2011 Workshops of International Conference on Advanced Information Networking and Applications

Detecting Sybils in Peer-to-Peer Overlays using Psychometric Analysis Methods

K Haribabu, Arindam Paul
Computer Science & Information Systems
Birla Institute of Technology & Science
Pilani, Rajasthan, 333031, INDIA
{khari, h2009420}@bits-pilani.ac.in

Chittaranjan Hota
Computer Science & Information Systems
Birla Institute of Technology & Science
Hyderabad, Andhra Pradesh, INDIA
hota@bits-hyderabad.ac.in

***Abstract*— Peer to peer networks are fast becoming the most popular file sharing media, guaranteeing complete user anonymity to the clients. However, modern P2P networks suffer from Sybil attacks, which forge multiple identities to influence the global decisions in the network. This paper suggests a novel solution to minimize Sybils influence using unique combination of Psychometric Tests, Color Tests & CAPTCHAs. Our survey has shown that the proposed approach is promising in detecting not just Sybils but Sybil groups. The results have shown that 30-50% of Sybil groups are detected.**

Index Terms— ***P2P Networks, Sybil Detection, Psychometric Analysis***

I. INTRODUCTION

P2P overlay networks are application-level logical networks on top of the Internet with their own topology and routing, primarily used for file-sharing. The major characteristics of P2P networks are shared provision of distributed resources and services, decentralization and autonomy. P2P overlay networks are categorized 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 e.g. Gnutella [1] and Kazaa [2]. While in structured, network topology is tightly controlled and content is placed at specified locations which make subsequent queries more efficient e.g. Content Addressable Network (CAN) [3], Chord [4], and Pastry [5].

In a Sybil attack, an entity in a peer-to-peer network masquerades itself as multiple simultaneous identities in the network [6]. Thus, a single user can exert a significant effect on the decisions or working of the entire network if the multiple identities created by the user form a significant fraction of the peer-to-peer network. Companies increase the rank of the web pages in Google search results [7] and some people associate certain search terms with popular personalities out of fun [8]. It does a lot of damage in P2P networks as compared to sensor [9] and ad-hoc networks because of their large global size and lack of physical constraints and as setting up an attack involves one physical entity and many different entities or abstractions. Other than computational [10-11], messaging [12], file-sharing [1], P2P networks are used to rank objects by popularity. Malicious users can use Sybil identities to distort the ranking process and make it work in their favor [13].

This paper attempts to identify the Sybil groups using psychometric ratings. The psychometric ratings are based on the answers to questions based on personal psychological nature. Since a single malicious user is creating multiple identities, he will be required to answer multiple questionnaires on behalf of the identities he has created, in contrast to a single honest user. Our solution is based on analyzing the questionnaires received from all the identities and clustering them based on the common psychological characteristics. The clusters are suspected to be originating from the same malicious user. The suspicion is verified by use of CAPTCHAs.

The rest of the paper is organized as follows. Sec. II discusses the related work done in Sybil detection, followed by our algorithm in Sec. III. Section IV talks about our survey results, and Sec V concludes our work

II. RELATED WORK

The goal of resource testing consisting of checks on computing and storage ability [14], and network bandwidth, as well as limited IP addresses [15] to determine if a number of identities possess fewer resources than would be expected if independent by computation puzzles. [16-17] present computation puzzle based approaches to test the resources of a Sybil entity. CAPTCHAs are automated puzzles forcing human effort but which are difficult for a computer to solve .The problem with these approaches is that even the honest nodes have to repeatedly solve the puzzles. Bazzi and Konjevod [18] have proposed that every identity is issued a geometric certificate by a set of beacons nodes in the network. Wang [19] has presented that an identity is identified by IP, MAC and a vector of RTT measurements from designated land marks. These approaches fail when the node changes its physical location. Castro et al. [20] argue that in a P2P overlay network, if a central authority distributes uniform node identifiers (IDs) then it is difficult for attackers to have any control over this process. Certainly this solution prevents

978-0-7695-4338-3/11 $26.00 © 2011 IEEE
DOI 10.1109/WAINA.2011.107

Sybils but it slows down the propagation of the network services to new users. Neural network based approach [21] collects the metrics for different behavioral aspects and classifies them using trained network. In Sybilguard [22], the authors have proposed a distributed algorithm for limiting entry of Sybil identities into a social network. They have used the principle that in a social network, the trusted edges between honest group and a Sybil group will be very few. The problem with these approaches is that they work only with networks that are evolved based on social trust relationships, which is generally not the case. Symon et al. [23] have proposed a Sybil monitor to observe and record the transactions of an individual node which record fake transactions.

