The Application of Graph Theory in BitTorrent Protocol and Older P2P Protocols

Danny Andrianto 13510011
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
danny.andrianto@students.itb.ac.id

Abstract—This paper will be discussing BitTorrent protocol and other P2P protocols which predeces it. Author will explain how BitTorrent works and compare it with the older P2P protocols by using graphs to show the connection between computers in each networks. From the graphs, author will analyze each protocol and determine its strength and weaknesses.

Index Terms—directed graphs, peer to peer file sharing, BitTorrent, protocols

I. GRAPH THEORY

1.1 Definition

Graph is an abstract representation of discrete objects and the relation between them. Graphs consist of verticees and edges. Vertice is the representation of the objects which is usually drawn as a dot or circle while edge is the representation of the relations which is usually drawn as lines connecting the verticees.

Graph is the collection of sets of vertices and edges which is written as:

Graph $G = (V,E)$
V=non-empty set of vertices
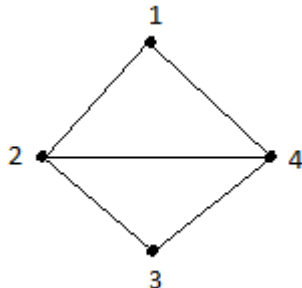$= \{v_1, v_2, ..., v_n\}$
E=set of edges
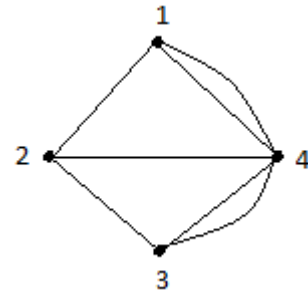$= \{e_1, e_2, ..., e_n\}$

1.2 Types of Graphs

1.2.1 Simple Graphs

Graphs which don't have any multiple or parallel edges, loops, and direction.
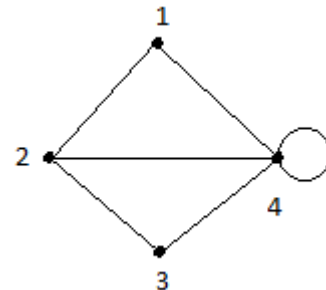

1.2.2 Multigraphs

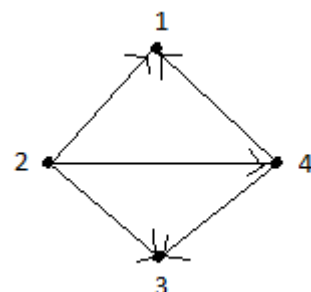Graphs which have multiple or parallel edges but no loops.

1.2.3 Pseudographs

Graphs which have loops.

1.2.4 Directed Graphs

Graphs whose each edges has direction orientation.

1.3 Basic Graph Terminology

1.3.1 Adjacent

Two vertices are adjacent if both of them are directly connected by an edge.

1.3.2 Degree

The degree of a vertex is the number of edges incident with it.

1.3.3 Initial and Terminal or End Vertex

If (u,v) is a directed edge, the vertex u is called the initial vertex of (u,v) and v is called the terminal or end vertex of (u,v).

1.3.4 In-Degree and Out-Degree

The in-degree of a vertex u is the number of edges with u as their terminal vertex and the out-degree is the number of edges with u as their initial vertex.

II. BitTorrent Protocol

BitTorrent is a type of peer-to-peer file sharing (P2P) protocol. P2P is different form the traditional file downloading in the way that users use a certain software to locate a computer which has the file that the user wants and download the file from the computer (peers), instead of a server. However, some users tend to disconnect after finisihing a download thus preventing other users from obtaining the files (leech). This is the problem which Bram Cohen (author of BitTorrent protocol) wished to solve.

Bram Cohen began writing the protocol in April 2001 and finally released the first BitTorrent client (BitTorrent) on 2 July 2001. He wrote the first BitTorrent implementation in Phyton as a free software. BitTorrent quickly gained its fame because its abilty to quickly share large files online. Finally in 22 September 2004, Bram formed BitTorrent, Inc. to maintain the protocol with his brother Ross Cohen and business partner Ashwin Navin. Right now, there are lots of BitTorrent besides the original BitTorrent software itself. Some of the most notables are uTorrent, Tixati, Vuze, and Deluge. It has been estimated that BitTorrent protocol accounted for roughly 43% to 70% of all internet traffic.

