Caracterização de carga de redes *Peer-to-Peer*

Juliano Santos, Leonardo Rocha, Diêgo Nogueira, Paulo Araújo, Virgílio Almeida, Wagner Meira Jr.
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
{juliano,lcrocha,diego,pauloh,virgilio,meira}@dcc.ufmg.br

15 de abril de 2002

Palavras-chave: Sistemas distribuídos, *Peer-to-Peer*, Caracterização, Gnutella

Resumo

As redes *Peer-to-Peer* surgiram recentemente com a proposta de viabilizar o compartilhamento de recursos computacionais via Internet. Entretanto, ainda se conhece muito pouco dessas redes. Em função de algumas de suas características, torna-se necessária a adoção de estratégias específicas para caracterização das mesmas. Apresentamos neste artigo uma metodologia de caracterização de carga de redes *Peer-to-Peer* fundamentada nessas características. Para validar a metodologia, realizamos um estudo de caso com a rede Gnutella, obtendo resultados relevantes sobre a mesma. Identificamos, dentre outras, as distribuições estatísticas que caracterizam os arquivos compartilhados, a disponibilidade dos *servents* e a freqüência das palavras nas consultas da gNet.

Abstract

Peer-to-Peer networks came up recently with the goal of making possible the sharing of computacional resources via Internet. So far, there is no definitive characterization of the behavior fo these networks. Because of some of their characteristics, the adoption of specialized strategies for their characterization becomes necessary. We present in this article a workload characterization methodology for Peer-to-Peer networks that address these characteristics. To validate the methodology, we performed a case study with Gnutella network, getting important results out of it. We identify, among others, the statistical distributions that characterize the shared files, the availability of the servents and the frequency of the words in the search patterns at gNet.

1 Introdução

O crescimento e a popularizaąão da Internet, juntamente com a evolução da indústria de *hardware*, resultou em um ambiente no qual vários clientes, conectados através da Internet, usam apenas uma fração de seu poder computarional [Var01]. Uma vez que a maior parte da carga nas atuais tecnologias na Internet se concentra nos servidores, existe nos clientes uma capacidade significativa de processamento e armazenamento de dados sub-utilizados [HC99].

A partir dessa constatação, começou a ser discutida, no final da década de 1990, uma nova arquitetura, alternativa à tradicional cliente/servidor, capaz de viabilizar o compartilhamento dos recursos das máquinas cliente, de forma a possibilitar uma maior utilização do seu poder computacional sem necessariamente aumentar o parque instalado [HC99]. A proposta é que os computadores conectados à Internet passem todos a exercer o papel tanto de clientes como de servidores, configurando o que vem se chamando de uma arquitetura *peer-to-peer* (ou simplesmente **P2P**) [Shi00].

Uma arquitetura P2P tem como definição os tipos de recursos que são compartilhados, como processador, memória e disco, e o protocolo de comunicação entre os seus integrantes. Uma rede P2P é uma instância de uma arquitetura P2P, composta por um conjunto de computadores que implementam as especificações dessa arquitetura. Em função de exercerem papéis tanto de clientes como de servidores, os computadores integrantes de uma rede P2P são comumente denominados *servents*.

O principal diferencial de arquiteturas P2P é a sua capacidade de permitir a implementação de aplicações distribuídas colaborativas, no sentido de que os servents que as compuserem teoricamente somem seus recursos para criar uma grande máquina distribuída, beneficiando-se mutuamente. Embora recentes, são crescentes as discussões sobre as aplicações possíveis para arquiteturas P2P [CGM00, Sul00] e sobre os seus potenciais ganhos e eventuais perdas [AH00]. São especialmente raros os trabalhos sobre redes P2P calcados em dados reais, permanecendo em aberto a real avaliação da sua eficácia em operação.

Este cenário motivou-nos a desenvolver uma metodologia para caracterização de carga de redes P2P, trabalho este inédito até então. Há várias metodologias de caracterização de carga propostas para a arquitetura cliente/servidor. No entanto, as redes P2P têm algumas características que tornam necessária a adoção de estratégias específicas. A primeira delas é o processamento completamente descentralizado, ou seja, não existe um centralizador de informações ou de tráfego. A segunda característica é a atuação variável do servent, podendo se comportar tanto como cliente quanto como servidor. A terceira e última é o dinamismo da rede P2P. Não se tem, em nenhum instante, a garantia de disponibilidade de um dado servent, ou seja, não se tem garantias de que ele estará no ar quando se desejar acessá-lo.

Este trabalho apresenta, portanto, uma metodologia para caracterização de carga de redes P2P, a qual, conforme mostraremos, contempla as três características mencionadas. Por caracterização de carga, entenda-se a caracterização dos recursos compartilhados e do tráfego de mensagens na rede para utilização desses recursos, ou seja, a carga gerada pela atividade da rede P2P. Adotamos a divisão da caracterização entre os lados cliente e servidor dos servents. Em função da arquitetura cliente/servidor ser amplamente estudada, a adoção dessa estratégia possibilita reduzir nosso trabalho a dois problemas conhecidos que são a caracterização de clientes e servidores em redes cliente/servidor. A metodologia introduz também um mecanismo de coleta de dados baseado na própria arquitetura P2P. Nele, informações da rede são obtidas a partir da observação do tráfego que circula pela rede e da geração de requisições específicas, por meio da implementação de um servent adaptado que coleta esses dados da rede.