To the best of our knowledge in this field, this is the first time psychometric tests are being used for detecting Sybils. Except [22], only this approach aims at finding Sybil groups instead of individual Sybil identities. The advantage in such detection is that the proliferation can be stopped at the root by ostracizing the whole group. In case of detecting individual Sybil identities, as soon as some Sybil identities are detected and ostracized, the malicious user can introduce more identities into the network compensating the loss.

III. Background

In this section we give a brief background about the Gnutella network and then talk about psychometric analysis methods.

A. Gnutella Protocol

Gnutella was the first fully decentralized peer-to-peer file sharing networks. Gnutella is a decentralized P2P system wherein the file transfer load is bore between the computer exchanging files, but the file searching traffic is distributed across the network. The protocol has a simple structure and uses broadcasts to locate files. Each node in the network can function as both client as well as server (hence referred to as servent). They can issue queries to other nodes as well as accept and respond to queries from other nodes, after matching the queries with the contents loaded in their own hard disks. The Gnutella version 0.6 network protocol employs six different descriptor types for communicating data between servents, namely - ping, pong, query, query hit, push and bye. To make Gnutella scalable, the network is enhanced with several changes, one of them being an ultra peer network.

B. Psychometric Tests and Color Tests

Personality Type or Psychological Type are terms most commonly associated with the model of personality development created by Isabel Briggs Myers, the author of the world's most widely used personality inventory, the MBTI or Myers-Briggs Type Indicator [24]. We divide the human psyche into 4 pairs of categories of functional types (extroversion-introversion, sensing-intuition, thinking-feeling, and judging-perceiving). Based on the response of a person to a question set, we rate a person based on the 4 categories. We have extended this MBTI model and we calculate a psychometric index based on the ratings given according to the MBTI model.

Apart from the MBTI psychometric analysis methods, we also used Luscher Color Test [25] analysis method. There is a set of eight colors and one is asked to rate the colors according to their preferences from most preferred to least preferred. According to psychologists, color has a deep rooted psychological influence. For e.g. colors like red signify passion and aggression while colors like blue signify calmness and peace. So, based on what color a person chooses, we can understand the psychological traits in a person. We try to extend this concept and try to generate a cumulative color index of a person.

IV. Proposed Solution

In this section we describe a psychometric analysis based Sybil detection solution. We try to solve the problem of Sybils in peer-to-peer networks by detecting their malicious behavior with the help of generating cumulative indices based on Myers Briggs Psychometric Test and Luscher Color Test. We model our solution on an 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.

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 ultra-peers are not Sybil nodes. i.e. the Sybil peers exist only at the lowest level.

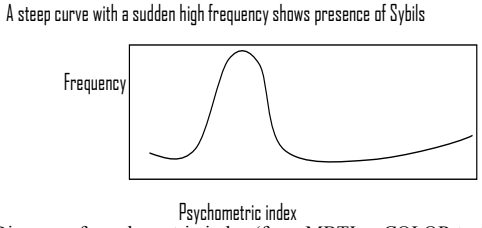

Fig 1: Diagram of psychometric index (from MBTI or COLOR test according to the theoretical model)

The psychometric index (plotted on the X axis) is a cumulative value i.e. derived out of the options selected by the peer. On the Y-axis, we plot the frequency i.e. the number of peers in the network having same psychometric index. In case of presence of Sybils, there will be a sudden peak/plateau or zone of high frequency indicating that many peers have similar psychological characteristics as illustrated by Fig 1.

