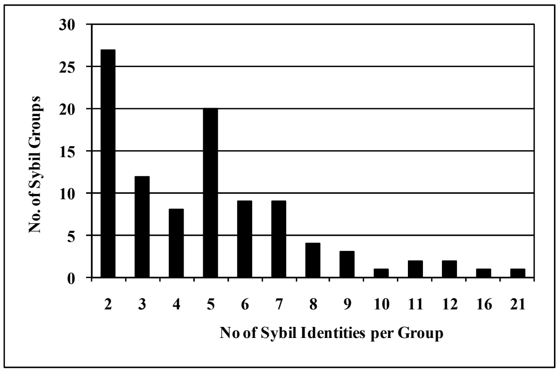


Figure 5 Sybil groups in the mapped network



6 Results analysis

The results are analysed with the objective to verify the effectiveness of the psychometric tests in detecting Sybil groups in an overlay network. Two types of data are analysed: rankings given to Luscher 8 colours and answers chosen for MBTI questionnaires. The similarity metrics for ranking data are Kendall's coefficient τ and Spearmans Correlation Coefficient ρ. DBSCAN algorithm is run on the ranking data collected in each questionnaire with ε taking values in the interval [0,1] each time incremented by 0.01 and *minPoints* taking values from 1 to 5. For each run, cluster goodness indices Rand Statistic, Jaccard Coefficient and Fawlkes and Mallows index are measured. Figures 6 and 7 show Jaccard coefficient and Rand Statistic for different combinations of ε and *minPoints*. It can be observed that the graphs reach the highest values at a particular combination. Table 4 shows the best recorded values of ε and *minPoints* of these indices. Similarly for the MBTI, the similarity indices are Cosine Similarity Metric and Pearson Correlation Coefficient. The best ε and *minPoints* are shown in Table 4. Rest of the analysis is carried out only at the best values of ε and *minPoints* as shown in Table 4.

Figure 6 Values of Jaccard coefficient for varied ε and minPoints using Kendall's τ

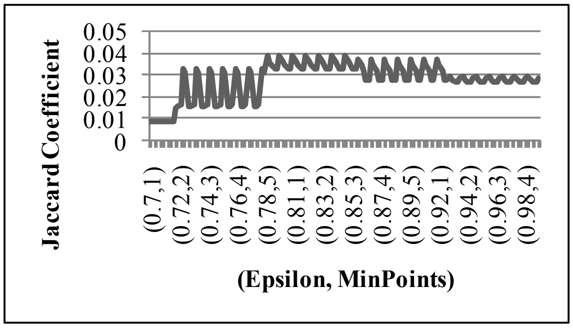
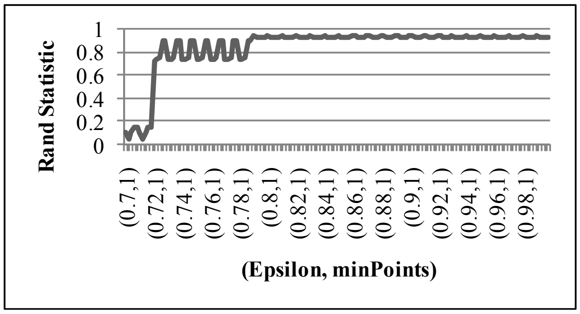
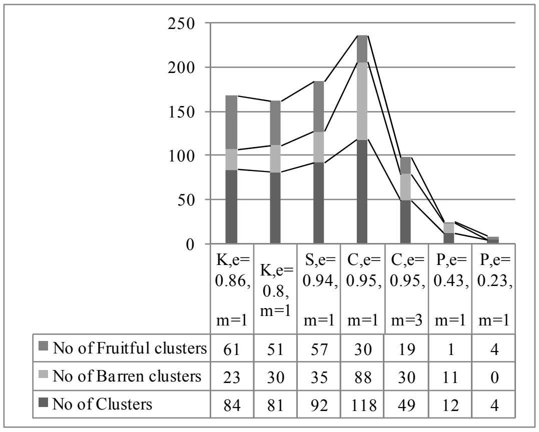