Todas as características das arquiteturas P2P são verificadas na medotologia proposta. A primeira, referente ao processamento descentralizado, e a terceira, do dinamismo da rede P2P, são atendidas pelo agente coletor de dados, que será descrito na seção 2. A segunda, que se refere à atuação variável dos servents é atendida pela estratégia de caracterização, detalhada nas seções 2.1 e 2.2, que segmentou os servents nas suas

características de clientes e de servidores.

Para validar a metodologia, elegemos uma rede P2P específica para conduzirmos o trabalho de caracterização. A Gnutella [Kup00, Ora00] foi a escolhida, por uma série de razões. A primeira razão foi se enquadrar entre a classe de aplicações P2P em que se observa maior crescimento, a de compartilhamento de arquivos. Outra razão é o fato de a Gnutella apresentar as três características de uma rede P2P. O Napster, por exemplo, outra arquitetura P2P de compartilhamento de arquivos muito popular, utiliza um índice centralizado [Ora00]. Por fim, a rede Gnutella tem um tráfego intenso, com volume de dados suficiente para subsidiar nossa caracterização.

Assim, este trabalho apresenta uma estratégia de caracterização para a rede Gnutella (gNet), que pode entretanto ser estendida para outras redes P2P. A aplicação desta estratégia a outras redes P2P depende do desenvolvimento de um coletor baseado na arquitetura que a rede implementa e da definição das variáveis a serem observadas nos aspectos de cliente e servidor desta rede. A estratégia proposta é aplicada sobre uma carga real da gNet, proporcionando algumas conclusões importantes sobre o funcionamento e o comportamento dos participantes desta rede como um todo.

O trabalho está organizado da seguinte forma: a próxima seção apresenta a metodologia de caracterização de carga de redes P2P proposta; a seção 3 apresenta a aplicação dessa metodologia à rede Gnutella, com a descrição da implementação realizada e o relato dos resultados observados nos experimentos realizados; por fim, a seção 4 apresenta algumas conclusões e trabalhos futuros.

2 Caracterização de carga de redes P2P

Nesta seção é apresentada uma metodologia de caracterização de carga, descrevendo as suas abordagens, características básicas e especificidades relacionadas à rede P2P.

Os servents de uma rede P2P não devem ser tratados isoladamente, e sim como componentes de uma máquina virtual distribuída, cujos objetivos e características são mais amplos do que os de cada servent específico. As redes P2P de compartilhamento de arquivos, por exemplo, têm por objetivo prover um repositório distribuído de arquivos, que permita a rápida disseminação e acesso à informação, além de prover mecanismos para mantê-la e preservá-la [CGM00, CSWH00]. As redes de processamento distribuído, por outro lado, pretendem construir servidores distribuídos sobre a Internet, usando os ciclos de processador e a memória disponíveis nos computadores conectados à rede para processar tarefas específicas [HC99].

De forma análoga aos seus objetivos, as características de uma rede P2P devem ser observadas como um todo. Diferentemente de uma rede cliente/servidor, em que normalmente se concentram nos servidores informações suficientes para se caracterizar o serviço e a comunidade de usuários, em P2P é necessário obter informações de todo o conjunto de servents para se conhecer as características de compartilhamento e uso dos recursos da rede [Var01]. Isto acontece em função do processamento ser totalmente descentralizado nas redes P2P, sendo esta uma das características fundamentais dessa arquitetura.

Como não existem servidores dedicados, da mesma forma que na arquitetura cliente/servidor, não há também nenhum compromisso dos servents com relação à disponibilidade, ou seja, não se pode estimar quando e por quanto tempo um servent estará conectado à rede. Argumenta-se [CGM00, CSWH00], entretanto, que a redundância oriunda da capacidade de processamento de todos os servents permite à rede manter

a disponibilidade dos seus serviços como um todo por meio da combinação das disponibilidades dos servents que a compõem, através de políticas de replicação. Esse dinamismo das redes P2P é outra característica fundamental da arquitetura.

Aliadas à capacidade dos servents de exercerem papéis tanto de clientes como de servidores, estas características são a base para a metodologia de caracterização de carga de redes P2P proposta neste trabalho. Nosso objetivo é permitir a sistematização do estudo dessas redes, norteando estudos e avaliações das mesmas. Assim, os grupos de pesquisa nessa área poderão contar com caracterizações baseadas em dados reais para observar e compreender ao que essas redes se prestam e como estão funcionando, contrapondo ao que se esperava das mesmas quando essas arquiteturas foram concebidas.

