

Peer-to-Peer File-Sharing

Expanding and Specifying the Model

SS21XX Final Project

Mike Kissinger
Jeremy Tout

Introduction

Since their inception, peer-to-peer networks have generated much interest in both the technical and legal fields. From a technological standpoint, they represent a robust strategy for distributing resources across a network (for example, the Internet) that eliminates central points of failure, and in many cases facilitates network performance. Very often, however, these shared resources are copyrighted media, mainly music and motion pictures. Because of the robust and decentralized nature of the networks, which often also protect user anonymity, enforcing copyright laws can be near to impossible. Additionally, because there is no central server in most peer-to-peer (P2P) networks, it is often difficult to obtain accurate network statistics, making the study of the behavior and performance of peer-to-peer networks a difficult task. In order to gain a better understanding of the behavior of such a system, models, in this case designed using system dynamics, can be employed.

The model we chose to expand is the model published in the *System Dynamics Review*, which uses stocks of users, content, bandwidth, and network performance to simulate the behavior of a simple pure peer-to-peer network. The model's assumptions about the network are simple because they assume bandwidth contributions proportional to users, and network throughput based on traffic per bandwidth availability. The term "pure peer-to-peer" is used to describe the model because it attributes all users the same status, meaning the model does not account for central servers or "super peers" that are sometimes used to facilitate searching and indexing of content within a network.

Because super peers have different responsibilities than normal users, differentiation of each of their functions may serve to specify the model and demonstrate differences between pure and “mixed” networks, where some peers are granted special status.

Early file sharing networks, including Napster, used a “hybrid” topology that used central servers for keeping track of the network’s content. This facilitates fast searches and content indexing, but left Napster vulnerable to legal action because the owners of the servers could be implicated in facilitating copyright infringement. The service has since been revived, although it is primarily in name only; Napster users are now charged for downloads and pay royalties to copyright owners. The unfeasibility of the central server model (at least for networks that operate outside the law) makes a completely decentralized network desirable, because it allows a greater degree of freedom and anonymity.

A “pure” network (an example of which is Freenet, an open source peer-to-peer client) has an egalitarian structure in which each user contributes content and bandwidth. Each node shares responsibility equally for discovering and transferring shared content as well as relaying search results and requests. In this case, the peer-to-peer model from the *System Dynamics Review* is appropriate, because it assumes all users contribute to bandwidth and content more or less equally. This is, of course, not always the case if users can withhold resources and have incentive to do so. However, this uniform topography with evenly distributed responsibility makes modeling the network a relatively simple proposition.

Gnutella is a file sharing protocol that combines central (but dynamically distributed) search and indexing responsibilities with a distributed peer-to-peer network. Because these super peer nodes bear much of the responsibility for maintaining records of the network's content, their function is critical for regular users to locate resources within the network. Even a content-rich network suffering from a lack or malfunction of super peers can be unattractive because media, though present, is inaccessible.

However, the inclusion of super peers serves to simplify the demands placed on normal user nodes, and centralizes the overhead of searching and indexing. This method should increase not only network performance, but also the array of network content available to any particular user versus normal peer-to-peer systems. By making minor adjustments within the model, we hope to adjust its dynamics to mimic those of a super peer based network.

Additionally, improving technology is likely to be a major contributor to the success of peer-to-peer networks and their improvement in the future. Because new technology is unpredictable in terms of its possible implementations, we are using a simplified approach that seeks to model the effects of the technology rather than the technology itself. Improvements in technology are likely to have effects on improving network traffic and performance, improving content availability and improving potential download success.

One such improvement is the imposition of an economy on the system. This economy tracks user contribution to and use of the network, allowing what had previously been folk views about the necessity of contribution to become institutional

rules that mandate user participation. Imposition of such an economy is likely to have a market effect on the free riding problem by preventing non-contributors from degrading network quality. This overall serves to reduce the effect of the free rider problem, or in the case of a well-implemented solution to eliminate it entirely.

Model Modifications

Because the roles and interactions of the super peers within the system may vary from network to network and because the simple inclusion of the super peers in the model changes the dynamics in unexpected ways, it may be more fruitful to model their influences through different methods. Our simple method assumes that super users are part of a technological adjustment to the network that improves user access to content by improving search efficiency and power. The various implementations of such technology may have differing impacts, so for simulation purposes, this effect may remain a variable that is left up to user definition.

