Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 9 (19): 3547-3554, 2009
ISSN 1812-5654

Modeling and Simulating Semantic Social Overlay Peer-to-Peer Systems

A. Modarresi, A. Mamat, H. Ibrahim and N. Mustapha
Faculty of Computer Science and Information Technology, University Putra Malaysia,
Jln. UPM, Serdang, 43400, Selangor, Malaysia

Abstract: The complexity of Peer-to-Peer (P2P) systems makes their analytical evaluation complicated. To conquer this problem, simulation studies are usually used to evaluate such systems. However, evolution of P2P systems, from simply a Gnutella-like network to advanced overlays, makes their comparison difficult in a similar condition. Using different inputs, outputs and datasets is the main cause of this problem. On the other hand, network simulators cannot be used for such systems because of high scalability and dynamicity of P2P systems. Most network simulators can simulate few nodes in high detail without considering dynamicity of systems. In this study, a generic model, parameters and datasets are presented and used to design a flow-based P2P simulator with the capability of implementing different P2P protocols to simplify the evaluation of P2P systems. Then, the behavior of a semantic social overlay P2P system is investigated and compared with two various types of overlays, namely random and interest-based systems to show the applicability of the simulator. Although three different types of overlays have been chosen, the generic model and selected parameters used in the proposed simulator provide a uniform environment to evaluate and compare different types of overlays in similar conditions.

Key words: Peer-to-peer computing, social network overlay, peer-to-peer modeling, simulation design

INTRODUCTION

Since introducing Napster different overlays, from completely random to semantic centric and social network concepts have been devised. Some samples of this broad range are shown by Crespo and Garcia-Molina (2005), Klampanos and Jose (2004), Tempich *et al.* (2004), Hasse *et al.* (2004a), Khambatti *et al.* (2004) and Sun *et al.* (2006). The main objective of all these structures is to find a particular data, based on query criteria in an acceptable amount of time and with logical resource consumption. In all these models, one of the most important factors that affects the mentioned goal is the location of peers in the overlay and their connections to other neighbors. In a simple random topology, peers find their neighbors randomly, but in a more advanced model such as semantic overlays, some kinds of metadata are gathered and used to find more effective neighbors. In social network models, social network concepts like communities, hubs and bridges are considered in the structure of the model.

Characteristics and topology of the overlays define some internal complexities for systems. The policy of settlement for a node in the overlay, routing algorithms and the way of establishing a connection with other nodes are some of these characteristics. These complexities along with input variables and output results must be identified and measured clearly. Considering all these elements together makes comparison of different models difficult. Some studies use different internal parameters to measure and investigate the efficiency of the overlay while other studies use different queries and data distribution to simulate. Some of these differences can be found in the mentioned structures.

In this study, a generic model is introduced and a flow-based simulator is designed to simulate different overlays with similar parameters. The parameters are chosen in a way that, they can show efficiency of the model as well as the performance of the structure. Then, a semantic social P2P network is simulated by the implemented simulator and compared with an interest-based and random model. The aim of this study was not introducing a methodology or framework for P2P systems, but to present a simulator that can simulate many types of overlays in similar conditions.

Overlay networks impose special requirements to simulation environments. These kinds of networks differ from computer network and distributed systems in several aspects. Some of these aspects like autonomy of nodes, decentralization control and heterogeneous shared resources do not impose special requirements to the simulator, but many of them like scalability and

Corresponding Author: Amir Modarresi, Faculty of Computer Science and Information Technology, University Putra Malaysia, Jln. UPM, Serdang, 43400, Selangor, Malaysia Tel: +60-17-2092620

dynamicity do. Moreover, overlay networks do not need a precise implementation for the transport layer and other lower layers of the network stack; though implementing them needs more memory which causes less scalability. Hence, some state-of-the-art simulators like NS-2 (Network Simulator), OPNET (Optimized Network Engineering Tools) (OPNET Technologies Inc., 2008), GloMoSim (Global Mobile Information System Simulation Library) (UCLA Computing Library), OMNeT++ (Objective Modular Network Test Bed) cannot be used as proper simulators. These simulators are precise packet-level and simulate systems with precise details of network protocols. However, such accuracy decreases the speed of simulation and its scalability. Moreover, considering a packet as a designated event in the system makes events queue longer and decreases simulation speed (Ahn and Danzig, 1996).