A. Questionnaire preparation

The questionnaire will evaluate the psychometric index of the peers based on MBTI or Luscher Color Test. The

questionnaire consists of 3 kinds of questions. Firstly, there are questions based on each of the 4 categories of personality traits according to MBTI model [26] which gives us the information about the psychological orientation of a person. For e.g., we can ask a person whether he likes to hear less and talk more or he likes to hear more and talk less. Based on which he likes more, he can give a score ranging from 1-5 with 1 standing for strongly wanting to talk less and hear more.In the following tables, we have enumerated some of the questions which help us to know about each of the traits in a person.

Table 1: Sample questions for testing introvert quality

Quality	Questions
Extroversion-Introversion	1) Do you talk more than listen or vice-versa?
	2) Do you have high energy or quiet energy?
	3) Do you want to stay behind scenes or want a public role?

Table 2: Some questions for testing intuition quality

Quality	Questions
Sensing-Intuition	1) Do you focus on details or do you see the big picture?
	2) Do you work at a steady pace or bursts of energy?
	3) Do you trust gut instincts?

Table 3: Some questions for testing feeling quality

Quality	Questions
Thinking-Feeling	1) Do you appear cool and reserved or warm and friendly?
	2) Do you value honesty and fairness more or harmony and compassion?
	3) Do you tend more to see faults or you are quick to compliment others?

Table 4: Some questions for testing perceiving quality

Quality	Questions
Judging-Perceiving	1) Do you work first, play later or play first, work later?
	2) Do you like to make and stick to plans or keep flexible plans?
	3) Do you like freedom to be spontaneous or find comfort in schedules?

Secondly, we have 2 sets of color test (each color tests have 8 colors) in each questionnaire in which we ask the peers to fill in their preferences of colors from most preferred to least preferred. There are 2 sets of color tests (with the colors in each set arranged in a different order) so as to reconfirm the color choice of the peers. Each color is associated with some particular trait [27], [28]. These associations are findings of the research in psychology field. For e.g. the orange-red color represent "force of will" & correspond to desire, domination, aggression, controlled passion, concern for others. Also, the questionnaire includes some human-solvable questions (i.e. questions which ask for simple details or some simple numerical questions written in a linguistic way) that are mixed with the psychometric questions. These are helpful in counteracting the automated programs which try to answer our questionnaires.

B. Questionnaire Evaluation

The leaf peers respond to the questions and then, the ultra-peers evaluate the questions and thus determine the psychological traits of the peer. For e.g. there are 3 questions based on a particular trait in a questionnaire, say Extroversion-Introversion with each question having a score of 1-5 with 1 pointing to introversion and 5 for extroversion. Then, if a person gets an Introversion-Extroversion (ie) score of 6 out of 15(i.e. 40% score), then the person can be said to be 40% introvert and 60% extrovert. Similarly, the same technique is used for other 3 traits.

Again for the color tests, a person can choose any possible order of colors among the 8 colors. As there are 2 sets of color tests and 8 colors in a color test, we can have $(8!)^2$ possible cases. So, this can give us a very accurate insight into the psyche of the person based on the choice of the colors. To give a simplified e.g., if a person chooses orange red as most preferred and black as the least preferred, Luscher Color test will predict the person as one who fights against restrictions or limitation, and insists on developing freely as a result of one's efforts. On the other hand, if he chooses blue as most preferred and green as least preferred, we can say that stress and tension induced on the person is beyond his capacity and he wants to escape in to a peaceful and problem-free situation.

C. Leaf nodes' behavior

A leaf node on receiving a questionnaire as a query will answer it back to the ultra-peer. A protocol is established that each questionnaire should be answered in a particular time interval or else, a node is given a warning that it might be labeled as a malicious node. In case, the node still does not answer even after some time interval after first warning, the node is then black-listed or treated as a malicious node and taken out of the network. For the node to again come back, it has to again join as a new node and follow the protocol.

D. Ultra-peer's behavior

Each ultra-peer in the network will generate a questionnaire from a pool of questions including a sub-pool of questions on each personality traits. Then, it will send it to the leaf-nodes inside the network. It will then enforce the protocol (as mentioned above) and collect the response of the sub-ordinate peers. Then, it will evaluate the psychometric indices based on the MBTI and Color Test and pass it to the Central Server.