Figure 7 Values of Rand statistic for varied ε and minPoints using Kendall's τ



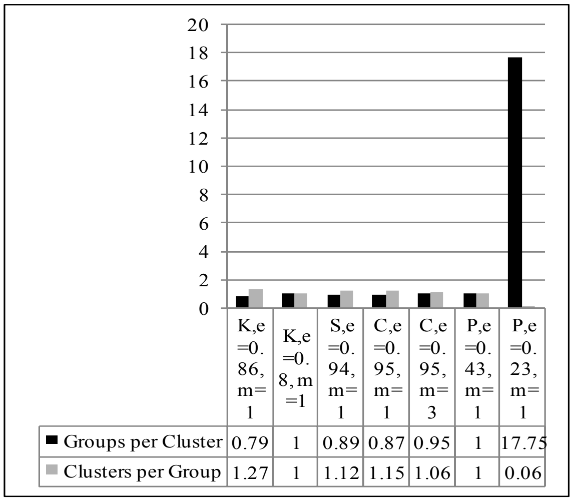
Overall the results show that our method is able to detect 51.5% of the Sybil groups of which 75% are strong Sybil groups. The percent of Sybil groups detected for each metric as shown in Figure 10. Although cluster formation with Pearson metric discover 71.7% of the Sybil groups, it has the limitation that all the Sybil groups are spread in just four clusters. And the percentage of false positives is too high summing to 25.7%. This is shown in Figure 12. That means that four clusters are crowded with unnecessary identities. So Pearson metric is not very useful in good cluster formation.

Figure 8 Cluster statistics for different metrics



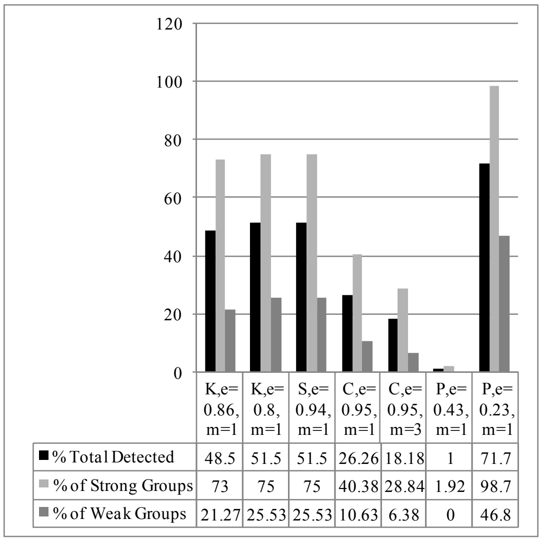
	K,e=0.86, m=1	K,e=0.8, m=1	S,e=0.94, m=1	C,e=0.95, m=1	C,e=0.95, m=3	P,e=0.43, m=1	P,e=0.23, m=1
No of Fruitful clusters	61	51	57	30	19	1	4
No of Barren clusters	23	30	35	88	30	11	0
No of Clusters	84	81	92	118	49	12	4

Figure 9 Spread of Sybil groups across clusters



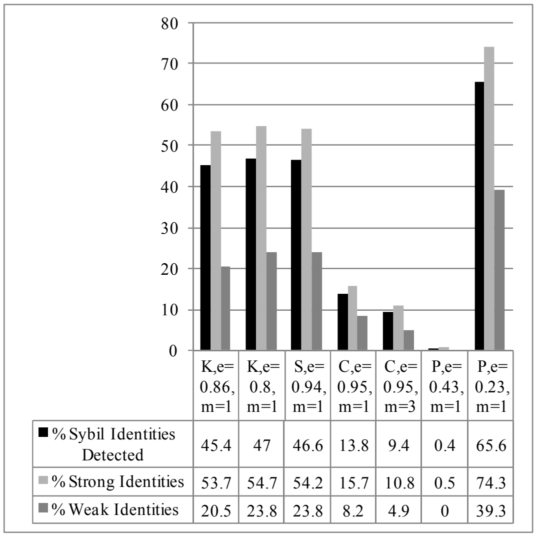
	K,e=0.86, m=1	K,e=0.8, m=1	S,e=0.94, m=1	C,e=0.95, m=1	C,e=0.95, m=3	P,e=0.43, m=1	P,e=0.23, m=1
Groups per Cluster	0.79	1	0.89	0.87	0.95	1	17.75
Clusters per Group	1.27	1	1.12	1.15	1.06	1	0.06