BitTorrent protocol uses a central server called tracker to track computers running BitTorrent client with a complete file (seed) or portion of the file which the user wants and identify a swarm. A swarm is a group of computers which simultaneously sending or receiving the same file. To maximize the transfer speed, BitTorrent then gathers the pieces of the file and download them simultaneously from different computers.

BitTorrent also uses the tit-fot-tat principles which means the more files the user uploads, the more download rate the user gets.

III. Older P2P Protocols

In this part, author will describe other P2P protocols which were popular prior to BitTorrent. They are the protocols which are used by the first generation P2P software, Napster, which is also the pioneer of P2P file sharing and the second generation P2P softwares, Gnutella and Kazaa (BitTorrent being the third generation).

3.1 Napster

Napster was the first P2P file sharing internet service which focused on sharing mp3 files. It started in 1999 when co-founder Shawn Fanning talked to his friends about the dificulties of finding MP3 back then. He then started to develop an idea to create a program which has the functions of search engine, file sharing, and IRC (internet relay chat). Napster's technology allowed users to easily share his/her mp3 files with other users (for free obviously). However, it resulted in massive violation rights and was finally shutdown in July 2001 because of its copyright issues. Napster would then reopened as a pay service.

Napster was the first service to allow the users to download files from other computer, not the server. Thus, Napster's central server worked as an index machine to list all available users and the mp3 file which the user requested. A Napster user only had to type the name of the mp3 file he/she was looking for and the program will list the mp3 file from various computers which have the file the user requested. When the user decided to download a file, the server will establish a connection to the computer which has the file the user requested and began downloading. Easy and simple, that's the reason it was so popular.

3.2 Gnutella

Gnutella is the first decentralized P2P protocol. Unlike Napster, there is no central database that knows all the files available on an Gnutella network. Instead it uses hostcahes which are computers that stays on 24 hours a day to provide initial connection for clients and a list of available computers. There are also many different clients to access the Gnutella network such as BearShare, Gnucleus, and LimeWire. Therefore it's more difficult to shut down Gnutella because there is no one server or software controlling the entire system.

Since there is no central database in Gnutella network, all the computers on the network tell each other about available files using a distributed query approach. It works by sending a request to other Gnutella machine it knows about (in most cases hostcaches). The Gnutella machine will then search if the requested file is in the hard disk. If it's not, then the computer will send another request to other Gnutella machines it knows about. When the requested file has finally been found, the computer which hosts it will send back the file's name and its IP address to the user's computer. The user can then download the file from the computers available on the network.

3.3 FastTrack

FastTrack is a P2P protocol which was used by Kazaa. It was created and developed by Bluemoon Interactive (which later created Skype). It was then introduced in

March 2001 by Consumer Empowerment. FastTrack protocol uses encryption and wasn't documented by its creators. Any rights to use this protocol will have to go through the company Sharman Networks. There has been many attempts to reverse engineer FastTrack protocol, however it only gives a rough line on how the FastTrack protocol works. However, the information might be outdated because there may have been recent changes.

The most notable technology used by FastTrack is supernodes. A supernode can consists of ten to several thousands of nodes. These nodes can act as a directory server for other clients and has no centralized control whatsoever. The supernode is then connected to other supernodes just like decentralized P2P network. Besides supernode, FastTrack also has bootstrapping nodes which act as the initial connection for clients to determine whether it is qualified enough to be a supernode. If it's qualified to be a supernode, the bootstrapping node will provide the IP adresses of other supernodes. If not, then it will only be provided with one supernode IP address. Ithe user's computer will first search the files he/she wants from the supernodes he/she part of before sending the request to other supernodes and ordinary nodes.

IV. THE APPLICATION OF GRAPH

In this next part, author will describe how the aforementioned protocols work with graphs.