E. Central Server

The central server maintains and periodically updates psychometric index ratings of all the peers within the network. It then tries to cluster the nodes based on their psychometric index values. In case of the presence of a Sybil, there will be presence of a huge number of nodes in a particular cluster.

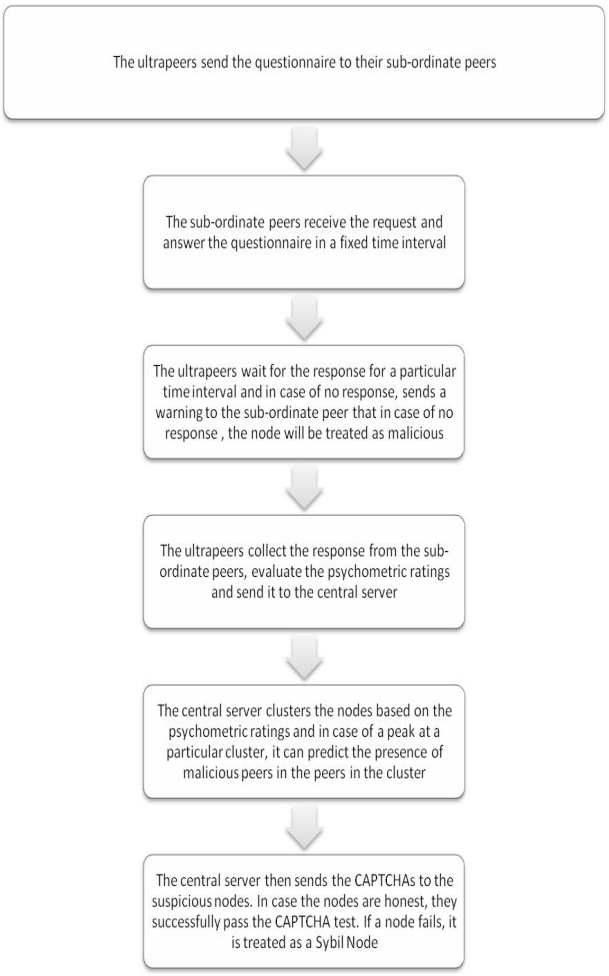

Fig 2: Flow diagram representing the step-wise procedure in the model

Then, we can mark all this nodes as suspicious and a CAPTCHA is sent as a final confirmation. . If the suspect node fails to pass the CAPTCHA test it is declared as a Sybil. An honest node might also get wrongly predicted as a Sybil, but it will always pass the CAPTCHA test, thus we can eliminate the presence of false positives.

F. Strengths and Limitations

The strength of this approach is that the malicious user who is managing huge Sybil identities in the network can't fool the psychometric tests as they are proven in the fast to reveal the truth even if the person is manipulative. Our approach identifies Sybil groups instead of Sybil identities.

One of the limitations are honest user's psychology may match with the psychology of a malicious user, thus falling in the same cluster as the malicious user. This is verified during the secondary testing. The test is based on psychological metrics, it might happen that the psychometric ratings of some Sybil identities coming from the same Sybil group may not fall within the same cluster. Therefore, a node may be Sybil but might not fall within the cluster and thus, may not get detected. This introduces false negatives in our proposed model. But these individual nodes are part of Sybil group where Sybil group itself is detected. Thus they also have to be uprooted once the group is detected.

V. SURVEY RESULTS

In this paper, the results are based on survey and not on simulation. This paper proposes a solution where people i.e. the users of the peer-to-peer network have to consciously answer the questions and thus, the responses cannot be generated by automated simulation. So we have taken to survey approach. We have conducted a survey on population of 50 and given 3 questionnaires to each. It is analogous to each person representing a malicious user and each user creating 3 virtual identities.

