December 2003

A Resource-Based Assessment of the Gnutella File-Sharing Network

Oleg Pavlov
Worcester Polytechnic Institute

Khalid Saeed
Worcester Polytechnic Institute

A RESOURCE-BASED ASSESSMENT OF THE GNUTELLA FILE-SHARING NETWORK

Oleg V. Pavlov and Khalid Saeed
Worcester Polytechnic Institute
Worcester, MA USA
opavlov@wpi.edu **saeed@wpi.edu**

Abstract

This paper reviews the growth behavior of a popular peer-to-peer network. We propose a dynamic hypothesis that the growth, overshoot, and collapse trajectories may be the result of complex causal interactions between inadequate resources, private provision of common goods, free riding, and membership dynamics. We draw parallels with other systems that are well-understood and known to exhibit similar trajectories. Computer experiments confirm that free riding by peers may lead to inadequacy of resources, decline in network performance, high attrition rates, and collapse. However, if freeloading tendencies are not strong, which is usually true in smaller groups, then the P2P system will function without oscillations. An experiment that considers improvements in search algorithms suggests that the reduction of total network traffic may not be sufficient to eliminate system fluctuations in the long run.

Keywords: Peer-to-peer technology, free riding, network resources, online file trading, system dynamics

Introduction

In 1999, 19-year old Shawn Fanning released Napster software. Its sole purpose was to allow music lovers to swap free MP3 files. Within 20 months, 65 million people shared more than 1 million titles (Leuf 2002, p. 191). In response to pirating on such a grand scale, a group of big record labels began a fierce legal battle to shut down Napster, Inc. They claimed that the network cost these companies millions of dollars in lost CD sales (CNN Money 2002). Meanwhile, Germany's communication conglomerate Bertelsmann AG invested $100 million in Napster.

The Napster-related publicity spurred a great deal of speculation about the approaching peer revolution in electronic commerce. One popular idea is that the peer-to-peer technology will democratize markets, allowing individuals to conduct their own auctions and eliminate the need for central auctions such as eBay or various business-to-business sites (Non 2000). Some early attempts in this direction include software from a company by the name Lightshare (**http://www.lightshare.com/**), which allows individuals or businesses to sell products from their computers without any intermediaries. Other attempts included adapting the P2P technology as a payment platform, although such experiments so far have had limited success (Elkin 2002). There are predictions that peer technology is the future of computing (Jovanovic et al. 2001) and of future distributed storage solutions (Leuf 2002, p. 106).

Napster eventually lost the case in court and had to shut down its network. However, a number of peer-to-peer networks that are more resilient to legal challenges still survive. Among them, a network based on the open-source protocol called Gnutella is the largest public peer-to-peer network in operation (Yang and Garcia-Molina 2002b). The technology allows formation of *ad hoc* computer networks without the centralized controlling body or central server. Advantages of such a system are distributed storage and processing cost, the autonomy of nodes, robustness to attacks, and load balancing (Yang and Garcia-Molina 2002a).

About 60 percent of Americans, or 174.6 million people, regularly use the Internet (Lenhart et al. 2003). After the adoption of the IPv6 protocol, millions of additional devices, such as appliances, will come on-line. The pressing issue is whether or not the

peer technology can scale to accommodate so many users. When Napster's future was legally questioned, many Napster users switched to Gnutella causing what was termed the Napster Flood in July/Agust 2000 (Ritter 2001). The sudden increase in membership did not last very long as the performance of the network very quickly deteriorated. Often music downloads failed and discouraged users left the Gnutella Network (**http://limewire.com/index.jsp/net_imrovements**). The analysis of the Napster Flood led some to conclude that the Gnutella-like network is not capable of handling the large volumes of traffic required for successful scaling (Ritter 2001). Proponents of Gnutella disagreed (Kabanov 2001), suggesting that when users switch to faster connections, such as cable or DSL, the problem will go away (Shirky 2000).

Overshoot and collapse behavior, similar to that observed in the case of Gnutella, is common in business and everyday life (Moxnes 1998; Senge 1990). We conjecture that excessive oscillations observed in Gnutella may be the result of forces similar to the ones present in socio-economic systems that display overshoot and collapse dynamics.

The following section describes the Gnutella Network in greater detail. Based on a review of systems that generate oscillating behavior, we propose a dynamic hypotheses. We describe our computer model and the computer experiments are presented. Finally, we offer our conclusions.

Gnutella Network

Gnutella's peer grid is a virtual network formed at the application level that is distinct from the underlying physical network (Ripeanu et al. 2002). A person can participate in the Gnutella network by either downloading a piece of software commonly referred to as a *servent* or by logging onto a dedicated Web site (Bolcer 2000). A Gnutella node forwards the search query to other nodes to which it is connected until the message travels the maximum allowed number of hops determined by the time to live (TTL) parameter. Hosts that contain the material in the query respond with a message that is traversed along the path on which it arrived. The original Gnutella protocol treats nodes equally, irrespective of their network connection speed, memory, or clock speed (Bolcer 2000).