Requirement of scalability limits the accuracy of overlay network simulators. Therefore, output parameters must be chosen carefully to cover some parts of this limitation. A simulated network overlay is usually represented by a graph with the capability of message passing without considering all low-level details of network stack. In some cases, even network bandwidth and delay are also dropped where the structure of overlay is studied.

Narses (Giuli and Baker, 2002) is a flow-based simulator to simulate large-scale distributed applications. It considers chunk of bytes instead of packets to avoid overhead of packet-level simulation. Narses simulates end users connections without considering whole bandwidth and intermediate routers. It does not consider any network congestion. In Spite of such characteristics, Narses can just simulate 600 nodes and no P2P protocols have been implemented (Brown and Kolberg, 2006).

The 3LS (3-Level Simulator) (Ting and Deters, 2003) is another overlay simulator that can simulate less than 1000 nodes and implement Gnutella protocol (Brown and Kolberg, 2006). In this simulator, network model is described in 2D space in terms of distance among nodes. The system is also separated to 3 models namely, networked, protocol and user.

PeerSim is another overlay simulator that can simulate a collection of internally developed P2P model, but it does not have any capability to simulate a semantic based overlay.

The P2PSim can simulate Chord, Tapestry, Kademlia, Accordion and Kelips (Brown and Kolberg, 2006). However, these implementations are exclusively performed in P2PSim and do not consider all features of these protocols (Brown and Kolberg, 2006).

A SOCIAL SEMANTIC SIMULATION MODEL

The main consideration in design of the proposed simulator is social semantic P2P overlays; however, other facilities have been taken into account to define different types of overlay in order to provide a uniform environment for comparison. Here, first a general view of a node in a social semantic model is presented and then the characteristics of the simulator are expressed. It is needless to say that a node in a simple overlay can be defined easily when semantic information is ignored.

The view on a peer in social semantic model is as follow:

- Each peer stores a set of items which are shared in the network. It is assumed that the interest of each peer is not random; therefore each peer shows a tendency to some particular interests. It is also assumed these interests are chosen from a shared ontology which defines the domain of the system. Each peer may have more than one interest, but few random items, shared by a particular peer, are not considered as its interests
- Each peer attends a community that its members have the same interests. If a peer has many interests, it attends several communities
- Each peer knows some similar peers in a designated community which are physically closer than other peers to the respective peer. This condition creates a sub-community inside a community. Many sub-communities may be created inside one specific community
- In each sub-community, there is a member who is more knowledgeable than the others. Such peers usually have more connections with other members. These members are called hubs. In this model each peer finds a hub in a proper sub-community from proximity points of view
- Hubs can be connected to each other and create connected sub-communities
- Each community is represented by the representative of that community. Each representative usually knows some information about the structure of the community and its sub-communities
- A bootstrapping node or a super peer which knows the shared ontology and its respective environment helps new nodes to join a proper community by introducing the representative of the community. This node covers the structural holes among communities

A peer can pose a query to find a particular item to all or some of its neighbors. Each query may be forwarded to