Economic Imposition

The first modification to the model is the addition of a variable named “economic free-riding scalar.” This variable serves as a multiplier that changes the value of the “adjusted free riding fraction”. With this new variable included, the equation for free riding becomes:

adjusted free riding fraction = economic free-riding scalar * free riding effect * minimal free riding fraction

“economic free-riding scalar” is intended to have values from 0 to 1; 0 indicating a total elimination of all free riding and 1 indicating a null effect of the economy on the free riding problem. Theoretically, if we wanted to suppose that economic tactics backfired, it could have a value such that effect of economy on free riding > 1 , but this seems unlikely to happen in any well-conceived system. The peer-to-peer network Mojo uses an economy to track users’ “mojo,” a value that represents their network contributions.

Although measures like this could in theory eliminate free riding altogether by enforcing a 1:1 ratio between contribution and credit, such stringent measures are likely to stifle a network and drive away users. Another problem with this solution is that a pure peer-to-peer network requires local storage of all network information, meaning that hacked software could easily project an image of a generous contributor without any actual contribution made to the network.

Measures of participation include the size and popularity of a users’ library, the throughput in terms of network traffic by a node, successful uploads from a node, as well as searching and indexing related activities. Choosing different measures may have differing effects. If users are concerned about hosting a large volume of content because they believe it makes them more susceptible to copyright holder lawsuits, then an economy based on network bandwidth credits may not prevent a decline in the overall

network library. Likewise, library size criteria that offer no incentive to complete uploads or contribute bandwidth may leave the network lacking in these areas. Because of this, total uploads (which gives a concrete idea of total useful network throughput) is probably the best choice as a main economic criterion.

This, of course assumes that users care about and are capable of understanding the workings of the network and the economy. Given that many users of popular peer-to-peer networks do not understand the workings of their own computers, it is unreasonable to assume knowledge of technical peer-to-peer issues in the user base. This should not however be confused with software that bypasses network features, which once developed are generally not difficult to learn and use. A finished client that is hacked to avoid restrictions placed on it by an economy nullifies the effect of the economy by essentially granting unlimited credit to users of hacked clients. These clients may be identical to standard clients, aside, of course from this hacked feature, and therefore there is no impediment to their adoption by the average user provided that the new client is widely known and available.

We did not include any separate variables to account for the bypassing of economic network features in our model. It would however be fairly simple to model this by assigning a decreased value to the “effect of economy on free riding,” or even by setting its value to a decreasing function to approximate the emergence and popularization of economy bypass measures.

Technological Improvement

The rapid technological progress made by computer technology in the past decade has been and continues to be stunning to many. The progression of Internet traffic dominance began in the 80s with ftp (file transfer protocol), and then shifted to http traffic as the World Wide Web became popular in the 1990s. Only recently has the dominance of Internet traffic shifted from centralized services to distributed peer-to-peer traffic. This traffic is however in terms of total bandwidth, which is not surprising given that peer-to-peer objects are often very large--a normal feature-length movie file is typically around 700Mb, depending on content and encoding. This size reflects that maximum capacity of most write-capable compact disc media. MP3 files, another dominant form of peer-to-peer traffic are significantly smaller than movie files, but nonetheless typically range from 3Mb to 10 Mb depending on file length and quality. By comparison, web objects are usually tailored for a variety of users, including low-end users with slow connections, and contain much simpler text-based and compressed image content. Web objects in general tend to be in the range of 1kb to 50kb, several orders of magnitude smaller than peer-to-peer objects.

The preference in the tradeoff between size and quality may depend on a user's resources and tastes. A user with a low end computer and a slow network connection may cherish small files for their reduced demand on his disk space, while an audiophile is willing to sacrifice volume of her collection if the quality of the music is improved. Additionally, a large file may be more difficult to download if a host user leaves the

network before the download is complete. A slow connection to the network may limit a user's ability to acquire large files, especially because dial-up users often piggyback their Internet connection over their telephone line. Digital Subscriber Line (DSL) services, which use high-frequency bandwidth that until recently was filtered from voice calls, have largely replaced dialup because declining prices have made the improved connectivity of DSL available to most dialup users.

The increased prevalence of widely available broadband connections such as DSL and digital cable services shows that improved technology can have great effects on general bandwidth availability and therefore on peer-to-peer networks. We sought to model the effect of technological improvement by adding a variable called "improved technology scalar." This improved infrastructure is not meant to reflect the overall increase in user-contributed bandwidth as a result of democratic adoption of broadband technology, but rather the technical improvements made in network features and structure. This variable, which affects "content reachable by a user" is intended to account for better searching and indexing, improved sharing algorithms, and any other changes that increase the ability of users to share and retrieve content. This variable also affects the normal latency, typical node bandwidth, and average new user contribution. These affects represent network improvements and improvements made to media processing and encoding such that it becomes much easier for the end user to share his media.