For instance, a user (U) wishes to download a movie with a BitTorrent client. First the user will send a request to a tracker (T) to find users who download the same file which uses the same BitTorent client and a seed (the computer which uploads a completed file).
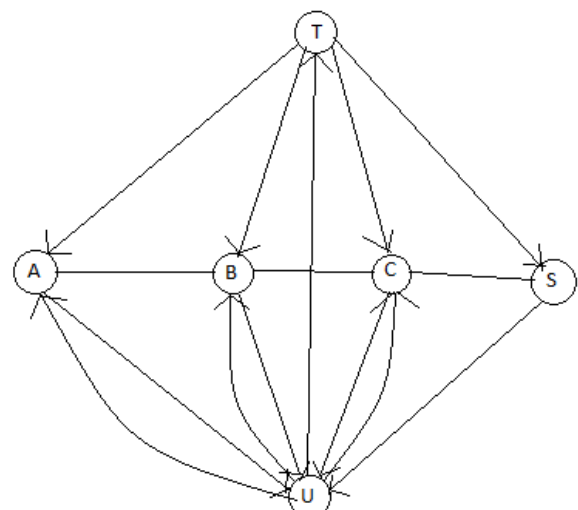

Then, the tracker will identify a swarm which consits of computers downloading the same thing and seeders (S).
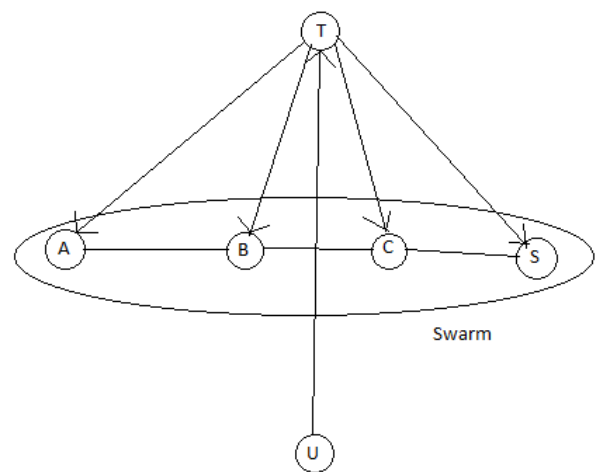
The user's computer then downloads different pieces of the file simultaneously from different computers


.
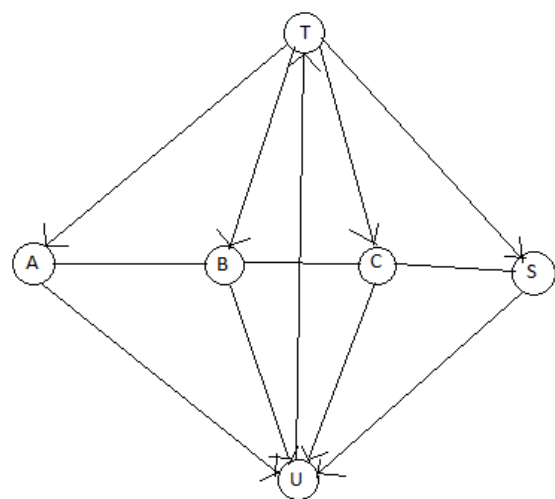
While downloading, the user's computer will also upload some pieces of the file to other users which are still downloading.

Once the download is complete, the computer becomes a seed.
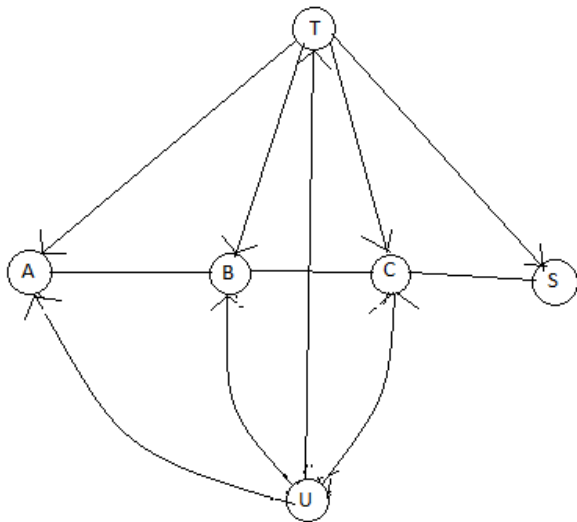
The more computers enter the swarm, the more connection established between computers, and the faster the transfer rate become (A swarm with 9 computers will be faster then 5 computers).
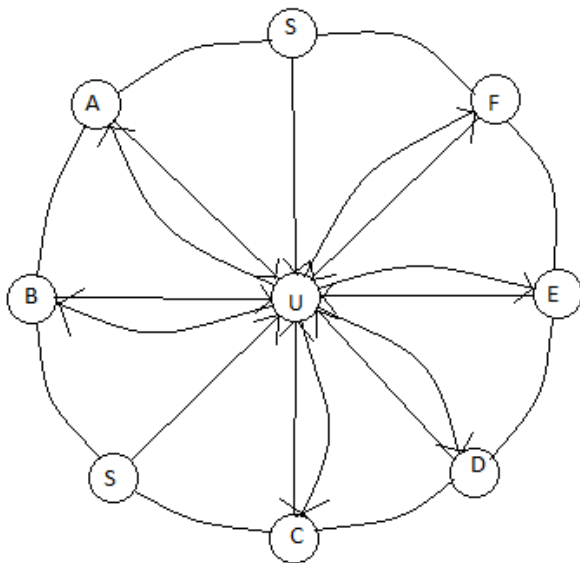