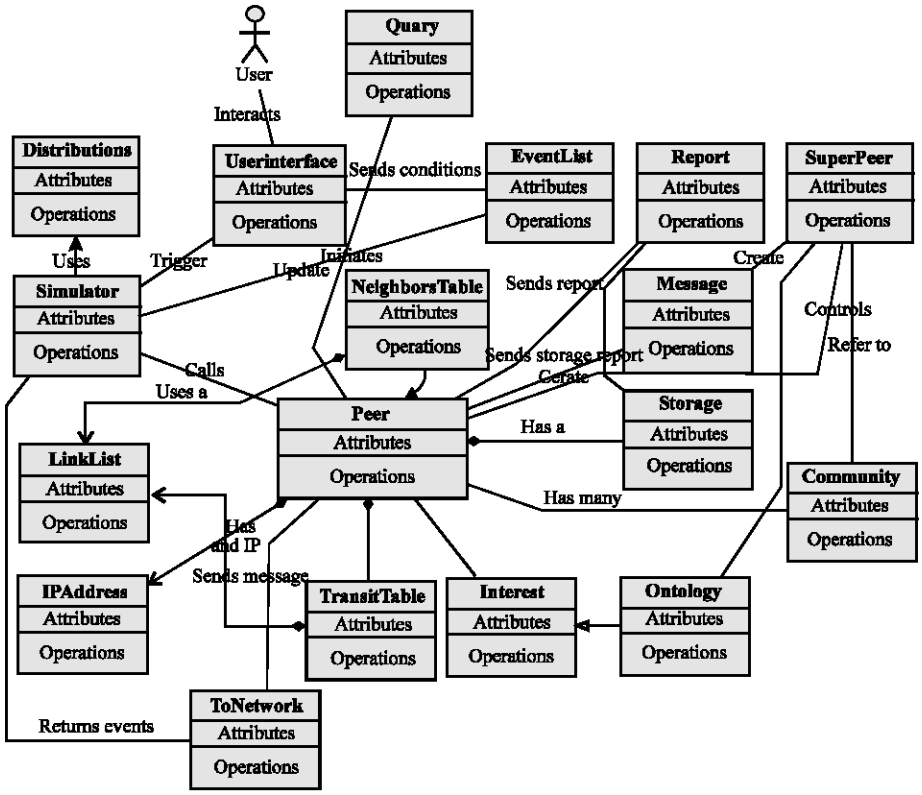

Fig. 1: The structure of the simulator

other nodes based on some routing strategies. Routing strategies usually use some routing information which is stored in the routing table of each peer. In a semantic overlay, this routing information may semantically identify some related nodes.

Peers may create new shortcuts to other nodes, if they receive proper answers from a particular node many times. Shortcuts may also be omitted if a peer can find a closer neighbor with good source of information or it is not useful for a long time.

The complete representation of this model has been introduced in (Modarresi *et al*., 2008).

The implementation model of the simulator is shown in Fig. 1. In this simulator each important entity is implemented by a corresponding object of the respective entity class. Super peer, community, ontology and interest classes are defined as parameters of the model. If they are not defined, a random model can be simulated easily. The combination of different simulation parameters can define models from completely random to social semantic network.

This is a flow-based simulator that considers a chunk of bytes, called a message, instead of packets in network stack. An overlay is models as a graph over the physical network; therefore what is represented in the model is an end-to-end connection between two end users. Distance among nodes is presented by the number of physical hops assigned randomly to each end-to-end connection; and network delay is the multiplication of average delay for a hop, assigned randomly to a connection, per number of physical hops created the end-to-end connection.

EVALUATION FUNCTIONS AND PARAMETERS

As it is mentioned by Schmitz and Loser (2006) the performance of a system can be measured by feeding some input parameters to the system and studying the behavior of the output parameters. This can be represented by a function F, as it is mentioned by Schmitz and Loser (2006) as follow:

$$(i_1, i_2, \ldots, i_n) \xrightarrow{F} (o_1, o_2, \ldots, o_m)$$

This function maps input parameters of the model to output behavior of the model. The function F is all settings and algorithms used in the system. In each type of model, different input parameters can be used. Some of these input parameters are general and used in all types

of models; however, some of them like number of communities, sub-communities or peer's interest belong to the some specific type of model.