With the "improved technology scalar" factored in, the variable "content reachable by a user" increases, because its equation becomes:

content reachable by a user = improved technology modifier * number of reachable nodes
* average shared content

It is assumed that the improvements in algorithms that run the network do not have a detrimental effect on its performance, but if such a situation were desired for modeling, “improved technology scalar” could be set less than 1. Additionally, if one wanted to assume that routing, indexing, searching and/or sharing capabilities increase with time just as bandwidth contribution does, one could create a simple growth structure to model this. Out of desire to keep our modifications simple and easily changeable by an end-user, along with the limited scope and time constraints of the project, we left it as a single variable that rather than modeling actual dynamics, merely represents how they might act.

Through manipulation of these variables, we hope to provide a tool for understanding the effects of technical progress in distributed peer-to-peer technology. Although there are currently no dynamics underlying the “improved technology scalar” and “economic free-riding scalar,” their simple implementation will hopefully allow for further experimentation in their effects. Assumptions about the effects of complex or fundamental changes to the network may be difficult to make accurately, and algorithm implementation is often subject to many hindrances and constraints, and is usually sudden and drastic, a different growth pattern from the steady increase in network users’ bandwidth contribution.

Because of the uncertainties and complexities associated with these types of improvement, it may be difficult to determine how the underlying dynamics should be modeled. Things like breakthrough research and revolutionary technology would clearly represent exogenous shocks to the model, and for this reason, step functions are amenable to this sort of modeling. Step functions are variables that begin with a value and then at some predetermined time, change that value to another. There is no smooth or continuous transition, which is intuitive, because the implementation of a new technology may require that its peers have the same technology and be compatible.

In our experiments we will use these variables to determine the potential effects of revolutionary improvement on peer-to-peer network behavior. By using step functions that allow the effects to be not only of varying magnitude, but also at varying times, we may determine new potential dynamics of peer-to-peer networks. The model is available online at <http://broadcast.forio.com/sims/ptopmods/> for public experimentation. Below is a summary of preliminary experiments that show the results that some of our changes had on model dynamics.

Experiments

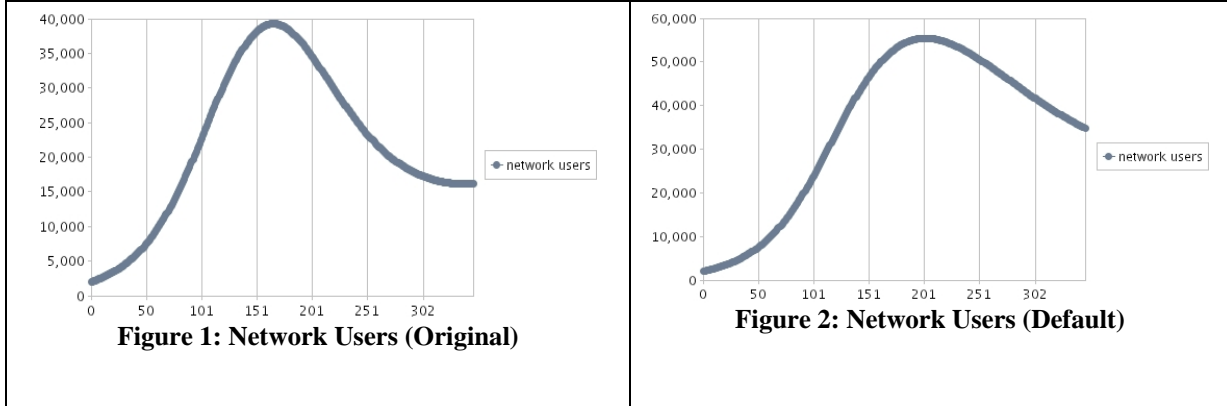
For the purpose of these experiments, we present side-by-side graphs with the experiment under evaluation and, as a control, a simulation with both of our new variables set to “1,” i.e. removing their effect to obtain results similar to the original Pavlov-Saeed model. In this manner we hope to illustrate the actual change involved

with our models, and not some sort of overarching behavior present regardless to the variable.