Gnutella has been compared to an Internet potluck (Kan 2001): nodes contribute to the network by providing files and by routing network traffic (Adar and Huberman 2000). Providing content to other peers is costly not only because acquiring the content imposes some fixed cost on the altruistic peer, but also because each additional upload from your computer slows down the serving computer and its own downloads (Adar and Huberman 2000; Yang and Garcia-Molina 2002b). Users clearly have an incentive to free ride with respect to content and bandwidth, which means "taking their share of it and keeping their own resources for themselves" (Marwell and Ames 1979).

Free riding may be accomplished in a variety of ways. By default, most peer software share all downloaded files (Golle et al. 2001; see also **http://www.limewire.com**). However, only about 30 percent of users share files on Gnutella and 20 percent of hosts share 98 percent of all the files available on the network (Adar and Huberman 2000; see Figure 1). As a result, 1 percent of the hosts provide 47 percent of the answers to file requests and 25 percent provided 98 percent of the answers (Adar and Huberman 2000). This is shown in Figure 2. Free riding can also be controlled by the desirability of the shared content. If the content is not desired by others, then the node free rides (Adar and Huberman 2000). Capacity offered to the network can be controlled by the number of allowed connections and by misstating the connection speed. An extreme case of free riding are browser-based Gnutella search web sites, e.g., asiayeah.com and gnute.com. Users that enter the Gnutella Network through such sites search the shared database but do not contribute content and do not route traffic.

A person may also choose not to contribute to the network simply by turning the computer off. There is a special term used to describe this type of behavior – *fishing* – the user logs into the network, downloads what he needs, and promptly leaves the system. Data presented in Figure 3 show that about half of the connections were 60 minutes or shorter and only 10 percent remain in the network for longer than 3 hours.

Systems That Lead to Growth, Oscillations, and Collapse

The Napster Flood of July 26-28, 2000, manifested itself in the 100 percent growth in the query rate on Gnutella (Bolcer 2000). After the surge, the rate fell and then rebounded to the Napster Flood level by August 14, 2000. During the Flood and after August 14, 2000, users complained that queries produced fewer responses and it took longer to get results (Bolcer 2000). Improve-

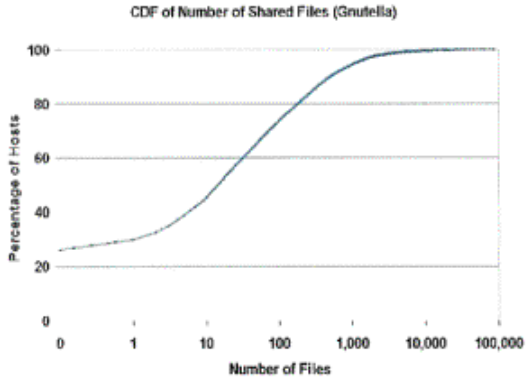

Figure 1. File Sharing in Gnutella
(Stefan Saroiu, P. Krishna Gummadi, and Steven D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," SPIE Proc. 4673, Multimedia Computing and Networking 2002 (2002). Reproduced with permission.)

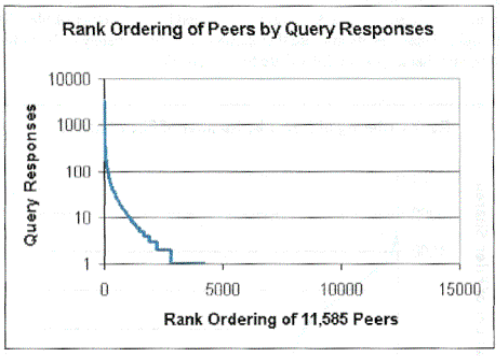

Figure 2. Query Responses
(E. Adar and B. A. Huberman, "Free Riding on Gnutella," *First Monday* (5:10), 2000. Reproduced with permission.)

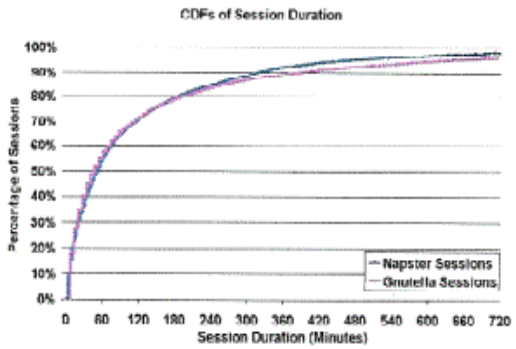

Figure 3. The CDF of Napster and Gnutella Session Duration
(Stefan Saroiu, P. Krishna Gummadi, and Steven D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," SPIE Proc. 4673, Multimedia Computing and Networking 2002 (2002). Reproduced with permission.)

ments in technology, such as more efficient search algorithms and the introduction of Pong servers, reduced overhead traffic, which allowed new growth. The network grew exponentially from about 2,000 users in November 2000 to about 48,000 by May 2001 (Ripeanu et al. 2002). However, as network statistics show (Figures 4 and 5), the pattern of growth and collapse is still present.