Modeling the evaluation function: Function F shows the internal behavior of the simulator to change input parameters to output results. This function can be modeled with many sub-models expressed as follow:

- **Content model:** Shared contents in P2P networks are not equally distributed among the nodes. Many studies like (Iamnitchi, *et al*., 2004) have shown that content is usually distributed among nodes by Zipf-like distribution. There are two main methods of distributing data among the nodes. One is using distribution functions like Zipf and the other is crawling and extracting data from real network and loading it to the model. The proposed model uses the first method. Different distribution functions can control content distribution, although the most real one is Zipf
- **Network model:** This simulator can construct different topologies from fully random to social semantic overlay expressed earlier. Social samentic simulation model has also the characteristics of small-world network (Watts and Strogatz, 1998). It shows high cluster coefficient and low characteristic path length (Modarresi *et al*., 2008). These values are not comparable with a random network which has a low cluster coefficient. On the other hand, the number of neighbors has a direct effect on network traffic and success rate. The fixed number of neighbors is not usual in real networks. Such a network defines a grid topology with high path length. This simulator can define the number of neighbors either as a fixed number or as a distribution function. Zipf distribution is more likely to real world conditions
- **Routing model:** Different routing algorithms can produce different results for specific input parameters. In addition, the efficiency of a routing algorithm depends on the topology of the network model. For example, a flooding-based algorithm can produce a high amount of traffic in a network whose peers have high number of neighbors. Same algorithm can produce less traffic and better results when the nodes are organized in an efficient overlay with few numbers of neighbors
- **Query model:** In an interest-based model, nodes usually pose queries about their interests, but there is little exact information to show which peer poses which queries. Cholvi *et al*. (2004) have shown that queries follow Zipf distribution. This simulator can produce linear and zipf distributions for query posing. Moreover, the ratios of queries which are related to the interests of peers are controllable

Input parameters: The main input parameters considered for a simulation model are as follow:

- **Number of nodes:** This parameter shows the scalability of the system. It also affects the output result. Certainly, finding a piece of data in a small network is easer that a large network, otherwise nodes are organized in an efficient overlay
- **Number of shared data:** This parameter also shows the salacity of the system. In a large network, if there is little amount of shared data and they are distributed randomly, the probability of popular data will increase. In contrast, this probability will decrease in a large network. It is obvious that finding popular data in a small network is much easier than unpopular data in a large network
- **Network topology related parameters:** Some input parameters are related to a specific topology. For example, in a social semantic system the type of ontology, as well as the number of communities and sub-communities are some of these types of input parameters

Output parameters: Different kinds of output parameters can be gathered, but output parameters which are measured depend on the application scenario. Some of these parameters are more general than others and can be helpful in each type of scenarios. Some of them are measured by this simulator as follows:

- **Sent messages:** This parameter reveals many things about the system. It shows the created network traffic. Few messages mean little network traffic. It also shows the effectiveness of the routing algorithm. A good algorithm can find results with fewer messages. It can also show the effectiveness of a particular algorithm over a specific overlay or how an overlay can affect the routing algorithm. One algorithm can produce better results on a particular overlay
- **Drop messages:** This parameter can show indirectly the cost of search. If more messages are dropped, it means that fewer answers will be retrieved

- **Repetitive messages:** It shows in what extend the overlay has been organized effectively and how peers can control loops in the structure of the overlays
- **Success rate:** It shows how many queries find proper answers in the network. Large networks with less efficient overlay reduce this value
- **Number of searched nodes:** This parameter shows the effectiveness of the routing algorithm and in what extend the algorithm forwards the messages toward the nodes that have higher probability to answer a query
- **Average number of hops:** This parameter also shows the efficiency of the routing algorithm and the overlay. Efficient algorithm and overlay produce less average number of hops to answer a query
- **Recall:** This parameter shows the proportion of retrieving answers to total relevant answers in the network. The higher recall rate means the higher quality of service

DATASET

One important part of each simulation study is the validity of data used in the simulation. According to (Sargent, 2005) data validity is defined as the correctness and adequateness of data which are used in model building, testing and experiences. ACM classification system (ACM, 1998) has been used for the ontology and DBLP data file used for data to match this ontology. These two data sources have been used in similar studies like Haase *et al*. (2004b) and Schmitz and Loser (2006) and their validity has been approved. In order to change the data source, a proper ontology must be provided and used in the simulator.

