The Structure of an Instant Messenger Network and its Vulnerability to Malicious Codes

Cheryl D. Morse and Haining Wang

Department of Computer Science
The College of William and Mary
Williamsburg, VA 23187
Email: {cddelo,hnw}@cs.wm.edu

ABSTRACT

The use of Instant Messengers (IM) has dramatically increased over the past few years. An IM network can be viewed as a graph, in which the nodes are the systems running IM and a link exists between two nodes if they are on each other's buddy list. Based on the trace collected at the William and Mary gateway, we observe that an IM network has a high degree of node clustering, a short average path length, and its node degree distribution following power-law. The IM graph can be modeled as a scale-free network. The scale-free nature of the IM network implies that if a virus were to infect the IM network, the virus would spread very quickly and would prevail. Simulations show that an IM virus could saturate the entire IM network in as few as six time steps.

1. INTRODUCTION

In recent years, more and more people use instant messengers (IM) for on-line communication over the Internet. These on-line users form an IM network, in which malicious codes may spread across the network. Investigating the topology of the IM network, determining whether it is a scale-free graph or a random graph, can reveal the communication characteristics of the IM network. Since America Online instant messenger (AIM) is the most widely used IM (with over 100 million users worldwide) [3], we use AIM as the object of this study.

Each user of AIM has a buddy list, associating with whom it like to communicate. The buddy list helps the user to know the status of its buddies, i.e., on-line or off-line. AIM messages are not sent directly from user to user, rather they are sent to an AIM server that forwards the messages to the recipient. This implies that each message, which is sent from one user on a local area network (LAN) to another user on the same LAN, must first leave the LAN and then re-enter the LAN. Figure 1 illustrates the path that messages traverse from source A to destination B.

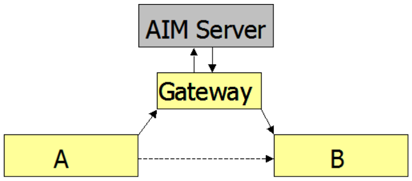

Figure 1: How a message is transmitted from user A to user B

2. TRACE-BASED MEASUREMENTS

In order to study the topology of an AIM network, we analyze the trace of AIM traffic collected at the gateway of the William and Mary campus network. The AIM trace consists of all inbound TCP messages with source port of 5190. Specifically, we single out the packets that reveal the structure of the AIM network. Note that when users login, log-off, or change their status, the AIM server sends the corresponding information to AIM clients. We utilize these information to construct the AIM graph. The topology metrics we used include the average shortest distance, clustering coefficient, and node degree distribution.

Based on the topology metrics mentioned above, we characterize the AIM network inside the William and Mary campus. At working hours on average, there are 2,309 IM users (nodes) inside the William and Mary campus, connecting to the AIM network simultaneously. Each node is connected to many other nodes at William and Mary varying from 1 to 150. The average node degree is 6.3. Note that the nodes may also connect to the nodes outside the campus network. However, it is impossible to determine the degree of the outside nodes accurately. Thus, our AIM network does not include the outside nodes, and we ignore these connections in the analysis.

Figure 2 shows the degree distribution of the nodes on the AIM network in logarithmic scale. The degree distribution can be approximated by a power law with an exponent of -1.66. This follows the property of a scale free network. The average shortest distance between any two nodes in the AIM network is 3.69 hops. To further validate if the AIM network, with 2,309 nodes and its average distance of 3.69 hops, follows the characteristics of a scale free graph, two scale free networks with 2,309 nodes are generated by using preferential attachment [1]. Table 1 shows the mean degrees, the average shortest distances and the clustering

Network	Mean Degree	Average Path Length	Clustering Coefficient
Generated Scale Free (I)	5.99	3.67	.0066
AIM Network	6.30	3.69	.0068
Generated Scale Free (II)	7.99	3.35	.0075
Random Graph	6.30	4.21	.0027

Table 1: Average path length, clustering coefficient, and mean degree of AIM and scale free networks

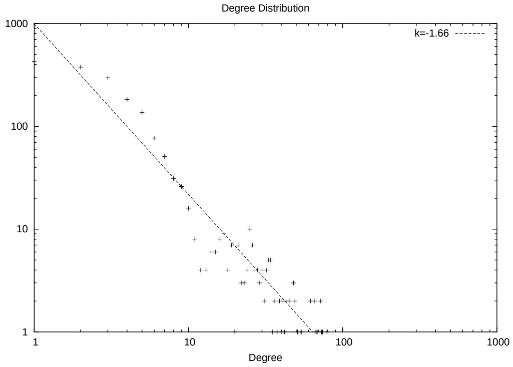

Figure 2: Degree distribution of AIM network with logarithmic scale.

coefficients of the AIM network and the two generated scale free networks, respectively.

As shown in Table 1, the AIM network has a relatively high degree of node clustering. Its value is close to that of the generated scale free graph with a mean degree of 5.99. In contrast, the clustering coefficient of the corresponding random graph is only 0.0027, which is significantly smaller than those of the scale-free graphs and that of the AIM network.

In summary, the node degree distribution of the AIM network follows a power law, the average shortest distance of the AIM network is short, and the AIM network has a high degree of node clustering. The AIM network meets all three requirements of scale free graphs, therefore, the AIM network can be approximately modeled as a scale free graph.

3. VULNERABILITY TO VIRUS

Malicious codes have the potential to spread quickly across a scale-free network and the ability to prevail. The standard model that is used to study the spread of a computer virus is the SIS epidemiological model [4]. In this model, the spreading rate of a virus can be computed as the rate, in which a healthy node becomes infected if at least one of its buddies is already infected. If the effective spreading rate is above a critical threshold, the virus will continue to spread and will never die out [2]. Because of the existence of hubs, the critical threshold of a scale-free network is zero [4]. This is due to the fact that a hub has a greater probability of becoming infected with a virus. Once it becomes infected, it will spread the virus to a significant portion of the network. With a critical threshold of zero, it is impossible for a virus to die out, it will prevail even if the spreading rate is extremely low [2].

While the AIM network is scale free, it is not clear how quickly a virus could spread across the AIM network. To estimate the time it would take to infect all nodes in the AIM network, we simulate the spread activity of a virus that saturates a network with the similar structure of the AIM network on the William and Mary campus.

Clock	Total Infected
0	1
1	3
2	171
3	1202
4	2274
5	2309

Table 2: Speed at which a virus spread across AIM network

Here we use the generated scale free network with a mean degree of 5.99 as the input to the simulation. At clock time 0, we assume that one node is infected with a virus. The initially-infected node then spreads the virus to other nodes it connected, and these nodes become infected with the virus at time 1. Table 2 shows the scenario in which the initially-infected node has the smallest degree. Even so, the saturation time is very short. Every node has been infected just after six time steps. Note that the total number of nodes here is 2,309.

4. CONCLUSION

We studied the structure of an IM network inside the William and Mary campus. We found that the IM network has a high degree of node clustering, a short average path length, and a node degree distribution that follows a power law. The IM network can be classified as a scale-free network. Since scale free networks have a critical threshold of zero, it would take an IM virus a matter of seconds to infect every node inside the network. Thus, the IM virus has the potential to wreak havoc across the IM network.

5. REFERENCES

[1] Albert-Laszlo Barabasi. *Linked: The New Science of Networks*. Perseus Publishing, Cambridge, Massachusetts, 2002.

[2] Zoltan Dezso and Albert-Laszlo Barabasi. Halting viruses in scale-free networks. *Physical E*, 65, 2002.

[3] American Online. AOL instant messenger - it's free. http://www.aol.co.uk/aim/.

[4] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14), 2001.