Overshoot and collapse behavior, similar to the one observed in the case of Gnutella, is common in business and everyday life (Moxnes 1998; Senge 1990). One of the extreme cases of overshoot and collapse that has been extensively studied and is well-understood is the tragic story of Easter Island. This is a small island in the Pacific a few thousand miles off the coast of Chile. After the arrival of the first residents on the island c. 400AD, the population grew exponentially. But as the local resources, such as trees and soil, were depleted, the population fell precipitously (Figure 6). Typically such behavior is generated by a system as shown in Figure 7.

It has been suggested that scarce resources in distributed computer systems may generate oscillations (Huberman 1998). Based on the knowledge of other systems that exhibit strong fluctuations, we propose a dynamic hypothesis that a peer-to-peer system can be characterized by a structure similar to the one observed in Figure 7. The carrying capacity is the bandwidth that gets depleted as more users join the network without adequately contributing to the common pool, that is, free riding. The causal

structure is shown in Figure 8. It also includes the main draw for user participation, the content, which also suffers from freeloading tendencies. Free riding leads to inadequate private provision of public resources, which eventually contributes to overshoot and collapse.

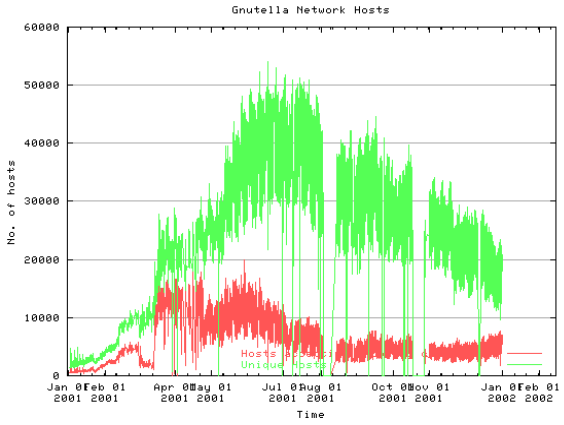

Figure 4. Gnutella Network Size for 1/1/2001–2/1/2002
(This graph is reproduced courtesy of Lime Wire, LLC, from **http://www.limewire.com/content/netset.shtml**)

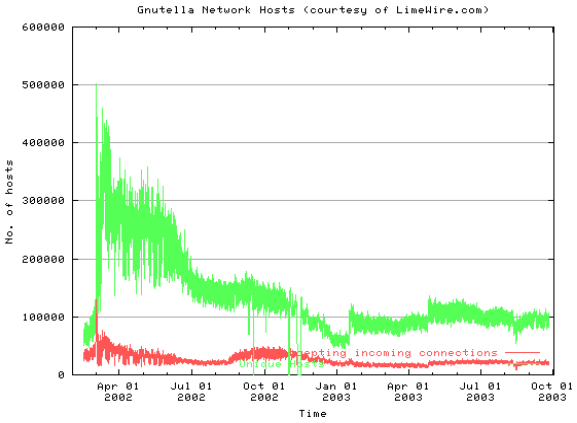

Figure 5. Gnutella Network Size for 2/1/2002–4/20/2003
(This graph is reproduced courtesy of Lime Wire, LLC, from **http://www.limewire.com/content/netset.shtml**)

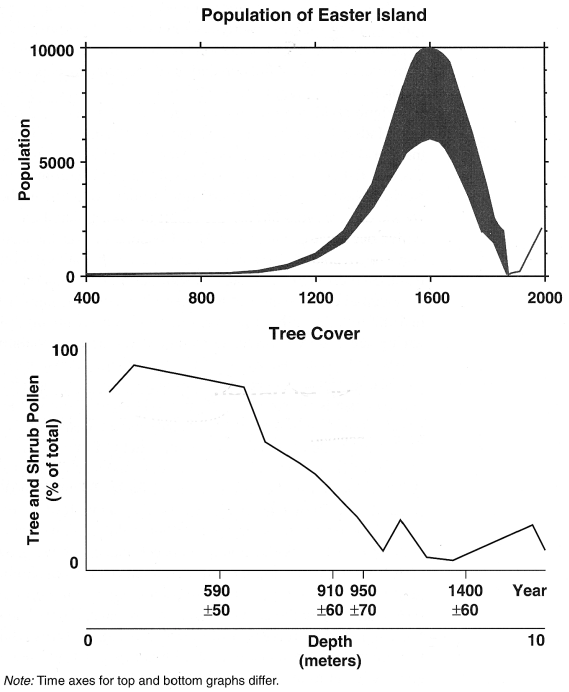

Figure 6. Overshoot and Collapse on Easter Island
(Source: J. D. Sterman, *Business Dynamics*, McGraw-Hill, Boston, 2000.