A metodologia divide-se em duas partes: caracterização cliente e caracterização servidor, em que são propostos dois conjuntos de variáveis para caracterizar, respectivamente, o uso e o compartilhamento dos recursos em uma rede P2P. Essas variáveis definem os requisitos para o desenvolvimento de um agente coletor de dados, o qual é um servent automatizado que, além de observar o tráfego espontâneo da rede, gera cargas programadas para obter informações específicas da mesma, coletando dados do tráfego gerado em resposta. Estas duas partes da metodologia são descritas em detalhes nas seções 2.1 e 2.2, respectivamente.

2.1 Caracterização cliente

O lado cliente dos servents de uma rede P2P trata das funcionalidades ligadas à utilização dos recursos da rede, ou seja, trata do lado consumidor dos servents. Ele tem duas características básicas: i)utiliza os recursos disponibilizados por outros servents da rede, e ii)gera as requisições das mensagens que trafegam na rede.

Além de usuário dos recursos compartilhados na rede por outros servents, o lado cliente é responsável pela geração da carga primária de mensagens na rede. Em outras palavras, ele é responsável por gerar as mensagens iniciais de todas interações entre servents, ou seja, é a parte ativa do servent. Por conseqüência, é a parte que define quais são os recursos demandados da rede e é a parte que precisa saber como obtê-los.

Assim, para a caracterização do lado cliente dos nós, abordamos dois conjuntos de variáveis:

Caracterização de demanda: estas variáveis definem os interesses dos servents de uma rede P2P, ou seja, os recursos que eles solicitam à rede. Em uma rede de distribuição de arquivos, essas variáveis poderiam tratar em alto nível dos tipos de arquivos ou temas de interesse do servent ou, mais especificamente, da relação de arquivos mais requisitados na rede. A caracterização dessas variáveis possibilita o mapeamento da demanda de recursos em toda a rede.

Caracterização do padrão de interação: estas variáveis tratam da forma como um servent consegue se inserir na rede, da visão que ele tem da mesma e da forma como ele pode "navegar" por ela, em busca de recursos. Elas estão diretamente relacionadas à vizinhança de cada um dos servents e à forma como os servents enviam e propagam suas mensagens. A caracterização dessas variáveis deve possibilitar o conhecimento da conectividade da rede e do volume de mensagens que deve trafegar para obtenção de um recurso.

A definição e avaliação desses dois conjuntos de variáveis possibilita a caracterização do lado cliente de uma rede P2P. Para cada rede que se pretende caracterizar, é necessário, portanto, identificar os dados necessários para se obter essas informações, e programar o agente coletor de dados para obtê-los. Na seção 3.2, apresentamos esse estudo para a gNet.

2.2 Caracterização servidor

O lado servidor dos servents de uma rede P2P está relacionado às funcionalidades ligadas ao provimento dos recursos da rede, ou seja, do lado fornecedor dos servents. Ele tem duas características básicas: i)disponibiliza e mantém recursos que podem ser utilizados pelos demais servents da rede, e ii)processa as requisições de outros servents, provendo aos outros servents os recursos solicitados.

O lado servidor dos servents é, portanto, responsável pela oferta de recursos na rede e pelo atendimento de requisições de outros servents, provendo acesso a esses recursos. Em outras palavras, ele é o provedor de recursos de cada servent, subdividindo-se entre o conjunto de recursos e a unidade de processamento das requisições.

Assim, para a caracterização do lado servidor dos servents, abordamos os dois seguintes conjuntos de variáveis:

Caracterização de oferta: estas variáveis definem o perfil de fornecimento dos servents de uma rede, os recursos ou tipos de recursos que eles provêem à rede. Em uma rede de distribuição de arquivos, de forma análoga ao descrito para o lado cliente, essas variáveis se referem tanto aos tipos ou temas dos arquivos como à relação de arquivos disponíveis na rede. A caracterização dessas variáveis possibilita o mapeamento da oferta de recursos em toda a rede.

Caracterização da capacidade de atendimento: estas variáveis descrevem a forma como os servents conseguem processar as requisições e prover os recursos solicitados pelas mensagens que chegam da rede. Elas estão diretamente relacionadas aos mecanismos de escalonamento de uso dos recursos e à capacidade computacional dos servents. A caracterização dessas variáveis deve possibilitar a análise da capacidade e da carga dos servents enquanto servidores de recursos, as quais estão diretamente ligadas à capacidade do servent como um todo.

Note que estas variáveis são complementares às utilizadas na caracterização cliente dos servents. A definição, avaliação e análise delas possibilita a caracterização do lado servidor de uma rede P2P. Em conjunto com as variáveis estudadas pela caracterização cliente, elas cobrem todo o funcionamento dos servents de uma rede P2P.

Como na caracterização cliente, para cada rede que se pretende caracterizar, é necessário identificar os dados necessários para se obter essas informações e programar o agente coletor de dados para obtê-los. Na seção 3.2 exemplificamos esse processo com a gNet.

3 Caracterização de carga da gNet

A gNet [Kup00] é uma rede P2P aberta, com aplicações clientes implementadas por várias equipes na Internet. A compatibilidade entre os clientes é garantida por um protocolo de comunicação simples definido pela arquitetura.