SIMULATION AND RESULTS

Three different overlays are defined and simulated with the designed simulator. In all of these structures 1000 nodes are created in each overlay. The first overlay is a random model, in a way that each node may connect to any other node in the system, without considering any criteria. In the second overlay, an interest-based model is simulated. In the model, the nodes establish a connection with other nodes which provide proper answers many times. In this way, all the nodes which have similar interests create a cluster. The third overlay is the one which has been explained in social semantic simulation model. This overlay uses social network concepts in its structure. In all experiments, DBLP bibliographic data file is used. In the third experiment, ACM classification (ACM, 1998) is also used as the ontology of the system. In order to provide semantic data, RDF statements are used to express the relativity between different pieces of data. In all these experiments Zipf distribution is used for number of neighbors of each node. The value shown in each chart as the number of neighbors is the maximum value that the distribution function can produce.

The number of sent messages during the simulation is shown in Fig. 2. These messages include queries, answers, control messages and all messages created during query routing process. Among all three overlays, the proposed overlay creates and sends the least number of messages. The reason is that peers in community-based overlay are clustered and configured more efficiently than other overlays. All the peers which have similar interest are grouped in one community; hence, for a particular query, a particular community is just searched.

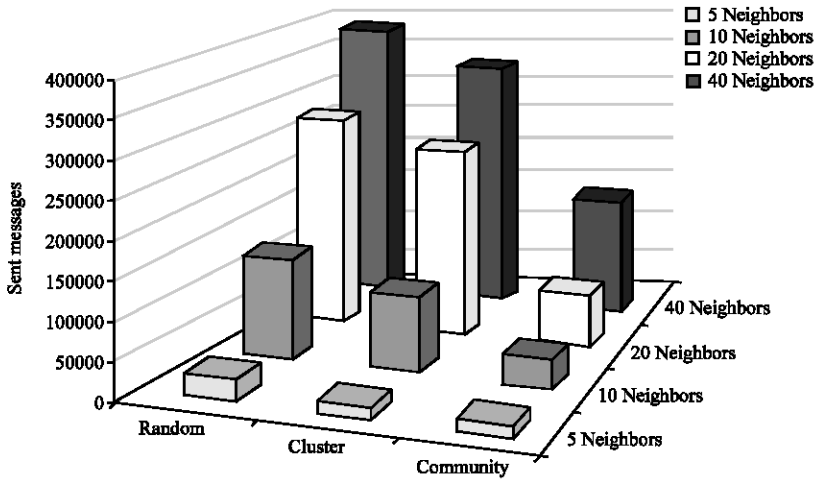

Fig. 2: Number of sent messages during simulation for three different overlays

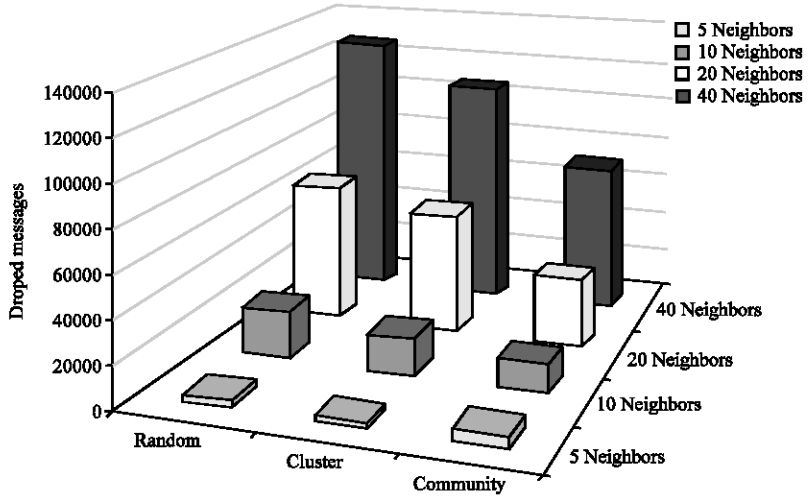

Fig. 3: Number of dropped messages during simulation for 3 different overlays

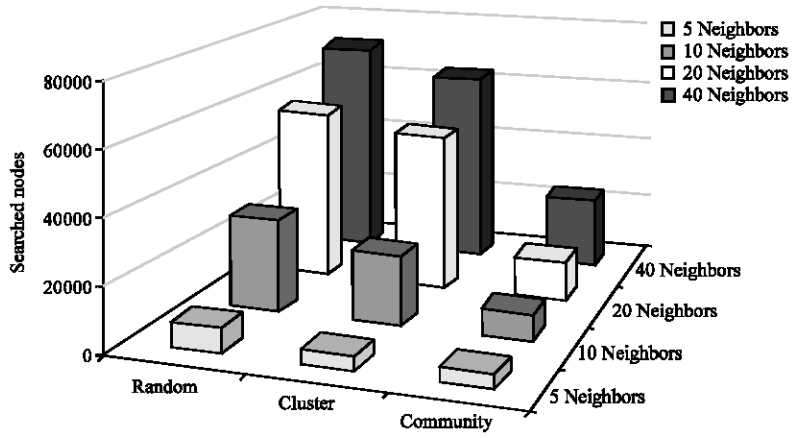

Fig. 4: Number of searched nodes in each overlay