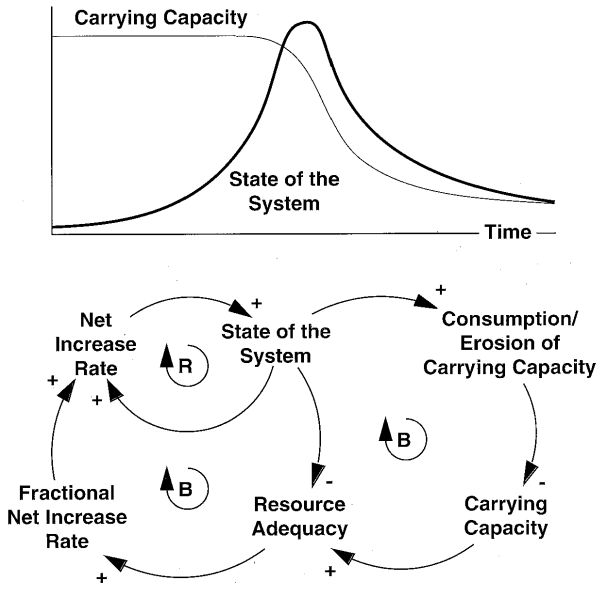

Figure 7. A Typical Pattern of Overshoot and Collapse and a Corresponding Causal Structure
(Source: J. D. Sterman, *Business Dynamics*, McGraw-Hill, Boston, 2000.

Group size has been recognized as an important factor for on-line communities (Butler 2001). In experiments, free riding typically exacerbates with crowding (see, for example, Isaac and Walker 1988). Additionally, as the group size expands, a sense of community disappears. Typically, a community stresses the importance of shared understanding, a sense of obligation rather than self-interest (Bender and Kruger 1982). Therefore, *contribution fraction* is negatively linked to the network size. Smaller contribution fraction leads to greater *free riding*. Figure 8 also identifies positive and negative feedback loops in the system.

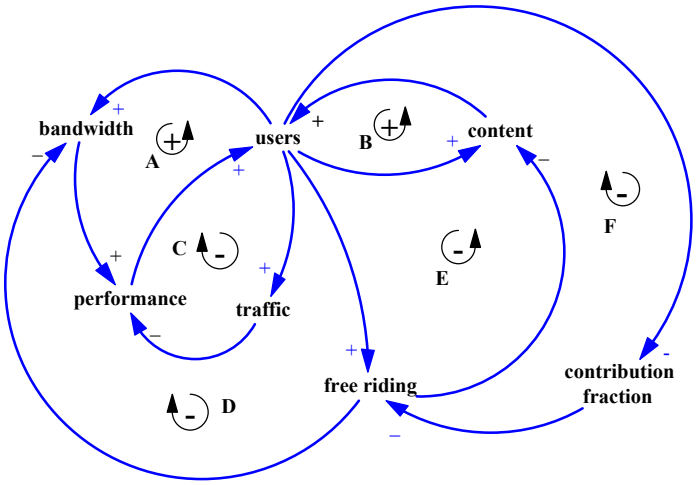

Figure 8. System Structure of the Gnutella Network

Model

We built a computer model corresponding to the causal structure described in Figure 8. The model consists of the following sectors: membership, shared content, capacity, and traffic.

Membership Sector

This sector, as shown in Figure 9, captures the average number of peers logged onto the system. The *attractiveness* of the network is enhanced by such attributes as *content attractiveness* (which in turn depends on the available content), shorter *latency* in responses to requests and downloads, as well as better chances of completed downloads. Users are also more likely to remain connected to the system longer when the system is attractive.

The *clustering coefficient* measures how connected a node's neighbors are. The lower the coefficient, the fewer connections neighbors have between them. The lower the coefficient, the lower the probability that you will see the same group of nodes next time you log onto the Gnutella. We use clustering coefficient as a proxy for the closeness of the peer community. The *contribution fraction* measures the average share of the individual content and bandwidth that is being made available by peers to the rest of the network.

Shared Content Sector

Figure 10 presents a rendition of the shared content sector. This sector keeps track of the files available through the network. The maximum content a new peer can bring to the group is the number of files on its hard drive. This is when *contribution fraction* is equal to one.