Figure 10 Percent of Sybil groups detected



	K,e=0.86, m=1	K,e=0.8, m=1	S,e=0.94, m=1	C,e=0.95, m=1	C,e=0.95, m=3	P,e=0.43, m=1	P,e=0.23, m=1
% Total Detected	48.5	51.5	51.5	26.26	18.18	1	71.7
% of Strong Groups	73	75	75	40.38	28.84	1.92	98.7
% of Weak Groups	21.27	25.53	25.53	10.63	6.38	0	46.8

Figure 11 Percent of Sybil identities detected

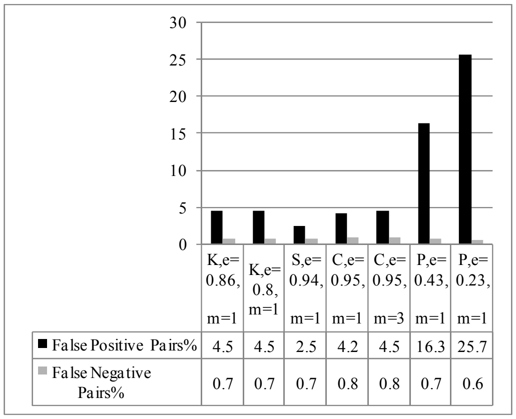


	K,e=0.86, m=1	K,e=0.8, m=1	S,e=0.94, m=1	C,e=0.95, m=1	C,e=0.95, m=3	P,e=0.43, m=1	P,e=0.23, m=1
% Sybil Identities Detected	45.4	47	46.6	13.8	9.4	0.4	65.6
% Strong Identities	53.7	54.7	54.2	15.7	10.8	0.5	74.3
% Weak Identities	20.5	23.8	23.8	8.2	4.9	0	39.3

Table 4 Values of epsilon and minPoints for DBSCAN algorithm at the best cluster formation stage

Test	*Neighbour Distance Metric*	*Rand Statistic*			*Jaccard Coefficient*			*Fowlkes &Mallows Index*		
		Epsilon	*Min Points*	*Value*	*Epsilon*	*Min Points*	*Value*	*Epsilon*	*Min Points*	*Value*
Luscher Short Colour Test	Kendall’s τ (K)	0.86	1	0.948	0.80–0.85	1	0.03937	0.86	1	0.104
	Spearman’s ρ (S)	0.94–0.95	1	0.9682	0.94–0.95	1	0.05782	0.94–0.95	1	0.126
MBTI	Cosine Similarity Metric (C)	0.95–0.99	1	0.95	0.95–0.99	3	0.01213	0.95–0.99	1	0.03
	Pearson Correlation Coefficient (P)	0.43–0.44	1	0.829	0.23–0.29	1-5	0.01329	0.43–0.44	1	0.046

Figure 12 False positives in cluster formation with different metrics



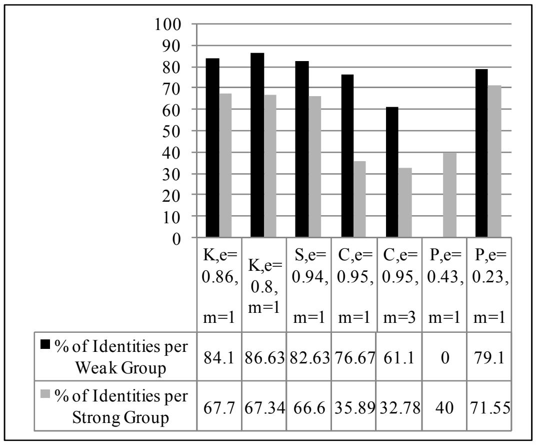
	K,e= 0.86, m=1	K,e= 0.8, m=1	S,e= 0.94, m=1	C,e= 0.95, m=1	C,e= 0.95, m=3	P,e= 0.43, m=1	P,e= 0.23, m=1
■ False Positive Pairs%	4.5	4.5	2.5	4.2	4.5	16.3	25.7
▨ False Negative Pairs%	0.7	0.7	0.7	0.8	0.8	0.7	0.6

The results show that Lusher's colour test gave better results compared to MBTI test. This may be due to the fact that in every questionnaire the colour set remained the same but the MBTI questions were not. The rankings given to colours were mostly consistent but the options chosen for the questions were not as consistent.