From the graphs, we can see that BitTorrent uses a rather complex protocol which allows users to trade files and download simultaneously from different hosts. This method proved to be successful in increasing the transfer rate. However, it is relatively more difficult to share our own file because we will have to create our own .torrent file for the files we want to share. In order to download a file, we also have to find the right .torrent file and there is a chance that the file might be fake or infected. Fortunately, there are many sites which host good torrent files such as isohunt, piratebay, and torrentz.
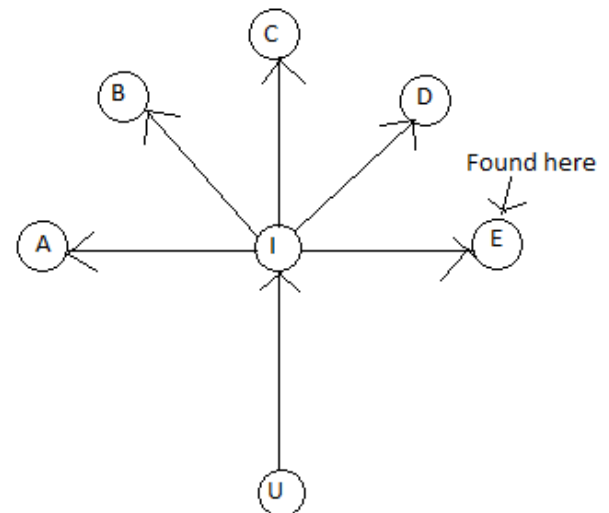
In Napster, the user (U) will first connect to a Napster central index server (I).
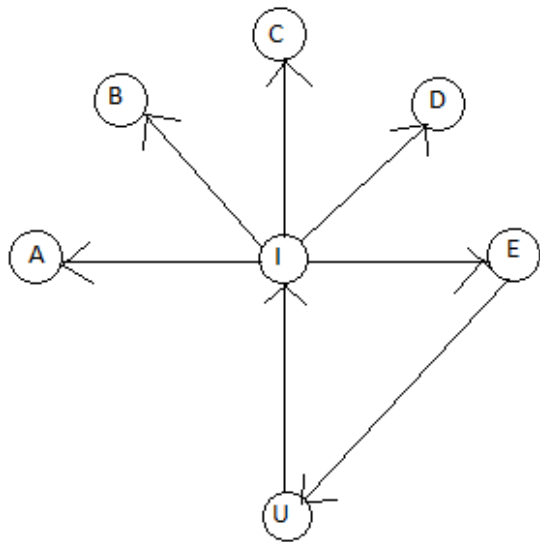

This server is connected to other computers running Napster software and will provide the list of available hosts (A, B, C, D, E, F).
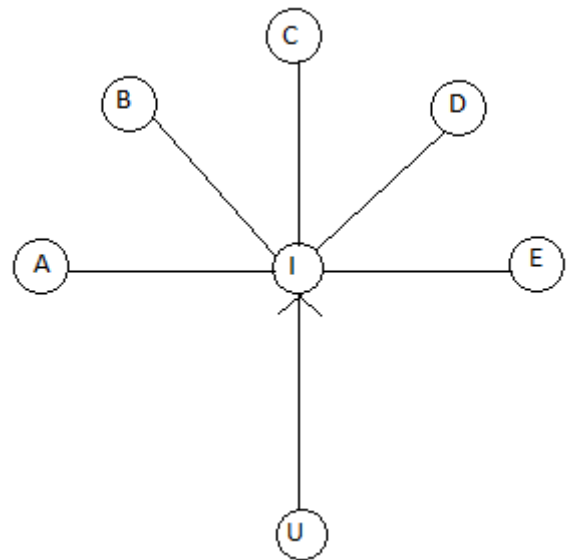
The server then search the computer with the file that the user is searching for.