Some assumptions have been made in the survey:

- We gave each person 3 sets. The sets are distinct but having similar composition of questions. It is assumed that all three questionnaires have the capability to bring out the 4 qualities of the person.
- People are given the questionnaires at different times so as not to influence the choices of one with another.
- We surveyed people over different gender, profession and age groups and the data is assumed to give a holistic picture but there can be some variations as people of the age-group of college goers are in majority.
- It is also assumed that population being surveyed have thoughtfully answered the questions. Also, they have sufficient intellectual and emotional maturity to answer the psychometric as well as innocent questions.

We have chosen the people for our survey as students, and faculty in the university campus. After collecting the filled in questionnaires, the data is tabulated. Each questionnaire carries questions for two types of tests namely psychometric test and color test. For the psychometric test, we evaluate the percentage for each of the 4 psychological traits in a person. For each trait, the score obtained is projected onto [-1, 1] interval. This will give closer view of overlaps of the four qualities amongst different questionnaires. We designate first one-third as +1, next one-third as 0 and rest as -1. For e.g. a person having ratings {-1, -1, 0, -1} for four traits, it indicates that he is more of an extrovert, intuitive, perceiver person having balance between thinking and feeling. The first rating refers to introversion-extroversion, second to sensing-intuition, third for thinking-feeling and last for judging-perceiving. To get a single metric that combines all four trait ratings, we use a composite function where the weights are arbitrary prime numbers i.e. 11111, 7, 111, and 5977.

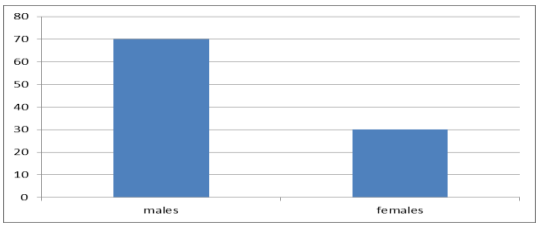

Fig 3: Bar-graph comparing the percentage of male and female participants in the survey (70% and 30% respectively)

The color test is evaluated with main aim being to characterize the order in which the colors are chosen. The order is characterized by computing a metric by giving weights in the order of their selection. We evaluate the color tests based on some weights which we give to every color. For the color set, the weights are the first 8 prime numbers (2, 3,…, 17, 19).

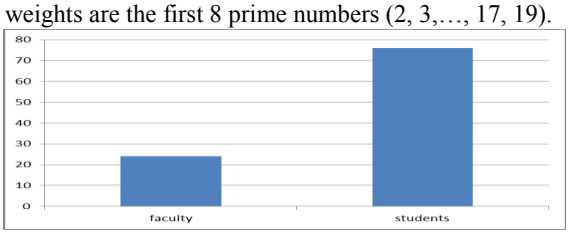

Fig 4: Bar-graph comparing the percentage of faculty and student participants in the survey

The weights for the order are taken as powers of 2. The prime numbers and powers of 2 are chosen as weights so that their combination will produce a unique value.

The color test value is computed as

$$\sum_{i=1}^{8} colorWeight(i) * orderWeight(i)$$

Like this for each set, the color test values are computed.

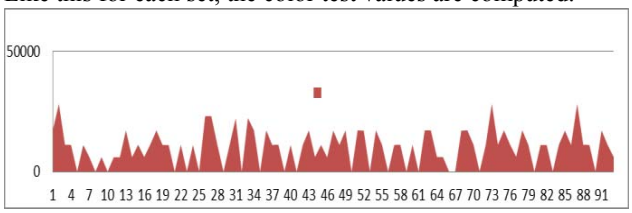

Fig 5: Graph of the psychometric index(MBTI) vs. the questionnaires which we have been given to the nodes.

The vertical axis refers to the psychometric index according to the MBTI model and the horizontal axis refers to the set of questionnaires. Fig 5 is based on MYERS BRIGGS TEST shows plateaus and troughs and almost of similar width (generally of 3) signifies that the papers given to same person have almost similar psychometric rating which when done in a real scenario with a Sybil having several pseudonymous entities along with many honest nodes can tell us about the presence of Sybils.