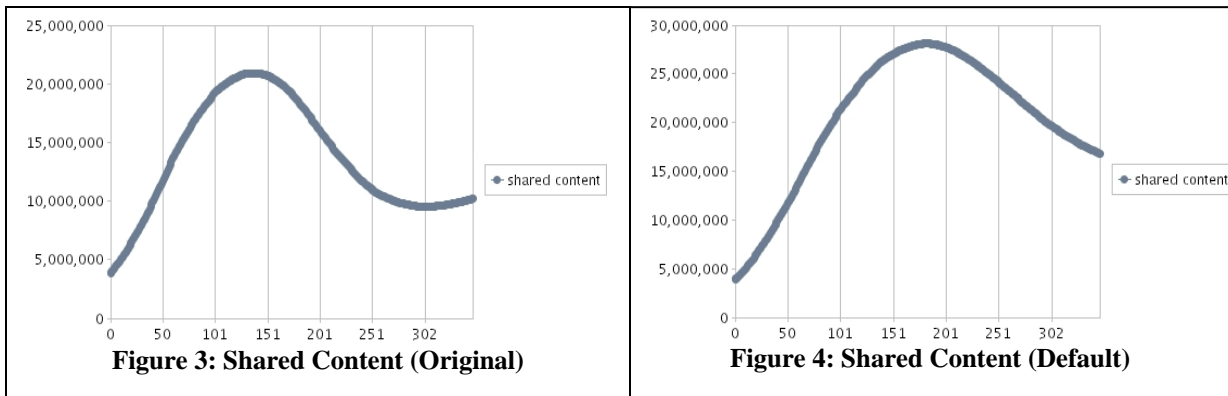
In all of the experiments, we have omitted graphs and discussion of the traffic measurement, due to the fact that in all cases it closely mimicked the behaviors observed in both the network users and shared content stocks.

Default Settings

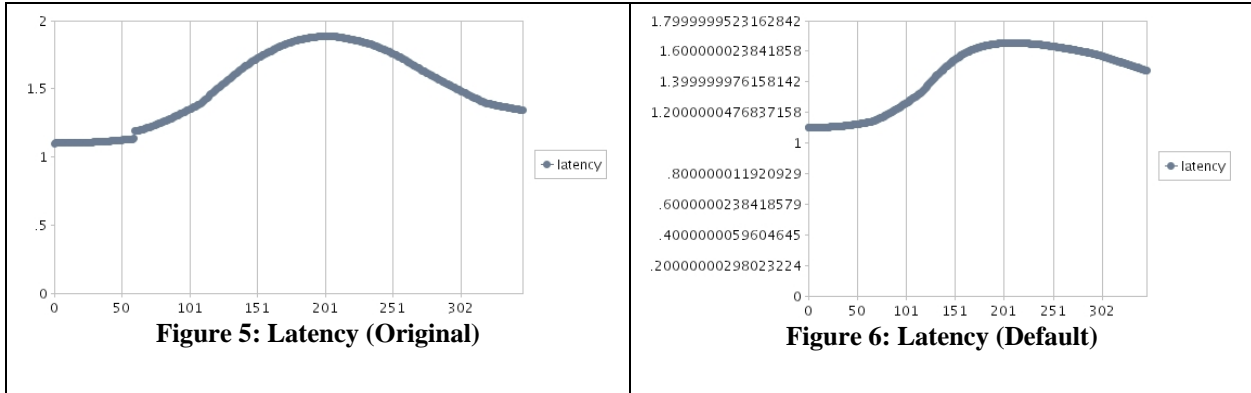
The default settings that we chose for the simulation are to set “economic scalar on free riding” to 0.98 and to set “improved technology scalar” to 1.05. All other variables were left unchanged, set to their values in the original model. This represents a 2% reduction in free-riding behaviors due to a reduction of economic pressures involved with the purchase alternatives to downloading, and an overall increase in the efficacy of our technologies (such as would be encountered in a linearly advancing state of the art) of 5%. While neither of these changes are a realistic model of exactly how things dynamically react in the real world, they do provide some interesting insight as to the general trends that these factors might have on the development of peer-to-peer file-sharing networks.