The number of dropped messages is another parameter to show the performance of a model. It can also show the effectiveness of the routing process. In most computer networks, many routes may exist between two nodes. The number of dropped messages shows how many repeated messages are received for a particular query. If a routing algorithm can discover such paths and avoid them, the number of dropped messages decreases. However, the structure of the overlay has a direct effect on this parameter. Figure 3 shows the values of this parameter for the running experiments.

The number of searched nodes for answering queries during the simulation is another parameter which can be considered. The best way for answering a query is a direct search. In these methods answers are retrieved directly. When the direct search is not possible, undirected search with blind scheme is often used. In contrast to direct search, undirected methods must search many nodes to find proper answers. The more searched nodes means the more network traffic and resource consumption. This parameter can show how effectively a routing algorithm can find an answer and how different overlays can affect on a specific routing algorithm. The number of total nodes during the simulation is shown in Fig. 4. Certainly, this parameter must be considered with other parameters like success rate. If the success rate is high and the searched nodes are low, it means that routing algorithm can forward queries to the nodes which have higher probability to answer a particular query.

In Fig. 5, the average path length for answering queries during simulation is shown. Smaller average value means shorter characteristic path length in the model and better organization nodes in the overlay. Moreover, it shows that in how many steps in average, a node is reachable.

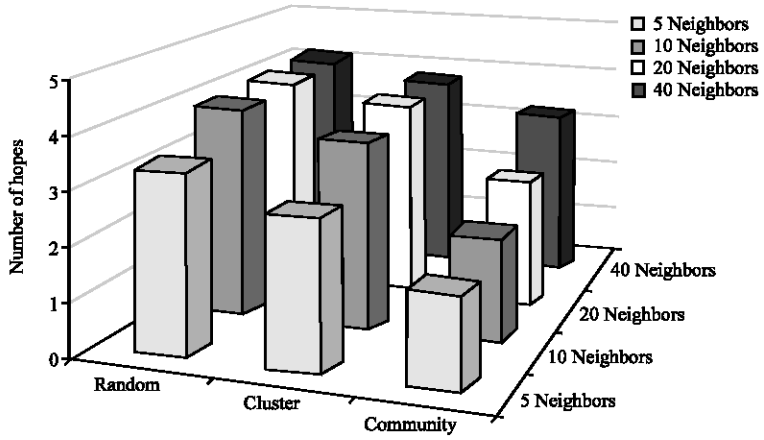

Fig. 5: Average length for answering queries per each overlay

DISCUSSION

The presented simulator can simulate various types of P2P overlays from completely random to social semantic network. This is in contrast to other simulators like Peersim or P2Psim which can implement a collection of previously developed models. Table 1 (Brown and Kolberg, 2006) shows supported protocols for each corresponding simulator.