The vertical axis refers to the psychometric index according to the MBTI model and the horizontal axis refers to the set of questionnaires. Thus we are able to spot that how same person have similar psychometric ratings. Thus, when done in a real situation where many entities are Sybil identities and many honest nodes, we will be able to successfully detect Sybils by mapping their psychometric rating.

There are certain limitations of the survey model.

- The age and educational qualification of the survey population are mostly similar (age: 18-25 and education: engineering background), so this may lead to some bias in the result.
- The survey has been conducted on a small population i.e. on 50 people so this may cause some bias in the model.
- We have given 3 sets of similar yet distinct questionnaire to each of the 50 people surveyed. An improved model will be giving one set of questionnaire to majority of the surveyed population and then giving 10-20 sets of questionnaire to the remaining population so as to divide the population into ordinary and Sybil nodes.

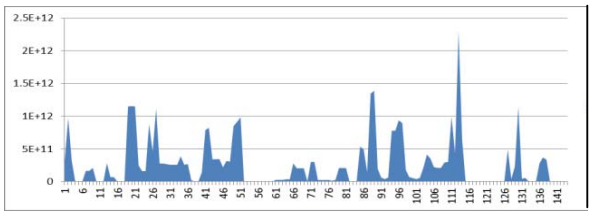

Fig 6: Area Plot of psychometric color index vs. the questionnaires given to various nodes.

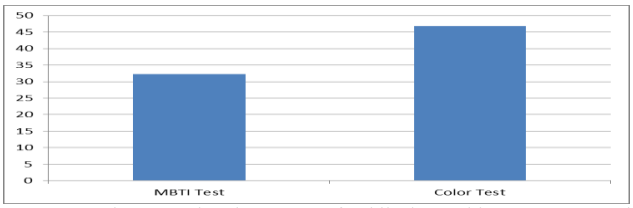

Fig 7: Bar-graphs comparing the percent of Sybils detected by MBTI test and Color Test (32% and 47% respectively)

VI. Conclusion

The paper presents a novel approach to limit the corruptive influence of Sybil entities on partially decentralized P2P networks. The approach is based on psychometric tests. Peaks that we get when the frequency of the psychometric indices are plotted with the psychometric scores from color and MYERS BRIGGS psychometric tests indicate the presence of Sybil groups. So, here we present a very powerful technique to detect Sybil groups and more so as it is based on the psychology of the users involved - honest or malicious. Also, we plan to improve our approach by designing the questionnaire in such a way that we minimize the effect of human manipulative tendencies while filling the questionnaires. Also we plan to improve our survey techniques by approaching a larger population and requiring a particular user to answer large sets of questionnaires.

ACKNOWLEDGEMENTS

We thank Ganeshan Subramanian for kindling this idea that psychometric tests can be used to detect sybils.