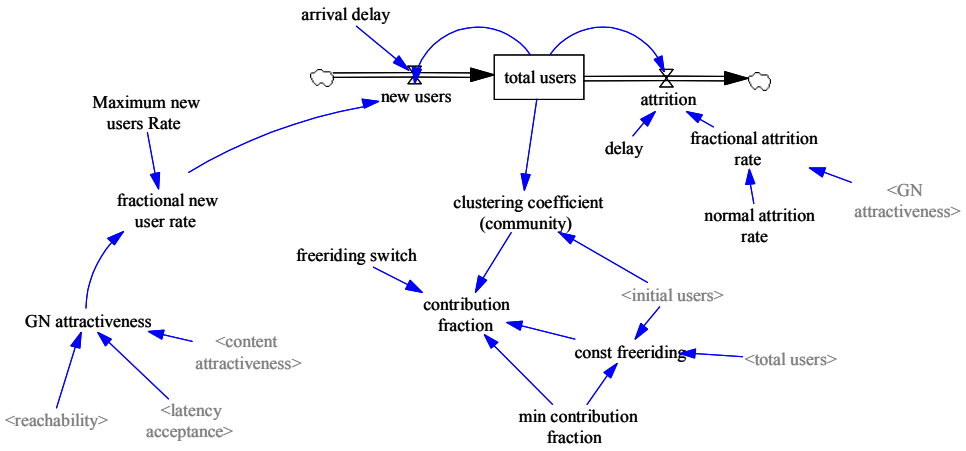

Figure 9. Membership Sector

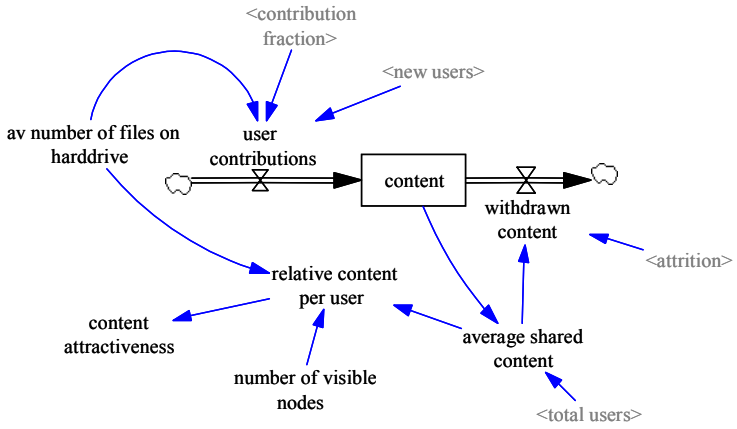

Figure 10. Shared Content Sector

A feature of the Gnutella topology is that due to the limited connectivity and the finite time to live (TTL) parameter, the potential reach of each node is much smaller than the entire network (Leuf 2002, p. 199). For the data collected for a seven month period starting in November 2000, Ripeanu et al. (2002) found that average node connectivity is 3.4 and is independent of the network size. Ritter (2001) estimates that for a network in which each node has on average three edges and TTL is set to 7 (a typical number in Gnutella), at best 21 nodes are visible from each peer. This is why it is the *relative content per user*, rather than the absolute value of *content*, that determines the *content attractiveness*.

Capacity Sector

Similarly to the shared content, shared bandwidth increases with each additional new peer, and diminishes when peers leave the network. At best, each peer contributes all its available bandwidth, which is the *average node bandwidth*. The sector is shown in Figure 11.

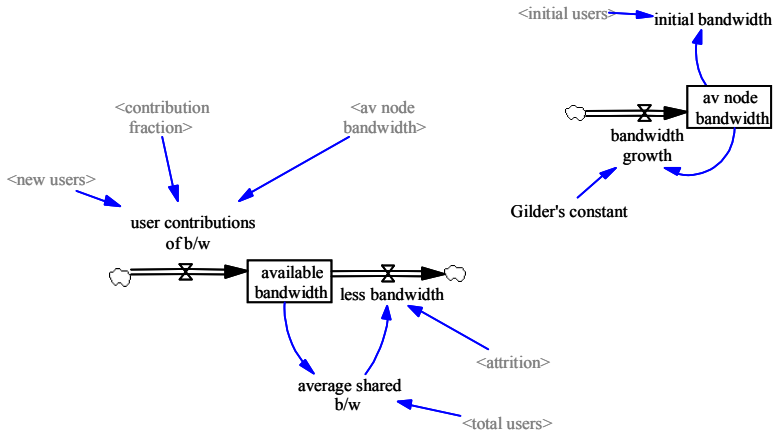

Figure 11. Capacity Sector

Traffic Sector

This sector models traffic in the network (see Figure 12). Following, Ritter (2001) we assume a liner relationship between the number of queries and the size of the network, that is, each peer submits some average number of requests to the system. Additionally, we assume, following Yang and Garcia-Molina (2002b), that some average aggregate bandwidth (in bytes) is generated by a representative query. All search requests add up to the total network *traffic*.