Finally, the server will establish a connection with the computer providing the file and the file transfer starts.

We can clearly see that the protocol which Napster uses is a simple one. It gives users easy access to the files thay want. However, we can also see that there is one major weakness of this system. By having one central server, it is easy to shutdown the system. This is also the reason why peer-to-peer file sharing services nowadays use decentralized system. By having no central server which controls almost everything it is almost impossible to shutdown the system.

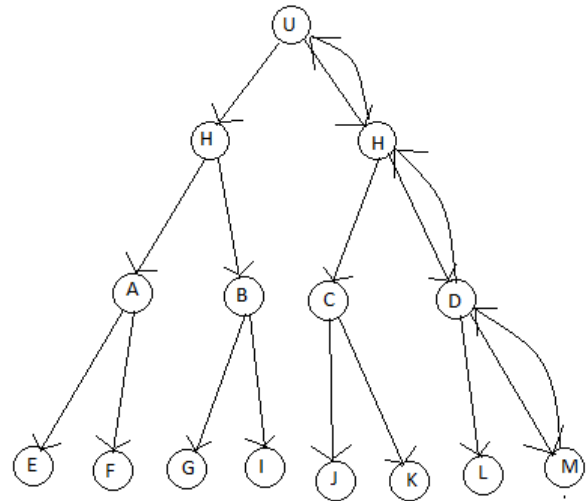In Gnutella, the user (U) will first be connected to several hostcaches (H).
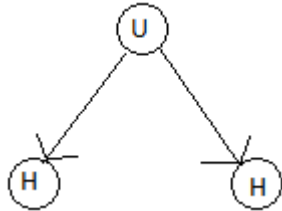
Then the hostcahes will start sending requests to other computers they know.

Found here
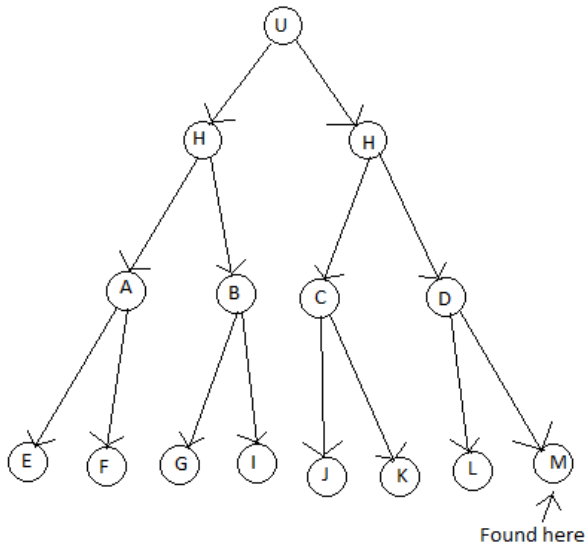
When the computer whch hosts the file the user is looking for is found. It will send back its information and the file transfer will start.


Gnutella uses decentralized and it proves to be more successful then its predecessor Napster. However, this kind of peer-to-peer connection can easily have problems with leechers. Leechers can really limit the performance of this system because replies follow the same path it was sent, and if a computer disconnected the path might be lost.