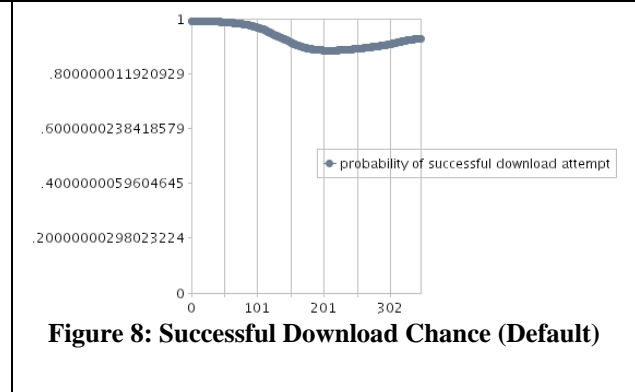
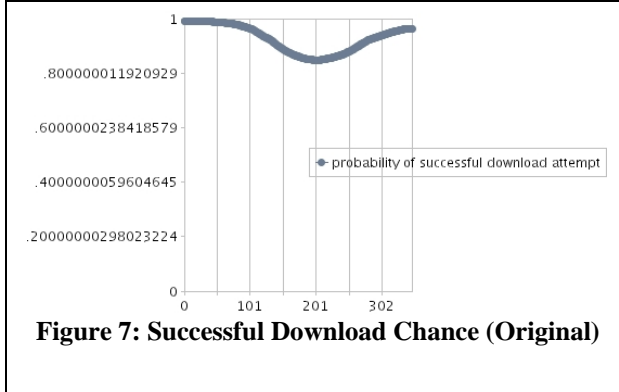
In the above two figures, we see the change caused in the pattern of network users caused by the variable manipulations in our default arrangement. The user peak is slower but higher, with a peak of around 58,000 as opposed to the 40,000 seen in the original. The fact that this increase is of a greater magnitude than either of our variable manipulations indicates the effects which the dynamic, reinforcing nature of the model have caused in synergy with our changes. In addition, the fall-off from the peak is significantly less sharp, indicating that not only would the users stock stabilize at a higher number, but also that it would stabilize at a higher proportion of the peak user figure than in the original model.



For this experiment, we see changes in the shared content versus that of the original that are very similar to the changes we saw with the network users. The peak is higher (proportionately by about the same amount), occurs later, and drops off more slowly. It is somewhat unsurprising that the content stock behaves in this way, due to the fact that the content is provided by the users on roughly an equal basis in this model. Therefore, the content will always follow the users.



Latency behaves in a somewhat less predictable manner. With our modifications, latency spikes up much more quickly, although it peaks earlier and never reaches quite the same height as that of the original. Conversely, it drops off less slowly, again in contrast with the behavior exhibited with both the stock of network users and that of shared content. These results show that the point of “latency crisis,” so to speak, is less severe with the better functioning network modeled with our modifications to the peer-to-peer system.



The final criterion that we will evaluate in this experiment is the chance of a successful download. The results here are the most puzzling, having in fact the opposite effect that we hypothesized. The probability of a successful download attempt with our modifications is, on the whole, less than that in the original model. Our model bottoms out around the same point in time, although to a lesser degree. However, the successful download chance with our default variable settings “recovers” to approach 1 much less slowly than in the original. It is possible that with a longer time scope that the default settings might allow the probability to eventually approach 1 more closely, but we do not have any true evidence of this with the current simulation.

It is hard to determine what the possible cause of this behavior might be or what ramifications this might have. The hypothesis that we have come up with is that with the wild period of fluctuation that we have observed in the growth behavior of peer-to-peer networks, especially in the content, users, and traffic stocks, that there is a period of “growing pains” where the network is performing at a sub-optimal level. When the stocks level out to a more constant level, the probability of successful download chance

becomes once more close to 1. This signifies that a mature, well-ensconced network is of more utility than a new one.

The final measurements noted in the Pavlov-Saeed simulation, and so reproduced here, are the numbers of minimum and maximum traded content, and also minimum and maximum online nodes. They are reproduced in the table below.

Simulation	Min. TC	Max. TC	Min. Nodes	Max. Nodes
Original	3,708	63,961	2,000	39,255
Default	3,708	91,611	2,000	55,405

These numbers clearly indicate two things. First, the minimum traded content and minimum online nodes features are either set to an artificial minimum in the model (this appears to be the case with the minimum online nodes), or are simply not affected by the changes we implemented in the model to simulate the economic reduction in free-riding and the effect of improved technology. The latter seems to be the case in the matter of the minimum traded content due to the odd number, but it is also possible that the minimum traded content number is a derivation of some hard-coded minimum in the original model.

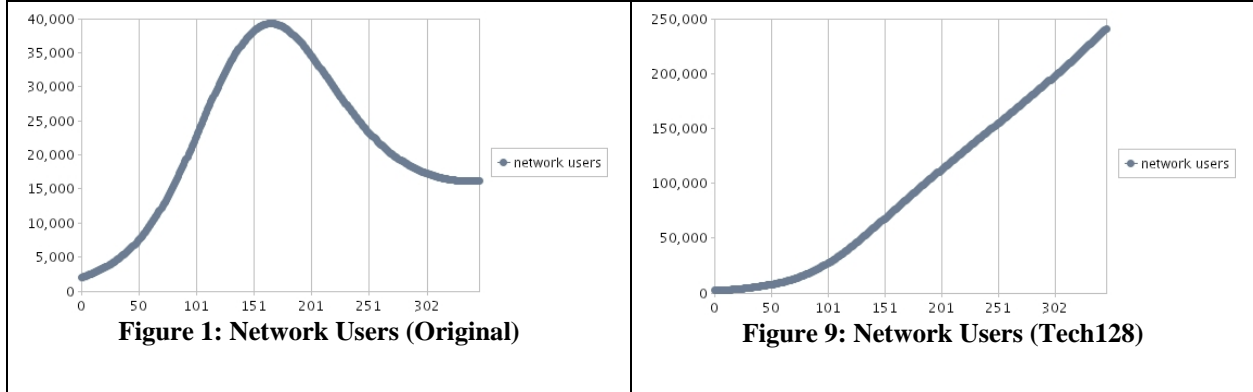
The second thing indicated by these numbers is present in both the maximum traded content and maximum online nodes figures—mainly, that the network grew. This is totally consistent with our other findings, and thus unsurprising. The maximum traded