A gNet é composta por servents autônomos. Cada servent tem uma relação de arquivos, os quais são o conjunto de recursos disponibilizados para a rede. Assim, os recursos compartilhados na gNet são arquivos. Cada arquivo é identificado por um texto, e está disponível para *download* por outros participantes da rede.

Um servent contacta outro por meio de uma mensagem do tipo PING, que é respondida pelo servent conectado com uma mensagem do tipo PONG. Os PINGS são propagados via sucessivos *broadcasts* para todos os vizinhos dos servents que a mensagem conseguir alcançar até uma determinada profundidade. Eles empregam um mecanismo análogo ao de TTL utilizado em roteamento de redes TCP/IP, sendo cada servent por onde a mensagem passa equivalente a um "*hop*" [Kup00, Rit00].

As consultas são feitas por meio de mensagens do tipo QUERY, enviadas aos vizinhos do servent requisitante de forma análoga ao PING, utilizando o mecanismo de TTL. A consulta é especificada por um texto e processada como uma busca aproximada sobre os nomes dos arquivos. Os servents que disponibilizam arquivos que contêm o texto procurado retornam uma mensagem do tipo QUERYHIT, com os nomes dos arquivos encontrados.

O *download* dos arquivos é feito por meio de uma conexão HTTP direta entre o servent requisitante e aquele selecionado dentre os que disponibilizam o arquivo. A mensagem de PUSH substitui essa conexão HTTP nos casos em que o servent que está provendo o arquivo está protegido por algum *firewall*.

A gNet tem um grande número de usuários no mundo todo, gerando um tráfego intenso [Cli00], e sendo normalmente referenciada como exemplo de rede P2P. A distribuição dos arquivos e o mecanismo de busca distribuído implementado pelo conjunto de servents são os seus principais diferenciais.

Esta seção apresenta a caracterização de carga da gNet segundo a metodologia proposta. Inicialmente, a seção 3.1 apresenta a implementação realizada para o agente coletor de dados, detalhando ainda o ambiente utilizado para realização do experimento. As seções 3.2 e 3.3 apresentam as caracterizações cliente e servidor da gNet.

3.1 Coleta de dados

Seguindo a metodologia proposta, implementamos, para a gNet, um agente coletor de dados. Utilizamos como base para essa implementação o código-fonte de uma aplicação Gnutella existente, aproveitando a implementação do protocolo Gnutella. A aplicação utilizada foi o Gnut, desenvolvido por um grupo independente de desenvolvedores coordenados por Robert Munafo[1], com código aberto na linguagem C[2].

Essa aplicação foi adaptada para gerar requisições automáticas segundo as especificações definidas pelas caracterizações cliente e servidor da gNet, conforme será detalhado nas seções 3.2 e 3.3. Os servents contactados foram obtidos a partir de uma relação de servents conhecidos na gNet, obtida no site http://www.gnutella.org [3]. Durante os experimentos, essa relação foi incrementada a partir de outros servents que contactaram o nosso servent, fato possível a partir da forma de propagação de mensagens na gNet [Kup00].

Os experimentos foram realizados a partir de duas estações de trabalho localizadas

[1] HTTP://HOME.EARTHLINK.NET/~MROB/PUB

[2] Ele pode ser obtido na URL HTTP://WWW.GNUTELLADEV.COM/SOURCE/GNUT ou ainda em HTTP://WWW.GNUTELLIUMS.COM/LINUX_UNIX/GNUT.

[3] A relação de servents não está mais disponível neste endereço.

no laboratório e-Speed, do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais. Essas estações utilizam o sistema operacional Mandrake, versão 7.2, uma das distribuições do Linux disponíveis no mercado. Elas têm conexão dedicada à Internet, via RNP, rede que abriga principalmente as instituições de ensino superior e centros de pesquisa brasileiros, o que possibilita a realização de experimentos contínuos sem restrições de conectividade.

Foram observadas variações no número de servents conectados à rede, no conteúdo das pesquisas e no volume de requisições realizadas no decorrer de um dia e de uma semana. As variações observadas durante uma semana foram detectadas, em menor escala, nas referentes a um dia, portanto, para que a base de dados coletada não atingisse um volume muito grande, inviabilizando o tratamento dos dados, optamos por realizar as análises sobre os dados coletados em um dia. Assim, os resultados apresentados e distutidos neste trabalho são relativos às 24 horas do dia 02/10/2001. Outros períodos análogos foram analisados, e resultados compatíveis ou similares foram obtidos.

Para que a localização do servent na rede não influenciasse nos dados obtidos, construimos uma base de endereços, proveniente de listas disponibilizadas na Internet, referenciando os servents mais conhecidos e que permanecem conectados por mais tempo na rede. Estes servents se encontram espalhados por todo o mundo. O servent adaptado contacta os endereços desta lista, evitando que os dados coletados apresentassem apenas características locais.