Considering different input parameters in the proposed simulator let the users add different capability to their models by a user interface during the definition of the overlay. This simulator provides an environment to compare different types of overlays in similar conditions. However, this is nearly impossible in other previous simulators like Narses, Peersim or P2PSim which can simulate few specific groups of overlays. Moreover, most of these simulators do not have any capability to define semantic overlays. Providing such an environment can be a proper solution to one of the main problems in the P2P area. Since different overlays are studied with different inputs, outputs, datasets and particularly specific designed simulators, the comparison of models to show the efficiency and performance of them will be very hard or even impossible (Schmitz and Loser, 2006). It must be mentioned that, the proposed simulator has been designed for unstructured P2P systems and cannot support any structured P2P protocol like CAN (Ratnasamy *et al*., 2001) and Chord (Stoica *et al*., 2001).

The proposed simulator also shows comparable scalability with highly scalable simulators like Peersim. It can simulate nearly one million nodes when simple overlays without semantic are simulated. This number decreases when either semantic overlays are studied or lots of data for sharing in the system are loaded into the storage of each node. This is due to the fact that, in the first case, Sesame query engine consumes extra memory and in the second case much memory is used to store shared data. Table 1 also shows the maximum number of nodes for each corresponding simulator. As it is observed, the proposed simulator is comparable with PeerSim and more scalable than other mentioned simulators in Table 1, when a simple overlay is simulated. It is also as good as many of them when a semantic overlay is considered.

Table 1: P2P protocols and maximum nodes supported by each simulator

Simulator	P2P protocols	Max. nodes
Narses	None	600
3LS	Gnutella	<1,000
NeuroGrid	Gnutella, NeuroGrid, Pastry, FreeNet	300,000
PeerSim	Collection of internally developed P2P models	$>10^6$
P2PSim	Chord, Accordion, Koorde, Kelips, Tapestry, Kademlia	3,000
Omnet++	None	1,000
NS2	Gnutella	N/A

N/A: Not available

In order to show the capability of the proposed simulator, three different types of overlays have been simulated in previous section. In addition, this simulator can captures different output variables during simulation to provide more flexibility for comparison. Although, many P2P simulators have been designed, in present knowledge, neither of them can simulate both social semantic and simple overlay simultaneously.

CONCLUSION

In this study, a general model for a social semantic P2P overlay was introduced. According to this model, a simulator with the ability to simulate many different overlays was designed. Moreover, many general output

parameters applicable for different types of overlay were introduced. The result was easier comparison of different overlays with the same dataset and input parameters in a similar environment.

It is attempted to change the structure of the implemented simulator to many APIs for easier integration with other applications and simulators as the future study.

REFERENCES

ACM, 1998. The ACM Computing Classification System: Association for Computing Machinery. ACM, New York, USA.

Ahn, J.S. and P.B. Danzig, 1996. Packet network simulation: Speedup and accuracy versus timing granularity. IEEE/ACM Trans. Networking, 4: 743-757.

Brown, A. and M. Kolberg, 2006. Tools for peer-to-peer network simulation. The Internet Engineering Task Force (IETF). http://tools.ietf.org/html/draft-irtf-p2prg-core-simulators-00.

Cholvi, V., P. Felber and E. Biersack, 2004. Efficient search in unstructured peer-to-peer networks. Proceedings of the 60th Annual ACM Symposium on Parallelism in Algorithms and Architectures, Jun. 27-30, Barcelona, Spain, pp: 271-272.

Crespo, A. and H. Garcia-Molina, 2005. Semantic overlay networks for P2P systems. Proceedings of the 3rd International Workshop on Agents and Peer-to-Peer Computing, Jul. 19, New York, USA., pp: 1-13.