REFERENCES

[1] Gnutella protocol 0.6. http://rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html
[2] Kazaa. http://www.kazaa.com
[3] Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, S.: A Scalable Content Addressable Network. In: Proceedings of the 2001 ACM Annual Conference of the Special Interest Group on Data Communication (SIGCOMM), pp. 161--172, ACM Press, (2001)
[4] Stoica, I., Morris, R., Liben-Nowell, D., Karger, D., Kaashoek, M.F., Dabek, F., Balakrishnan, H.: Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. IEEE/ACM Transactions on Networking 11, 17--32 (2003)
[5] Rowstron, A., Druschel, P.Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. In: Proceedings of IFIP/ACM International Conference on Distributed Systems Platforms (Middleware), pp. 329--350, Springer-Verlag, London (2001)
[6] Douceur, J.R.: The Sybil attack. In: Proceedings of the 1st International Workshop on Peer-to-Peer Systems, pp. 251-260. Springer New York (2002)
[7] Bianchini, M., Gori, M., Scarselli, F.: Inside page rank, ACM Transactions on Internet Technology, 5(1):92–128, (2005)
[8] Viglucci, A., Tanfani, J., Getter, L.: Herald special report: Dubious tactics tilted mayoral votes. Miami Herald, February 8, 1998
[9] Dinger, J., Hartenstein, H.: Defending the Sybil Attack in P2P Networks: Taxonomy, Challenges, and a Proposal for Self-Registration. In: Proceedings of the First International Conference on Availability, Reliability and Security (ARES 2006), pp. 756 - 763, IEEE Computer Society (2006)
[10] Anderson, D.:SETI@home in Peer-to-Peer: Harnessing the Benefit of a Disruptive Technology, O'Reilly & Associates, CA (2001)
[11] Larson, S.M., Snow, C.D., Shirts, M., Pande, V.S.: FOLDING@home and GENOME@home: Using distributed computing to tackle previously intractable problems in computational biology, Computational Genomics, 2002.
[12] Miller, J.: Jabber: Conversational technologies in Peer-to-Peer: Harnessing the Benefits of a Disruptive Technology, O'Reilly & Associates, CA (2001)
[13] Riley, D. Techcrunch: Stat gaming services come to youtube, 2007. http://www.techcrunch.com/2007/08/23/myspace-style-profile-gaming-comes-to-youtube.
[14] K.Haribabu, Chittaranjan Hota, Saravana S. Detecting Sybils in Peer-to-Peer File Replication Systems, in Proc. International Conference on Information Security and Digital Forensics, ISDF 2009, City University, London, Sept 2009, pp. 152-164, Springer, 2009.
[15] Brian Neil Levine, Clay Shields, and N. Boris Margolin, "A Survey of Solutions to the Sybil Attack," Tech report 2006-052, University of Massachusetts Amherst, Amherst, MA, October 2006
[16] Borisov, N.: Computational Puzzles as Sybil Defenses. In: Proceedings of the Sixth IEEE International Conference on Peer-to-Peer Computing, pp. 171 – 176. IEEE Computer Society Washington (2006)
[17] Rowaihy, H., Enck, W., Mcdaniel, P., AND LA Porta, T. Limiting Sybil Attacks in Structured Peer-to-Peer Networks. In Proc. 26th INFOCOM Conference (St. Louis, MO). IEEE Computer Society Press, Los Alamitos, CA., 2596–2600.
[18] Bazzi, R. A. AND Konjevod, G. On the Establishment of Distinct Identities in Overlay Networks. In Proc. 24th Symposium on Principles of Distributed Computing (Las Vegas, NV). ACM Press, New York, NY, 2007, 312–320.
[19] Wang, H., Zhu, Y., AND Hu, Y. An Efficient and Secure Peer-to-Peer Overlay Network. In Proc. of 30th Local Computer Networks. IEEE Computer Society Press, Los Alamitos, CA.,2005, 64–771.
[20] Castro, M., Druschel, P., Ganesh, A., Rowstron, A., Wallach, D.S.: Secure routing for structured peer-to-peer overlay networks. In: Proceedings of the 5th USENIX Symposium on Operating Systems Design and Implementation, pp. 299 – 314, New York: ACM Press (2003)
[21] Haribabu, K, Arora, D, Kothari, B, Hota, C. Detecting Sybils in Peer-to-Peer Overlays using Neural Networks and CAPTCHAs. In Proc. International Conference On Computational Intelligence and Communication Networks.. IEEE Computer Society (2010).
[22] Yu, H., Kaminsky, M., Gibbons, P.B., Flaxman, A.: SybilGuard: Defending against Sybil attacks via social networks. In: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, pp. 267 – 278. ACM Press New York (2006)
[23] Jyothi, B. S., Janakiram, D., SyMon: Defending Large Structured P2P Systems Against Sybil Attack , In: Proceedings of International Conference on Peer-to-Peer Computing, Seattle Washington, USA, Sept 9-11, 2009, 21-30
[24] Myers Briggs Foundation http://www.myersbriggs.org/
[25] Luscher Color Test http://www.luscher-color.com/
[26]Myers Briggs Type Indicator Form G http://dluz.com/Rion/GeneralInterest/Meyers_Briggs/Meyers-Briggs_FormG_Question_Booklet.pdf
[27]Color Psychology http://www.viewzone.com/luscher.html
[28]Color: Meaning, Symbolism and Psychology http://www.squidoo.com/colorexpert