3.2 Caracterização cliente

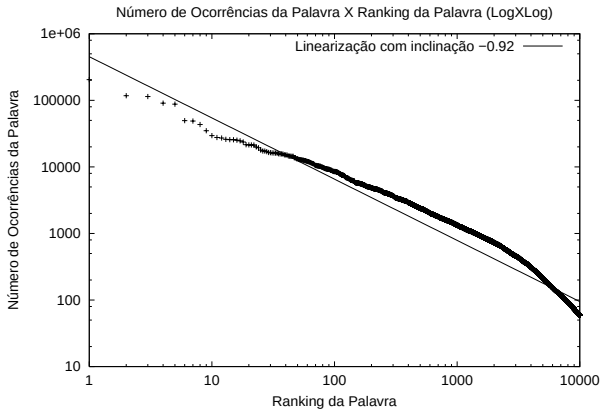
A caracterização cliente trata da utilização dos recursos compartilhados pela gNet, ou seja, da busca e obtenção de arquivos na rede. Seguindo a metodologia proposta, ela se subdivide entre os recursos que se demanda da rede, em que se caracteriza o perfil de consumo dos servents, e a forma como obter um recurso, em que se caracteriza os meios para se obtê-los.

3.2.1 Caracterização de demanda

Na Gnutella, os recursos compartilhados pelos servents são arquivos. Para se avaliar os arquivos demandados pelos servents da rede Gnutella, realizamos um conjunto de experimentos que tratam basicamente das mensagens de QUERY, utilizadas para buscar por arquivos. Nosso objetivo com esses experimentos é caracterizar a demanda dos servents, ou seja, quais os tipos de informações requisitadas por eles à rede.

Caracterizamos as palavras mais procuradas na rede, de forma a termos visões da popularidade desses termos e da distribuição de suas freqüências. O experimento que dá suporte a esta análise é passivo, trabalhando basicamente com a coleta das QUERYs recebidas pelo servent que desenvolvemos. Todas as QUERYs recebidas são consideradas nas análises.

O objetivo deste experimento é observar a distribuição de freqüência de ocorrência das palavras procuradas na gNet. Para tanto, foram armazenadas todas as QUERYs recebidas pelo nosso servent (2.992.390) durante 24 horas e extraídas as palavras distintas que as compunham (94.642). As freqüências (ou número de ocorrências) das palavras foram registradas também. De posse desses dados, foi construído um *ranking* de freqüência dessas palavras. As 10 palavras mais freqüentes podem ser visualizadas na Tabela (b) da Figura 1.



Posição	*Palavra*	*Ocorrências*
1	the	204.405
2	divx	117.563
3	avi	113.961
4	of	90.829
5	mp3	87.675
6	mpg	49.520
7	in	48.840
8	and	43.293
9	star	34.960
10	to	29.518

Figura 1: (a) *Ranking* de freqüência das palavras pesquisadas na gNet e (b) 10 palavras mais freqüentes

O gráfico (a) da Figura 1 apresenta a distribuição de freqüência do *ranking* de palavras, o qual se aproxima de uma distribuição de *Zipf* [Zip49] com inclinação igual a -0,92. Resultado similar a este foi apresentado em [Sri01].

3.2.2 Caracterização de padrão de interação

Na Gnutella, os arquivos são obtidos por meio de um mecanismo de busca distribuído, em que a consulta é inicialmente enviada a um conjunto de servents conhecidos, os quais processam a busca nas suas próprias bases de arquivos, retornando uma relação com os nomes dos arquivos encontrados, e propagam a consulta pelas suas próprias vizinhanças.

A revisão desse processo é importante neste contexto para destacar algumas características determinantes nesse mecanismo de busca. A primeira delas trata da conectividade dos servents, ou seja, das vizinhanças dos mesmos, as quais são determinantes para o alcance das consultas. O alcance de uma consulta mede quantos servents são contactados quando ela é enviada para a rede. O mecanismo de distribuição de consultas da Gnutella pode contactar dezenas de servents em uma busca, o que aumenta a possibilidade de sucesso na busca. No entanto, isso não necessariamente se traduzirá em velocidade na pesquisa. Um outro fator que deve ser considerado é o tempo de latência entre o envio e o retorno das mensagens. A composição desses fatores nos permitem analisar o processo de busca de arquivos como um todo. Em particular, analisamos a latência da comunicação entre os servents.

São dois os principais requisitos avaliados em um sistema de busca: o primeiro deles é a efetividade da busca, ou seja, se o usuário encontra ou não o que procura. A vizinhança e o alcance dos servents são determinantes no primeiro requisito, uma vez que determinam o tamanho da base de arquivos onde se fará a busca. No entanto, se o tempo de resposta das consultas aos servents não for tolerável, de nada adianta dispor de uma base infinita de arquivos. Para analisar esse requisito, realizamos o seguinte experimento para medir a latência entre o envio da mensagem e o retorno de sua resposta.