In FastTrack, the users computer (U) is first connected to a bootstrapping node. The node then determine whether the computer is qualified to be a part of a supernode (S). If it is the user's computer will be a part of a supernode and connected to other supernodes.
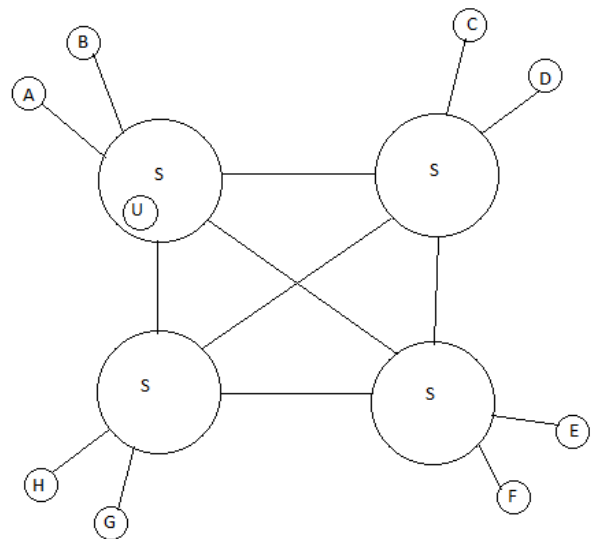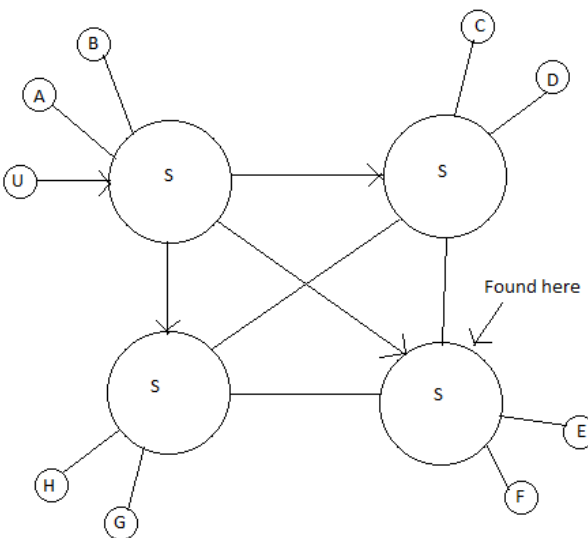

If it isn't the user's computer will only be connected to a supernode.
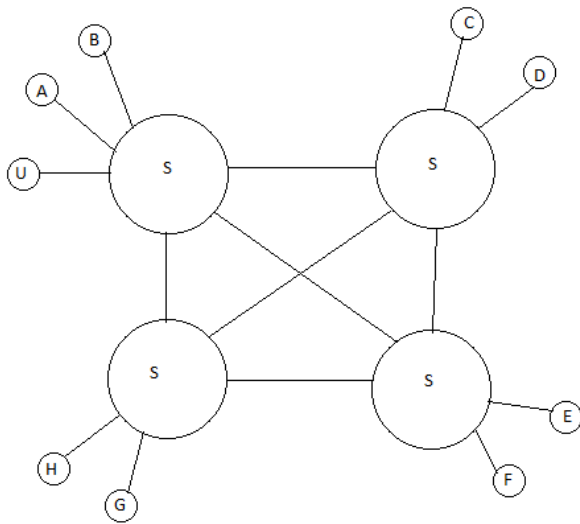
The next step is almost the same as Gnutella in which it will search for the file in the. It works the same way whether the user's computer is part of a supernode or not.
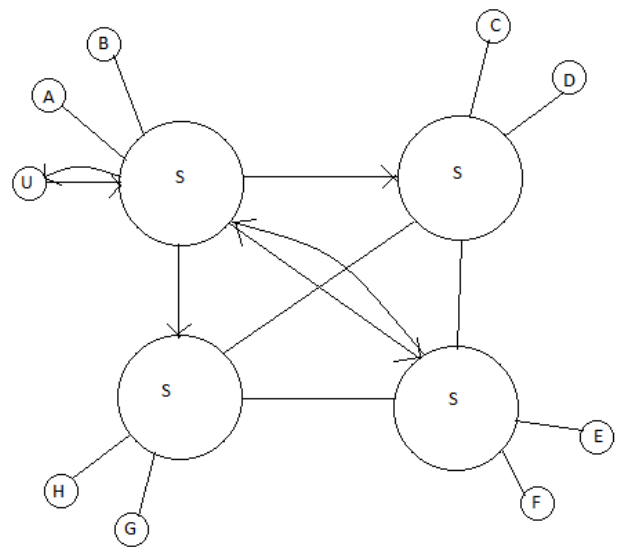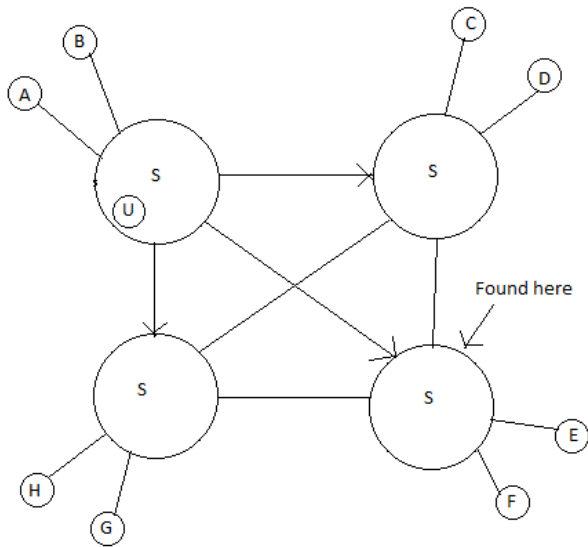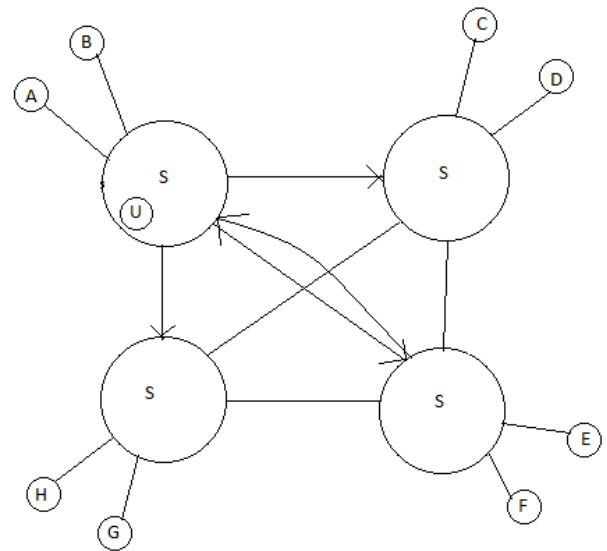
From the graphs, we can see that by becoming a part of supernode a user will have faster replies because the path which the data transfer take is shorter. This protocol also has the same problem with Gnutella in case of leechers. However, supernodes make it possible to maintain the path even if a computer is disconnected.