The quality of the Sybil groups detected can be observed from two perspectives. One is that if the ratio of clusters to groups is near to 1, it means that the whole cluster is dedicated to one Sybil group. This eases the further process of issuing computational puzzles simultaneously to all identities in one Sybil group. As shown in Figure 9, most of the metrics except the Pearson metric, the ratio comes near to 1.

It can be observed that it is difficult to detect weak Sybil groups compared to strong Sybil groups. This is due to the fact that there is small number of identities per group. The consistency of our approach is proportional to the number of identities in a Sybil group. Second perspective is to see how much of a Sybil group is discovered. If most of the identities of a Sybil group are discovered means all the identities will be issued computational puzzles putting real constraint on the resources of the Sybil entity. Figure 13 shows that the for Luscher's test, roughly 65% of the group is detected in strong group category.

Figure 13 % of Identities per Sybil group discovered



	K,e= 0.86, m=1	K,e= 0.8, m=1	S,e= 0.94, m=1	C,e= 0.95, m=1	C,e= 0.95, m=3	P,e= 0.43, m=1	P,e= 0.23, m=1
■ % of Identities per Weak Group	84.1	86.63	82.63	76.67	61.1	0	79.1
▨ % of Identities per Strong Group	67.7	67.34	66.6	35.89	32.78	40	71.55

It can be observed from Figures 10 and 11, detecting Sybil groups has advantages compared to detecting Sybil identities. For example, using Kendall's $\varepsilon = 0.80$, the number of Sybil groups discovered are 51 and the number of Sybil identities in all these groups are 340. For the same, the number of Sybil identities discovered is only 231 which is 47% of the total. If we go by discovering Sybil groups, we can end up removing 69% of the Sybil identities from the network.

7 Conclusion

In this paper we have presented a novel approach GAUR for detecting Sybil groups using psychometric tests. A survey is conducted amongst students and faculty in the campus. The experimental results have shown that 75% of the strong Sybil groups were detected. This shows that it is feasible to use psychometric tests to detect Sybil groups. The effectiveness of the test is also shown by the completeness of 67%, i.e. 67% of the identities in Sybil groups are detected. It is also shown that the greater the number of identities in a group, the better are the chances of detecting it. Also it is shown that detecting Sybil groups is more advantageous than detecting Sybil identities. The work can be improved by adapting better clustering algorithms like DENCLUE.

References

Bazzi, R.A. and Konjevod, G. (2005) 'On the establishment of distinct identities in overlay networks', *Proceedings of the 24th Symposium on Principles of Distributed Computing*, Las Vegas, USA, pp.312–320.

Bazzi, R.A., Ri Choi, Y. and Gouda, M.G. (2006) 'Hop chains: secure routing and the establishment of distinct identities', *Proceedings of the 10th International Conference on Principles of Distributed Systems*, Bordeaux, France, pp.365–379.

Bellman, R.E. (1957) *Dynamic Programming*, Princeton University Press.

Bhattacharjee, R. and Goel, A. (2005) 'Avoiding ballot stuffing in eBay-like reputation systems', *ACM SIGCOMM Workshop on the Economics of Peer-to-Peer Systems*, August 2005.

Bianchini, M., Gori, M. and Scarselli, F. (2005) 'Inside pagerank', *ACM Transactions on Internet Technology*, Vol. 5, No. 1, pp.92–128.

Borisov, N. (2006) 'Computational puzzles as Sybil defenses', *Proceedings of the 6th IEEE International Conference on Peer-to-Peer Computing*, IEEE Computer Society, Washington, pp.171–176.

Briggs, M.I. and Myers, P.B. (1980) *Gifts Differing: Understanding Personality Type*, Davies-Black Publishing, Mountain View, CA.