Mais especificamente, quantizamos a latência entre o envio de um PING e o retorno do respectivo PONG, de forma a examinar a latência de comunicação entre os servents da gNet. Para tanto, foram enviados PINGs com TTL=1 para todos os servents da

nossa base de endereços durante 24 horas perfazendo um total de 1.823.972 PINGs. Registramos, para cada PONG, o tempo transcorrido entre o envio do PING respectivo e a chegada da resposta. Com esses dados, construímos histogramas referentes à latência entre o PING e o PONG. Estes histogramas podem ser visualizados na Figura 2.

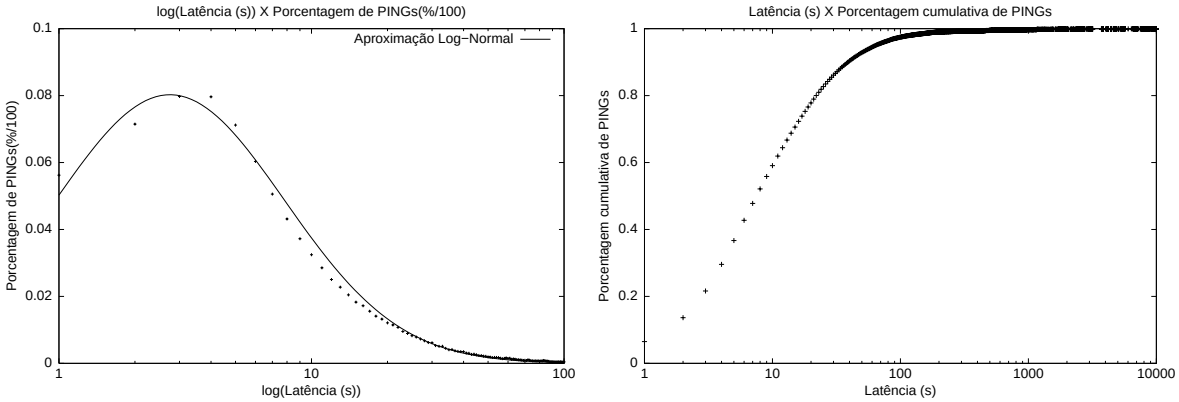


Figura 2: (a) Distribuição da freqüência da latência entre PINGs e PONGs e (b) Distribuição acumulada da latência

Analisando o gráfico (a) da Figura 2 observamos que a distribuição dos pontos se aproxima muito de uma curva da distribuição *Log-Normal* [Mat], cuja fórmula é:

$P(x) = \frac{1}{Sx\sqrt{2\pi}} e^{-\frac{(\ln(x)-M)^2}{2S^2}}$

A linearização realizada resultou nos valores de S e M iguais a 1 e 2,1, com variância de 0,18% e 0,12% respectivamente. Os dados utilizados para esta aproximação, compreendendo as respostas recebidas em até 200 segundos, abrangem 98,9% dos PINGs respondidos, conforme pode ser observado no gráfico (b) da Figura 2.

3.3 Caracterização servidor

A caracterização servidor trata do provimento dos recursos compartilhados pela gNet, ou seja, do compartilhamento de arquivos e processamento das requisições de outros servents da rede. Seguindo a metodologia proposta, ela se subdivide entre os recursos que se provê à rede, em que se caracteriza o perfil de provimento dos servents, e a forma como processar as requisições por um recurso, em que se caracteriza os meios para se provê-los.

3.3.1 Caracterização de oferta

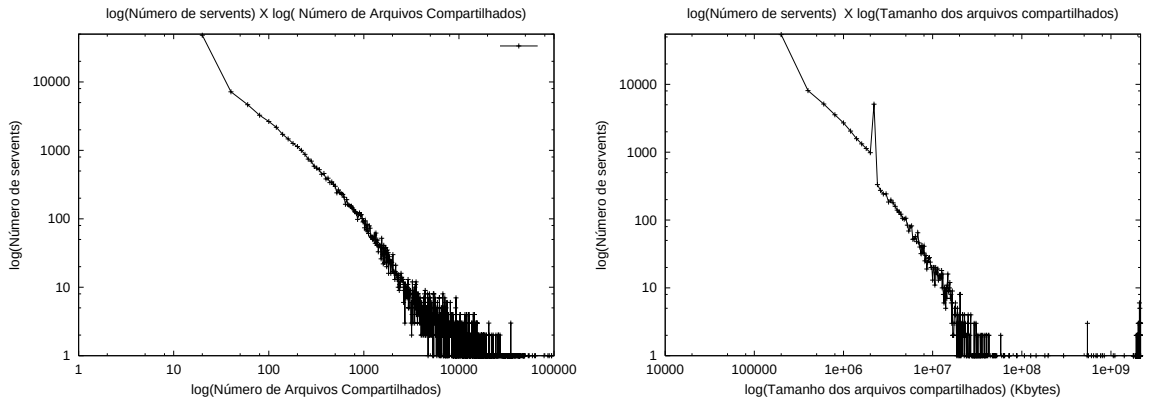
Na Gnutella, cada servent utiliza uma determinada área em disco para compartilhar um conjunto de arquivos para a rede. O número de arquivos, o espaço em disco e os tipos de arquivos disponibilizados são características independentes de cada servent. Portanto, não há na rede uma base de informações sobre o conjunto total de recursos compartilhados, seja quantitativa, tratando de números, ou qualitativa, tratando da diversidade de conteúdos.