content grew to a degree greater than did the maximum online nodes, thus perhaps meaning that users are sharing more per capita.

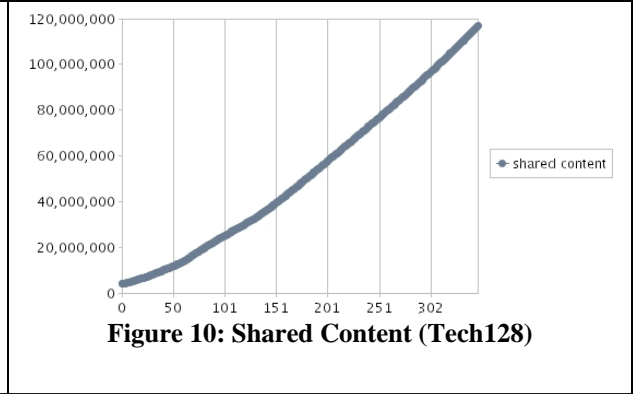
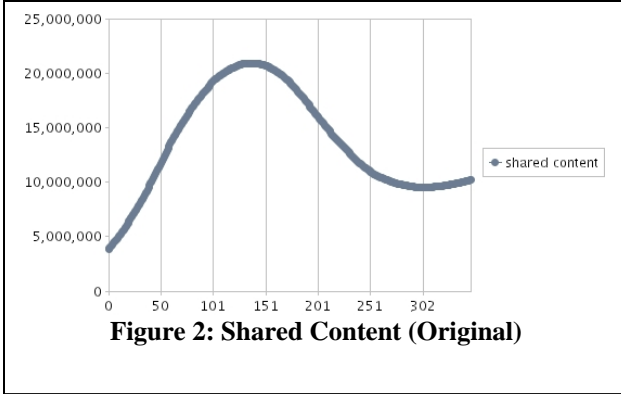
In conclusion, the model did indeed behave exactly as designed and expected for this experiment. The economically-dictated 2% reduction in free-riding, combined with the 5% improved technology factor, led to a richer network with a greater number of users, greater amount of content, lower latency, and more content shared per user. These findings would be consistent with a real-world situation in which the economy was on an upswing (or CD prices were lowered) and technology continued to advance at a fairly slow-to-moderate pace.

Scenario: The Technological Breakthrough

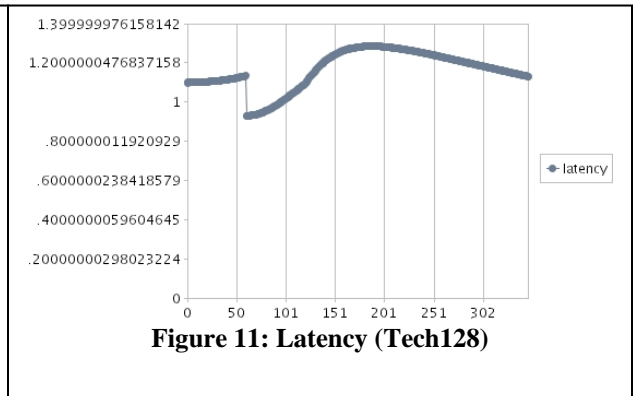
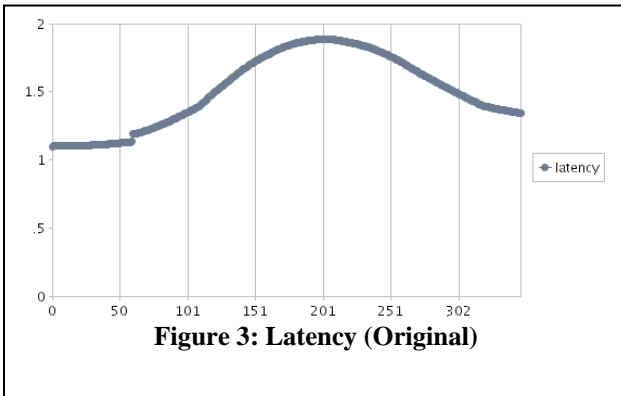
By far the most interesting experiment that we conducted on the model was what we called the “technological breakthrough.” In this experiment, we attempted to model at what point a technological breakthrough would significantly change the behavior of the entire system, and what these affects would be. For this experiment, our economic free-riding scalar is set to 1, thus negating its effect. The value that we found best shows the “threshold behavior” or change in dynamics is to set the improved technology scalar to a value of 1.28.