Buford, J., Yu, H. and Lua, E.K. (2009) *P2P Networking and Applications*, Morgan Kaufmann.

Camarillo, G. (Ed.) (2009) *RFC 5694 Peer-to-Peer (P2P) Architecture: Definition, Taxonomies, Examples, and Applicability*. Available online at: http://tools.ietf.org/search/rfc5694 (accessed on 6 March 2011).

Castro, M., Druschel, P., Ganesh, A., Rowstron, A. and Wallach, D.S. (2003) 'Secure routing for structured peer-to-peer overlay networks', *Proceedings of the 5th USENIX Symposium on Operating Systems Design and Implementation*, 9–11 December, Boston, USA.

Kamvar, S.D., Schlosser, M.T. and Garcia-Molina, H. (2003) 'The eigentrust algorithm for reputation management in p2p networks', *Proceedings of the 12th International World Wide Web Conference (WWW)*.

Cox, L.P., Murray, C.D. and Noble, B.D. (2002) 'Pastiche: making backup cheap and easy', *SIGOPS – Operating Systems Review*, Vol. 36, pp.85–98.

Craig, A., Depken II and Gregorius, B. (2010) 'Auction characteristics, seller reputation, and closing prices: evidence from eBay sales of the iPhone', *International Journal of Electronic Business*, Vol. 8, No. 2, pp.170–186.

Danezis, G. and Mittal, P. (2009) 'SybilInfer: detecting Sybil nodes using social networks', *NDSS*, San Diego, CA.

Danezis, G., Lesniewski, C., Kaashoek, M.F. and Anderson, R.J. (2005) 'Sybil-resistant DHT routing', *Proceedings of the 10th European Symposium on Research in Computer Security*, pp.305–318.

Dinger, J. and Hartenstein, H. (2006) 'Defending the Sybil attack in P2P networks: taxonomy, challenges, and a proposal for self-registration', *Proceedings of the 1st International Conference on Availability, Reliability and Security (ARES 2006)*, IEEE Computer Society, pp.756–763.

Dingledine, R., Mathewson, N. and Syverson, P. (2004) 'Tor: The second-generation onion router', *Proceedings of USENIX Security Symposium*, pp.303–320.

Douceur, J.R. (2002) 'The Sybil attack', *Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS'02)*, 7–8 March, Cambridge, USA, pp.251–260.

eDonkey2000 (2000) *Protocol Specification*. Available online at: http://hydranode.com/docs/ed2k/ed2kproto.php (accessed on 6 March 2011).

Ester, M., Kriegel, H., Sander, J. and Xu, X. (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise', *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp.226–231.

Fowlkes, E.B. and Mallows, C.L. (1983) 'A method for comparing two hierarchical clusterings', *Journal of American Statistical Association*, Vol. 78, pp.553–569.

Gnutella (2000) *Gnutella Protocol Specification Version 0.4*. Available online at: http://rfc-gnutella.sourceforge.net/developer/stable/index.html (accessed on 6 March 2011).

Gyongyi, Z. and Garcia-Molina, H. (2005) 'Web spam taxonomy', *1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pp.39–47.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) 'On clustering validation techniques', *Journal of Intelligent Information Systems*, Vol. 17, pp.107–145.

Haribabu, K., Arora, D., Kothari, B. and Hota, C. (2010) 'Detecting Sybils in Peer-to-Peer Overlays using Neural Networks and CAPTCHAs', *Proceedings of IEEE International Conference on Computational Intelligence and Communication Networks, CICN 2010*, Bhopal, pp.154–161.

Haribabu, K., Hota, C. and Saravana, S. (2009) 'Detecting Sybils in peer-to-peer file replication systems', *Proceedings of International Conference on Information Security and Digital Forensics, ISDF 2009*, City University, London, pp.152–164.

Haribabu, K., Paul, A. and Hota, C. (2011) 'Detecting Sybils in peer-to-peer overlays using psychometric analysis methods', *Proceedings of 25th IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA-2011)*, Singapore, pp.114–119.