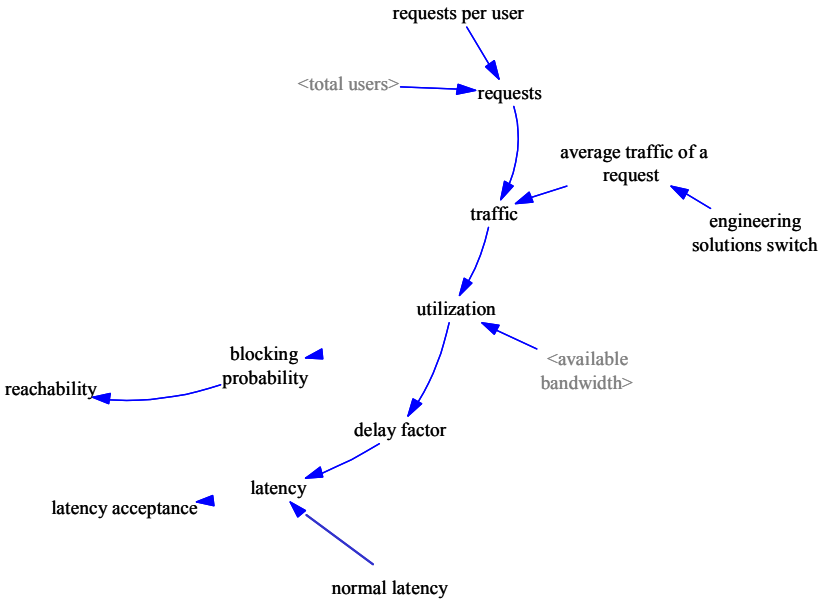

Figure 12. Traffic Sector

Utilization is the ratio of traffic to the connection capacity, which is measured in terms of bandwidth. Once a node's bandwidth is saturated, a number of things might happen (Leuf 2002, p. 121). First, a connection might be dropped. This would lead to lost return paths, unfulfilled requests, and repeat of request broadcasts. Second, the node may simply ignore some of the request traffic. Third, the node can buffer some messages and wait until bandwidth frees up, but this would slow down the performance of the computer and also contribute to the latency along the path. In general, network theory suggests that delay (*latency*) and network traffic are related as in Figure13. Lower levels of latency improve current *latency acceptance* variable. The importance of latency can be understood if one remembers that fast response times contributed to Napster's explosive popularity (Leuf 2002, p. 130).

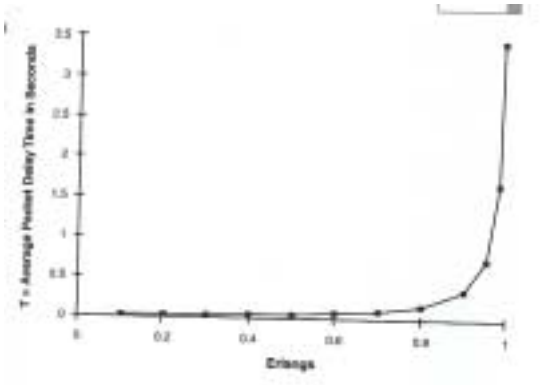

Figure 13. Average Packet Delay
(Source: J. A. Pecar, and D. A. Garbin, *The New McGraw-Hill Telecom Factbook*, McGraw-Hill, New York, 2000, p. 429.)

Base Run

Figure 14 shows Gnutella data from Figure 4 overlaid by the simulated trajectory generated by our model. One immediately notices the absence of high frequency fluctuations in the simulation. The fast oscillations in the data come from hourly variations in online usage: more people are on the Internet around midnight than at 6 o'clock in the morning (Kotz and Essien 2002). In this model, we do not replicate hourly variations in order to avoid the potential problem of stiffness that arises when time constants of significantly different magnitudes are employed in a model (for a discussion, see Maron and Lopez 1991). Figure 14 demonstrates that the model is capable of replicating quite accurately the historic trend for the sample year.

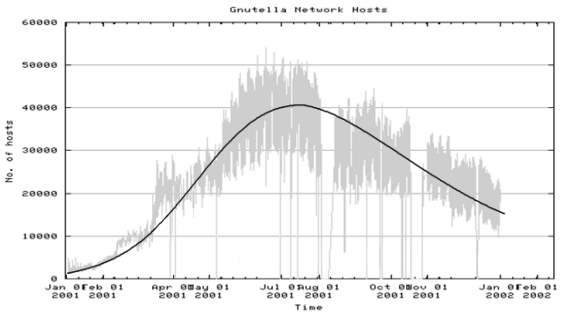

Figure 14. Simulated Trajectory

Computer Experiments

In the following experiments, we set the initial size of the Gnutella Network to 2000 nodes, which is about the size reported by Adar and Huberman (2000) for a day in August 2000.

Experiment 1: Altruistic Agents

Altruistic members contribute the entire content of their hard drives and all of their available bandwidth to the peer network. In the model, this is expressed by setting the *contribution fraction* in the membership sector to 1 (*freeriding switch* = 1). Note that this means that the marginal request contribution from each additional member is the same and equal to *average requests per user* (see the traffic sector). The generosity of members allows the system to expand exponentially (Figure 15). Realistically, of course, such an altruistic network would be limited by the maximum number of potential users, such as, for example, the entire on-line population of the United States. When limited, the growth trajectory would be an S-shaped curve.

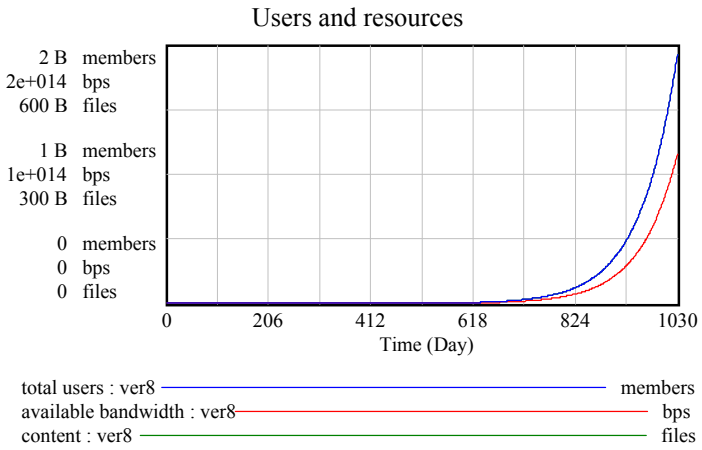

Figure 15. Exponential Growth When Members are Altruistic

Experiment 2: Free Riding Tied to Network Size

With this experiment, we would like to address the question of whether the dynamics of the system change if free riding behavior becomes more prominent with the group size (Figure 16). An additional balancing loop F (see Figure 8) passing through the *contribution fraction* adds to the complexity of the system. Interestingly, the system experiences a series of dampening oscillations, eventually, converging to a group of a smaller size, in which altruistic behavior dominates.