Assim, para caracterizar a oferta de arquivos na rede Gnutella, tratamos de duas variáveis basicamente: volume de dados e tipos de conteúdo dos servents que compõem a rede. Cada uma delas é analisada por uma série específica de experimentos, descritos a seguir.

3.3.1.1 Volume de dados disponibilizados pelos servents A rede Gnutella dispõe de um mecanismo para divulgação de informações sobre volume de arquivos disponibilizados por meio das mensagens PING e PONG. Todo PONG carrega o número de arquivos e o volume de dados compartilhados pelo servent que o enviou. Nós utilizamos esse mecanismo para avaliar o volume de dados compartilhado na rede Gnutella.

O objetivo deste experimento é caracterizar o número de arquivos e o volume de dados compartilhados pelos servents. Para tanto, foram enviados PINGs para todos os servents da nossa base de endereços (120.535) durante 24 horas. Registramos, para cada um dos servents que nos enviaram respostas (90.282), os números de arquivos e quantidades totais de Kbytes compartilhados.

De posse desses dados, construímos um histograma do número de arquivos compartilhados pelos servents (Figura 3(a)) e um histograma do volume em Kbytes compartilhados pelos servents (Figura 3(b)). Visualizadas em escala LogXLog, ambas apresentam comportamento linear, configurando *power laws* [Ada00].



(a) Número de Arquivos

(b) Quantidade de Kbytes

Figura 3: Distribuição dos Arquivos pelos servents na gNet

Observa-se nos gráficos da Figura 3 que a grande maioria dos servents compartilham poucos arquivos, de tal forma que 73% dos servents não compartilham mais do que 100 arquivos e 600 Mbytes. Por outro lado, há alguns servents que compartilham mais de 90.000 arquivos e 2.000 Gbytes. Assim, embora seja uma rede P2P totalmente descentralizada, observa-se na gNet uma grande maioria de servents que compartilham poucos arquivos, e alguns poucos que centralizam grandes volumes de dados. Esse fenômeno já fora anteriormente relatado em [AH00]. O nosso trabalho contribui nesta questão ao identificar a distribuição *power law* presente tanto no número de arquivos quanto na quantidade de Kbytes compartilhados pelos servents da gNet, complementando esse trabalho anterior.

A análise dos dois gráficos indica ainda a existência de uma relação de proporcionalidade entre o número de arquivos e a quantidade de Kbytes compartilhados, em função das curvas apresentarem desenhos semelhantes. Para verificar essa relação, construímos um gráfico em que cada ponto representa o número de arquivos e a quantidade de Kbytes de um servent específico, o qual pode ser observado na Figura 4.

Observando o gráfico da Figura 4 pode-se perceber que os pontos convergem para uma reta, ou seja, convergem para uma relação linear entre o número de arquivos

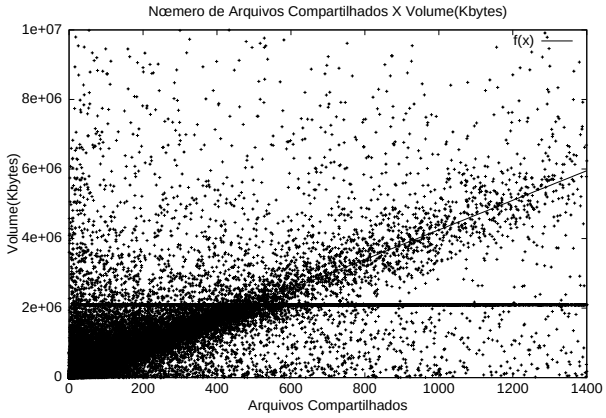


Figura 4: Relação entre número de arquivos e quantidade de bytes compartilhados pelos servents

compartilhados e o tamanho desses arquivos. A aproximação desses pontos gerou a reta exibida no gráfico, cujo coeficiente de inclinação é 4.250. Ou seja, o tamanho médio dos arquivos compartilhados na gNet é de 4.250 Kbytes.

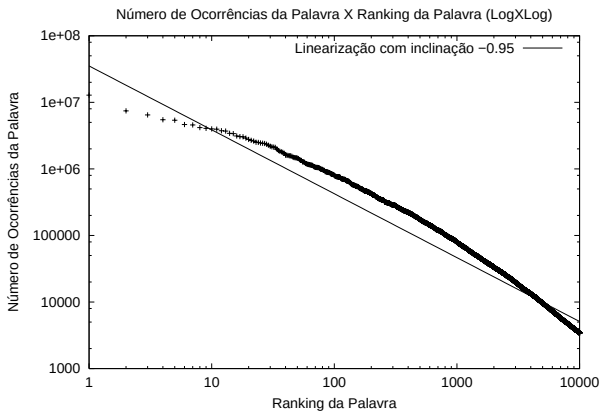
Em ambos os gráficos que tratam quantidade de Kbytes compartilhados, observa-se uma concentração de pontos em torno de 2.097.151 Kbytes, fora dos padrões descritos (um pico e uma reta, respectivamente). Esses desvios podem estar associados a problemas de representação de números inteiros nas implementações dos servents. A representação padrão de inteiros utiliza 32 bits, a qual limitaria os volumes de dados em bytes a 2147483647, e, em Kbytes, a 2.097.151.