It is immediately apparent the vast and far-reaching effects initiated by the relatively moderate 28% multiplier that we applied to key variables for this scenario. The entire behavior of the system is different. Instead of exhibiting behavior consistent with the limits to growth archetype, this system appears to grow without bound. Note that the slower progress in the first 60 or so steps is probably actually caused by the Forio model's method of always using the default variables for the first 60 steps. In actuality, the growth is probably close to linear from the start. The dynamic nature of the model is once more exhibited in the manner in which a fairly small change is so magnified by its interactions with other parts of the model that the network users, instead of peaking at around 40,000 and leveling off, continue to grow without limit up to 250,000! What must have happened here from a system dynamics perspective is that 1.28 was the point at which our technology modifier managed to overpower the balancing aspects of the peer-to-peer system (free-riding, etc.).

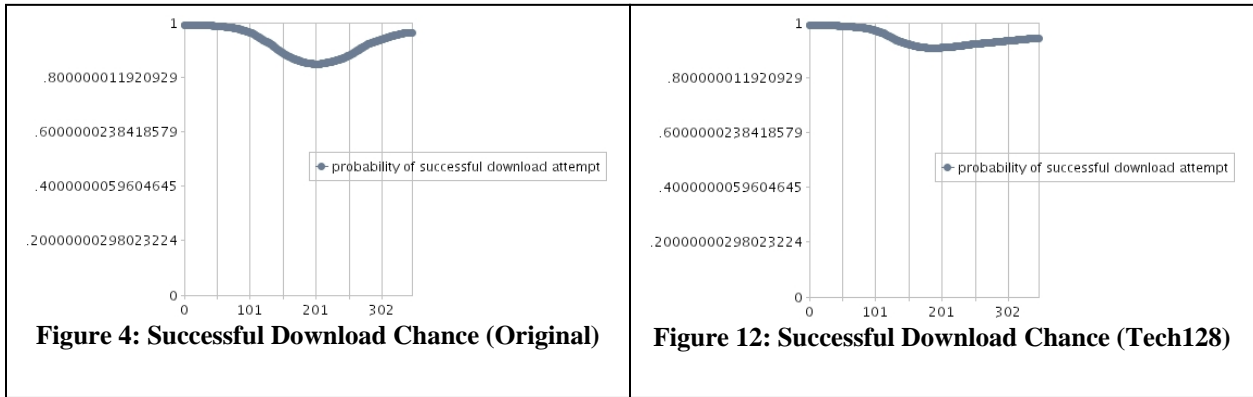


As in the first experiment, the behavior of the content mimics the behavior of the users. The linear growth pattern exhibited in Figure 10 is almost identical to that in Figure 9. Clearly, in the overall framework of the peer-to-peer model, regardless of what influence our variables have, content and users are linked in a fairly direct correspondence.



As previously noted, we see artifacts before the 60-steps point due to the way in which the Forio model behaves. The behavior of latency is similarly unpredictable in this experiment as it was in the last one. It remains below 1 until around the 100-step mark, at which point it rises to a peak (“growing pains”) of around 1.35 at about the 180-step mark. After this point, it falls off linearly. This indicates that it may not, or at least not in

the immediate future, level out. In this improved technology scenario, it is possible that latency will stabilize at a point near or below 1, a huge reduction in latency versus the original two simulations.



Once more, we are faced with enigmatic results in the area of the successful download probability. While the bottom-out effect is significantly lessened in our technological breakthrough scenario model, it is still far slower to recover than the original model. This serves as an affirmation of our hypothesis for the reason of this behavior in the first experiment. It is, however, worthy of note that the change to the chance of successful download was more incremental than anything else, despite the fantastical changes caused by this technology modifier in both the users and content stocks.