V. CONCLUSION

A directed graph can be used to draw the connection between computers in a peer-to-peer file sharing protocol. From the graphs, we can easily understand how the protocol works and spot the differences between each protocols as well as the srengths and weaknesses.

Author can then conclude that the reason why BitTorrent is so popular today is because nowadays people like to share and download large files such as games, movies, etc. With BitTorrent, people can easily find the files they want and download it with faster download speed. BitTorrent's complex decentralized network really supports downloading large files and there are many websites providing the torrent files for any files any user wants.

When found, the reply will be sent back following the path the request was sent.

REFERENCES

[1] http://www.zeropaid.com/bbs/archive/index.php/t-1961.html. Access date: 10 December 2011.
[2] http://www8.cs.umu.se/~bergner/thesis/html/node64.html. Access date: 10 December 2011.
[3] http://computer.howstuffworks.com/file-sharing.htm. Access date: 10 December 2011.
[4] http://computer.howstuffworks.com/napster.htm. Access date: 10 December 2011.
[5] http://computer.howstuffworks.com/bittorrent.htm. Access date: 10 December 2011.
[6] http://www.indiabroadband.net/general-offtopic-discussions/6783-how-exactly-torrent-works.html. Access date: 10 December 2011.
[7] http://en.wikipedia.org/wiki/Peer-to-peer_file_sharing. Access date: 10 December 2011.
[8] http://en.wikipedia.org/wiki/Napster. Access date: 10 December 2011.
[9] http://en.wikipedia.org/wiki/Gnutella. Access date: 10 December 2011.
[10] http://en.wikipedia.org/wiki/Fasttrack. Access date: 10 December 2011.
[11] http://en.wikipedia.org/wiki/BitTorrent_(protocol) . Access date: 10 December 2011.
[12] http://en.wikipedia.org/wiki/BitTorrent_(software) . Access date: 10 December 2011.
[13] Rosen, Kenneth H. 2007. *Discrete Mathematics and Its Application, 6th Edition*. New York: McGraw-Hill.
[14] Subramanian, Ramesh and Brian D. Goodman. 2005. *Peer to Peer Computing: The Evolution of a Disruptive Technology*. Hershey: Idea Group.

STATEMENT

I hereby declare that this paper is the work of my own writing, not an adaptation nor translation nor plagiarism.

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 12 December 2011

Danny Andrianto 13510011