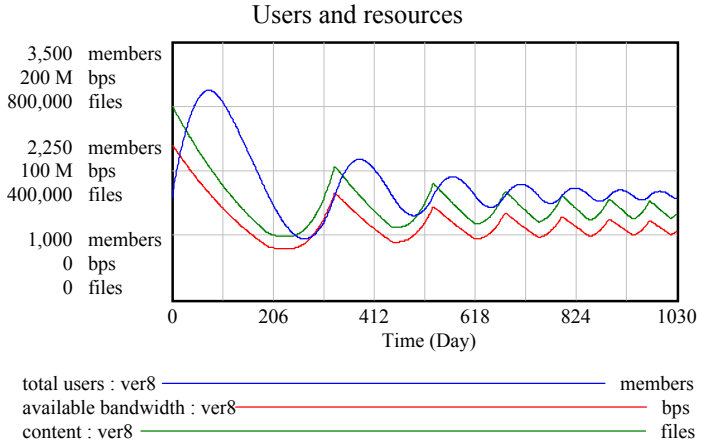

Figure 16. Oscillations When Free Riding Fraction Changes

This result suggests that the Gnutella protocol is capable of supporting smaller groups of users that do not exhibit strong freeloading tendencies. Network size may be regulated in various ways. LimeWire, one of the popular Gnutella *servents*, allows forming virtual clusters around a common interest. An economic solution may logically divide the entire network into smaller networks. For example, Odlyzko (1997) suggested, although for a different problem, a pricing system he called Paris Metro Pricing for the Internet. The idea is to let users self-select into subnetworks of various sizes based on their willingness to pay. More expensive networks will have fewer people and better performance.

Experiment 3: Engineering Improvements

Gnutella uses a breadth-first traversal (BFS) search algorithm. A query message propagates to the depth, which is equal to the TTL value. BFS wastes bandwidth because only very few nodes carry useful information, and therefore only very few nodes should be contacted (Yang and Garcia-Molina 2002b). Iterative deepening has been proposed as one of the possible improvements to the search engine. Better algorithms would also take into account the mismatch between the virtual topology of Gnutella and the physical connections between computers (Ripeanu et al. 2002). One of the major goals of such improvements is to reduce network traffic.

In this experiment, we assume that a new technological improvement is introduced into the network at time t = 550. The effect of the innovation is the reduction in the average traffic generated by each search request. As Figure 17 shows, reduction in traffic boosts performance and encourages new users (loop C in Figure 8). However, the engineering solution does not change the underlying structure of the system, and therefore does not change the dynamics of the system. The system still exhibits wide oscillations.

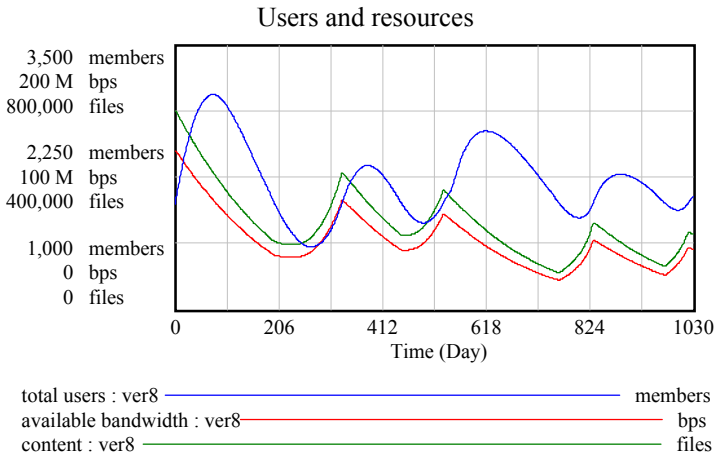

Figure 17. Engineering Improvements

Conclusion

This paper reviews the growth behavior of a popular peer-to-peer network. We propose a dynamic hypothesis that the overshoot and collapse trajectories may be the result of complex causal interactions between inadequate resources, private provision of common goods, free riding, and membership dynamics. We draw parallels with other systems that are well-understood and known to exhibit similar trajectories. Computer experiments confirm that free riding by peers may lead to inadequacy of resources, decline in network performance, high attrition rates, and collapse. However, if freeloading tendencies are not strong, which is usually true in smaller groups, then the system will function without oscillations. An experiment that considers improvements in search algorithms suggests that the reduction of total network traffic may not be sufficient to eliminate system fluctuations in the long run.

References

Adar, E., and Huberman, B. A. "Free Riding on Gnutella," *First Monday* (5:10), 2000.

Bender, T., and Kruger, S. *Community and Social Change in America*, Johns Hopkins University Press, Baltimore, MD, 1982.