Giuli, T.J. and M. Baker, 2002. Narses: A scalable flow-based network simulator. http://arxiv.org/abs/cs.PF/0211024.

Haase, P., R. Siebes and F. Van-Harmelen, 2004a. Peer selection in peer-to-peer networks with semantic topologies. Proceedings of the 1st International IFIP Conference on Semantics of a Networked World, 2004, Paris, France, pp: 108-125.

Haase, P., B. Schnizlera, J. Broekstrab, M. Ehriga and F. Van-Harmelenb *et al*., 2004b. Bibster-a semantics-based bibliographic peer-to-peer system. Web Semantics: Sci. Services Agents World Wide Web, 2: 99-103.

Iamnitchi, A., M. Ripeanu and I. Foster, 2004. Small-world file-sharing communities. Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, Mar. 7-11, Hong Kong, pp: 952-963.

Khambatti, M., K.D. Ryu and P. Dasguptaand, 2004. Structuring peer-to-peer networks using interest-based communities. Proceedings of the 1st International Workshop, Databases, Information System and Peer-to-Peer Computer, LNCS., 2944, Sept. 7-8, Berlin, Germany, pp: 48-63.

Klampanos, I.A. and J.M. Jose, 2004. An architecture for information retrieval over semi-collaborating Peer-to-Peer networks. Proceedings of the 2004 ACM Symposium on Applied Computing, Mar. 14-17, Nicosia, Cyprus, pp: 1078-1083.

Modarresi, A., A. Mamat, H. Ibrahim and N. Mustapha, 2008. A social network Peer-to-Peer model for peer clustering. Proceedings of the ITSim 2008 International Symposium on Information Technology, Aug. 26-28, Kuala Lumpur, Malaysia, pp: 1-7.

OPNET Technologies Inc., 2008. OPNET. http://www.opnet.com.

Ratnasamy, S., P. Francis, M. Handley, R. Karp and S. Schenker, 2001. A scalable content-addressable network. Proceedings of the ACM SIGCOMM Conference on Application Technology, Architectures and Protocols for Computer Communaction, Aug. 27-31, San Diego, CA. USA., pp: 161-172.

Sargent, R.G., 2005. Verification and validation of simulation models. Proceedings of the 37th Conference on Winter Simulation, Dec. 4-7, Florida, USA., pp: 130-143.

Schmitz, C. and A. Loser, 2006. How to model semantic Peer-to-Peer overlays? Proceedings of the GI Jahrestagung (Informatik 2006), Oct. 12-19, Dresden, Germany, pp: 1-8.

Stoica, I., R. Morris, D. Karger, M.F. Kaashoek and H. Balakrishnan, 2001. Chord: A scalable peer-to-peer lookup service for internet applications. Proceedings of the ACM SIGCOMM Conference on Application, Technologies Architectures and Protocols for Computing Communications, Aug. 27-31, San Diego, CA., USA., pp: 149-160.

Sun, Y., L. Sun, X. Huang and Y. Lin, 2006. Resource discovery in locality-aware group-based semantic overlay of peer-to-peer networks. Proceedings of the 1st International Conference on Scalable Information Systems, May 30-Jun. 1, ACM, New York, USA., pp: 1-40.

Tempich, C., S. Staab and A. Wranik, 2004. REMINDIN': Semantic query routing in Peer-to-Peer networks based on social metaphors. Proceedings of the 13th International Conference on World Wide Web, May 17-20, New York, USA., pp: 640-649.

Ting, N.S. and R. Deters, 2003. 3LS-a Peer-to-Peer network simulator. Proceedings of the 3rd International Conference on Peer-to-Peer Computing, Sept. 1-3, Linköping, Sweden, pp: 212-213.

Watts, D.J. and S.H. Strogatz, 1998. Collective dynamics of' small-world' networks. Nature, 393: 440-442.