Simulation	Min. TC	Max. TC	Min. Nodes	Max. Nodes
Original	3,708	63,961	2,000	39,255
Tech128	3,708	425,393	2,000	241,005

The findings with the traded content and online nodes figures are precisely as expected. Both of the minimum figures remain at their consistent levels throughout all of the simulations, indicating a hard-coded value of some sort or change beyond the scope of our experiments. Meanwhile, both of the maximum figures experienced huge increases. Unlike in our first experiment, the maximum traded content did not increase in greater proportion than did the maximum online nodes. This corroborates with the fact that we did not change the economic scalar of free-riding for this scenario; therefore users are not sharing more per capita.

Analysis of the Technological Breakthrough

Scenario

The implications of the experiment are many. While it is undeniably true that our approach to modeling technological improvements is basic at best, we feel that it does show in some way some of the likely effects. In this case, a mere 28% improvement in technology-related factors has caused this peer-to-peer network to spin out of control. Users and content are growing linearly without bound, indicating that we are moving towards this peer-to-peer network being an unstoppable monolith of data traffic, monopolizing the entire internet.

Everyone would use this peer-to-peer network for satisfying almost all of their media needs, and this would have far-reaching impact upon the economy, putting mass

media conglomerates out of business and possibly in fact stopping the production of new media. It would be a sad day indeed if the file-sharing networks grew so powerful that it was impossible for artists, producers, actors, and directors to make any sort of profit whatsoever on their works, and therefore to stop producing new work except perhaps occasionally for free.

What's more is how small of a technology factor increase that it took to cause these changes. While we cannot, in the scope of this course and this report, properly analyze our "28%" figure and track it to certain time-based patterns of technological growth, it seems clear that it is only a matter of a few years, a decade at most, before the state of the art advances sufficiently to cause this cataclysm.

This represents a clear dictate for the recording and movie industries: they absolutely must remove the incentives towards peer-to-peer file-sharing. While the media organizations have preferred to attempt this in the manner of disrupting the networks and bringing the force of the law to bear on the file-sharers, these devious technological schemes are rather ineffective except as a scare tactic—peer-to-peer is already so huge that the RIAA and MPAA are essentially powerless to directly impact the users in a meaningful way.

Furthermore, technological problems almost always have technological solutions. Tactics such as bogus downloads can easily be defeated by the software involved with the peer-to-peer file-sharing, or by simply introducing a new file-sharing paradigm. One example of this is Direct Connect, in which users connect to a central "hub" which provides search services and connectivity between all of the users on the hub. To gain

admittance to the hub, one generally needs an account on the hub with a password, which is usually only obtainable by knowing either the administrator of the hub or some other user within the hub who will vouch for you. Using this technique, it is simply impossible for the RIAA or MPAA to create any bogus downloads since they are not permitted entrance to the hub.

What the media industry must do, instead, is to remove the incentive behind file-sharing in the first place. They must offer a real alternative to it: legal, high quality music/movies/etc. at a price people are willing to afford. Few people actually want to be a criminal and steal things, even data, however it is simply no contest between downloading an OK-quality copy of a song for free, even if it takes a while, than to buy a \$18 CD just for one song. Media industry pricing would be moderated in this way by simple economics, if they were not organized into the cartel oligopoly that they are today. Instead, the price is set artificially, and peer-to-peer file-sharing is simply one of the externalities of this situation. People on the demand curve below the price point (which we would hypothesize includes many people, and almost all college students, for example) simply turn to file-sharing over the internet to get their desired music, movies, etc.

It is only through this reasoned economic leverage that the media companies will be able to institute a system under which all, or at least most, people will see that it is to their advantage to obtain their media directly from the companies rather than by illegal peer-to-peer downloading. This keeps money flowing in to the media creation and distribution process, and allows new media to be created. Like it or not, any media is

reproducible: if we can see or hear it, we can copy it. Thus, there will always be a bootleg alternative to legitimate media purchase. If the media companies continue to price media beyond what people are willing to pay, the simple supply/demand dynamics of the situation will topple the entire media industry.

Conclusions

We can come to some conclusions based upon the findings of our experiments and our analysis of the peer-to-peer model. Clearly, free-riding and technological improvements have huge effect on the dynamics of peer-to-peer file-sharing. Furthermore, the progress of peer-to-peer networks in and of itself creates economic pressures which are, in a way, counterweights to those imposed on the media market artificially by the big media conglomerates. Peer-to-peer in this respect serves as a sort of economic pressure towards equalization and equilibrium.

The most interesting conclusion we came to, of course, was facilitated by our “Technological Breakthrough” scenario. The results certainly seem to indicate that it is up to the media companies to save media. If they do not lower the prices of media, surely the growth of peer-to-peer networks implicated by the ever-present marching on of technology will force their hand.

To run our simulation or see directions on its use, please visit:
<http://broadcast.forio.com/sims/ptopmods/>