Bolcer, G. A. "Bandwidth Barriers to Gnutella Network Scalability," *FoRKed*, September 22, 2000 (available online at **http://xent.com/FoRK-archive/sept00/0657.html**).

Butler, B. S. "Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures," *Information Systems Research* (12:4), 2001, pp. 346-362.

CNN Money. "Napster Sold to Roxio for $5.3 Million," *CNN Money*, November 27 2002 (available online at **http://money.cnn.com/2002/11/27/news/deals/napster/index.htm**).

Elkin, N. "The Future of the P2P Transaction Market," *Entrepreneur.com*, January 2, 2002 (available online at **http://www.entrepreneur.com/Your_Business/YB_SegArticle/0,4621,296007,00.html**).

Golle, P., Leyton-Brown, K., Mironov, I., and Lillibridge, M. "Incentives for Sharing in Peer-to-Peer Networks," in *Proceedings of the Third ACM Conference on Electronic Commerce*, ACM Press, New York, 2001 (available online at **http://robotics.stanford.edu/~kevinlb/ec01-short.pdf**).

Huberman, B. A. "Computation as Economics," *Journal of Economic Dynamics and Control* (22:8/9), 1998, pp. 1169-1186.

Isaac, R. M., and Walker, J. M. "Group Size Eeffects in Public Goods Provision: The Voluntary Contributions Mechanism," *The Quarterly Journal of Economics* (103), 1988, pp. 180-199.

Jovanovic, M. A., Annexstein, F. S., and Berman, K. A. "Scalability Issues in Large Peer-to-Peer Networks—A Case Study of Gnutella," University of Cincinnati Technical Report 2001 (available online at **http://www.ececs.uc.edu/~mjovanov/Research/paper.html**).

Kabanov, M. "In Defence of Gnutella," Editor's Column, 2001 (available online at **http://gnutellameter.com/gnutella-editor.html**).

Kan, G. "Gnutella," in *Peer-to-Peer: Harnessing the Benefits of a Disruptive Technology,* A. Oram (ed.), O'Reilly, Cambridge, MA, 2001.

Kotz, D., and Essien, K. "Analysis of a Campus-Wide Wireless Network," in *Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking*, ACM Press, New York, 2002, pp. 107-118 (available online at **http://doi.acm.org/10.1145/570645.570659**).

Lenhart, A., Horrigan, J., Rainie, L., Allen, K., Boyce, A., Madden, M., and O'Grady, E. *The Ever-Shifting Internet Population: A New Look at Internet Access and the Digital Divide*, The Pew Internet and American Life Project, 2003 (available online at **http://www.pewinternet.org/reports/pdfs/PIP_Shifting_Net_Pop_Report.pdf**).

Leuf, B. *Peer to Peer: Collaboration and Sharing Over the Internet*, Addison-Wesley, Boston, MA, 2002.

Maron, M. J., and Lopez, R. J. *Numerical Analysis: A Practical Approach*, Wadsworth Publishing Company, Belmont, CA, 1991.

Marwell, G., and Ames, R. E. "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem," *American Journal of Sociology* (84:6), May 1979, pp. 1335-1360.

Moxnes, E. "Not Only the Tragedy of the Commons: Misperceptions of Bioeconomics," *Management Science* (44:9), 1998, pp 1234-1248.

Non, S. G. "Does the Peer-to-Peer Model Make Business Sense?," ZDNet UK, July 14, 2000.

Odlyzko, A. M. "A Modest Proposal for Preventing Internet Congestion," mimeo, AT&T Labs, 1997 (available at **http://www.dtc.umn.edu/~odlyzko/doc/networks.html**).

Pecar, J. A., and Garbin, D. A. *The New McGraw-Hill Telecom Factbook*, McGraw-Hill, New York, 2000.

Ripeanu, M., Foster, I., and Iamnitchi, A. "Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design," *IEEE Internet Computing* (6:1), 2002, pp. 50-57.

Ritter, J. "Why Gnutella Can't Scale. No, Really," February 2001 (available online at **http://www.darkridge.com/~jpr5/doc/gnutella.html**).

Saroiu, S., Gummadi, P. K., and Gribble, S. D. "A Measurement Study of Peer-to-Peer File Sharing Systems," in *SPIE Proceedings of the Multimedia Computing and Networking Conference*, San Jose, CA, January 2002.

Senge, P. M. *The Fifth Discipline: The Art and Practice of the Learning Organization*, Doubleday/Currency, New York, 1990.

Shirky, C. "In Praise of Freeloaders," O'Reilly Network, 2000 (available online at **http://openp2p.com/pub/a/p2p/2000/12/01/shirky_freeloading.html**).

Sterman, J. D. *Business Dynamics*, McGraw-Hill, Boston, 2000.

Yang, B., and Garcia-Molina, H. "Designing a Super-Peer Network," mimeo, Computer Science Department, Stanford University, 2002a.

Yang, B., and Garcia-Molina, H. "Efficient Search in Peer-to-Peer Networks," mimeo, Computer Science Department, Stanford University, 2002b.