Hu, Y.C., Perrig, A. and Johnson, D. (2002) 'Ariadne: a secure on-demand routing protocol for ad hoc networks', *Proceedings of ACM MOBICOM*, pp.12–23.

Jyothi, B.S. and Janakiram, D. (2009) 'SyMon: defending large structured P2P systems against Sybil attack', *Proceedings of International Conference on Peer-to-Peer Computing*, Seattle, 9–11 September, Washington, USA, pp.21–30.

Kazaa (2001) *Kazaa*. Available online at: http://en.wikipedia.org/wiki/Kazaa (accessed on 6 March 2011).

Lesniewski-Laas, C. and Kaashoek, M.F. (2010) 'Whānau: A Sybil-proof distributed hash table', *7th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2010)*, San Jose, CA, pp.111–126.

Levine, B.N., Shields, C. and Margolin, B.N. (2006) *A Survey of Solutions to the Sybil Attack*, Technical Report 2006-052, University of Massachusetts Amherst, Amherst, MA.

Luscher, M. (1969) *The Luscher Color Test*, Pocket Books, New York.

Margolin, N.B. and Levine, B.N. (2007) 'Informant: detecting Sybils using incentives', *Proceedings of Financial Cryptography (FC)*, Scarborough, Trinidad and Tobago, pp.192–207.

Napster (1999) *Napster*. Available online at: http://en.wikipedia.org/wiki/Napster (accessed on 6 March 2011).

P2P-Next (2008) *P2P-Next Generation Consortium*. Available online at: http://www.p2p-next.org/ (accessed on 6 March 2011).

Paul, J. (1901) 'Étude comparative de la distribution florale dans une portion des Alpes et des Jura', *Bulletin de la Société Vaudoise des Sciences Naturelles*, Vol. 37, pp.547–579.

Rand, W.M. (1971) 'Objective criteria for the evaluation of clustering methods', *Journal of American Statistical Association*, Vol. 66, pp.846–850.

Ratnasamy, S., Francis, P., Handley, M., Karp, R. and Shenker, S. (2001) 'A scalable content addressable network', *Proceedings of the 2001 ACM Annual Conference of the Special Interest Group on Data Communication (SIGCOMM)*, San Diego, USA, 27–31 August, pp.161–172.

Resnick, P., Zeckhauser, R., Friedman, E. and Kuwabara, K. (2000) 'Reputation systems', *Communications of the ACM*, Vol. 43, pp.45–48.

Pearman, R.R. and Albritton, S. (1997) *I'm Not Crazy, I'm Just Not You*, 1st ed., Davies-Black Publishing, Palo Alto, California.

Rowaihy, H., Enck, W., Mcdaniel, P. and LaPorta, T. (2007) 'Limiting Sybil attacks in structured peer-to-peer networks', *Proceedings of the 26th INFOCOM Conference*, St. Louis, MO, pp.2596–2600.

Rowstron, A. and Druschel, P. (2000) 'Pastry: scalable, distributed object location and routing for large-scale peer-to-peer systems', *IFIP/ACM International Conference on Distributed Systems Platforms*, 4–8 April, Middleware, New York, USA.

Stoica, I., Morris, R., Liben-Nowell, D., Karger, D., Kaashoek, M.F., Dabek, F. and Balakrishnan, H. (2001) 'Chord: a scalable peer-to-peer lookup service for internet applications', *SIGCOMM'01*, pp.149–160.

Wallach, D.S. (2002) 'A survey of peer-to-peer security issues', *Proceedings of Software Security – Theories and Systems*, November, Tokyo, pp.42–57.

Wang, H., Zhu, Y. and Hu, Y. (2005) 'An efficient and secure peer-to-peer overlay network', *Proceedings of the 30th Local Computer Networks*, Los Alamitos, CA, pp.64–771.

Yu, H., Kaminsky, M., Gibbons, P.B. and Flaxman, A. (2006) 'SybilGuard: defending against Sybil attacks via social networks', *Proceedings of the 2006 conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp.267–278.

Yurkewych, M., Levine, B.N. and Rosenberg, A.L. (2005) 'On the cost-ineffectiveness of redundancy in commercial p2p computing', *Proceedings of ACM CCS*, pp.280–288.