Caracterizamos as palavras mais presentes nos nomes dos arquivos compartilhados na rede, de forma a observarmos a distribuição de suas freqüências. A principal diferença entre este experimento e o proposto para análise dos interesses dos servents é que este é ativo, enviando uma série de consultas para a rede por meio de QUERYs com TTL 7 e analisando os nomes dos arquivos de resposta recebidos nos QUERYHITs.

O objetivo deste experimento é observar a distribuição de freqüência de ocorrência das palavras que compõem os nomes dos arquivos retornados pelas consultas da gNet, assim, não foi analisado neste experimento o valor semântico das palavras. A análise que leva em consideração o conteúdo semântico destas palavras, objetivando categorizar os servents de acordo com o conteúdo dos arquivos compartilhados, foi realizada e será apresentada em trabalhos futuros.

Para atingir o objetivo deste experimento aqui proposto, foram armazenados todos os QUERYHITs recebidos pelo nosso servent (1.090.388) durante 24 horas e extraídas as palavras distintas que os compunham (67.160). As freqüências (ou número de ocorrências) das palavras foram registradas também. De posse desses dados, foi construído um *ranking* de freqüência dessas palavras. As 10 palavras mais freqüentes podem ser visualizadas na Tabela (b) da Figura 5.

O gráfico (a) da Figura 5 apresenta a distribuição de freqüência do *ranking* de palavras, presentes nos nomes dos arquivos compartilhados, o qual se aproxima de uma distribuição de *Zipf* [Zip49] com inclinação igual a -0,95, de maneira análoga ao observado no gráfico da Figura 1.



Posição	*Palavra*	*Ocorrências*
1	the	12.812.431
2	sex	7.443.177
3	of	6.479.737
4	porn	5.484.807
5	and	5.394.169
6	xxx	4.642.787
7	teen	4.570.437
8	britney	4.176.157
9	in	4.074.423
10	madonna	3.986.030

Figura 5: (a) *Ranking* de freqüência das palavras que compõem os nomes dos arquivos na gNet e (b) 10 palavras mais freqüentes

3.3.2 Caracterização da capacidade de atendimento

A caracterização dos recursos que são oferecidos à rede permite o dimensionamento do volume de arquivos disponíveis na Gnutella. No entanto, essa informação não pode ser traduzida em capacidade de compartilhamento de recursos. Outros recursos, como capacidade de processamento e conectividade dos servents que provêem os recursos, são determinantes na capacidade de atendimento às requisições da rede, sendo também limitadores para a capacidade de compartilhamento de arquivos na rede Gnutella.

Nesse contexto, analisamos algumas variáveis que demonstram essa capacidade de atendimento: disponibilidade dos servents, volume de requisições recebidas e o volume de requisições atendidas.

3.3.2.1 Disponibilidade dos servents Os *servents* de uma rede P2P como a Gnutella tem a peculiaridade de não terem compromissos com disponibilidade, ou seja, eles podem entrar e sair da rede a qualquer momento. Para analisar a disponibilidade dos servents na rede Gnutella, realizamos um experimento onde enviamos PINGs e QUERYs periodicamente (a cada 200 segundos) para todos os servents da nossa base de endereços (120.535) durante 24 horas. Registramos, para cada um dos servents que nos enviaram alguma resposta (79009), o tempo total disponível durante o experimento, a partir das respostas enviadas para as mensagens que geramos. De posse desses dados, construímos o histograma do tempo de disponibilidade dos servents.

A distribuição obtida, exibida no gráfico (a) da Figura 6 em escala LogXLog, é uma *power law*[Ada00], conforme pode ser observado pelo seu comportamento linear nessa escala. Observa-se uma dispersão dos pontos no final da curva, mas o número de servents presentes nesse intervalo é pouco significativo, conforme pode ser observado pela distribuição acumulada apresentada no gráfico (b) da Figura 6. A análise desses dados mostra que 98% dos servents não ultrapassam 2.700 segundos de disponibilidade e, nesse intervalo, fica explícito o comportamento linear da curva.

Como pode ser observado no gráfico (b) da Figura 6, a grande maioria dos servents apresenta muito baixa disponibilidade, 84% dos servents (66876) não ultrapassam 10 minutos de disponibilidade (0,7% do tempo total). Por outro lado, pode-se observar também um servent com disponibilidade acima de 96% (23 horas e 4 minutos). Esse valor é próximo aos aferidos nos servidores de redes cliente/servidor (acima de 99%).

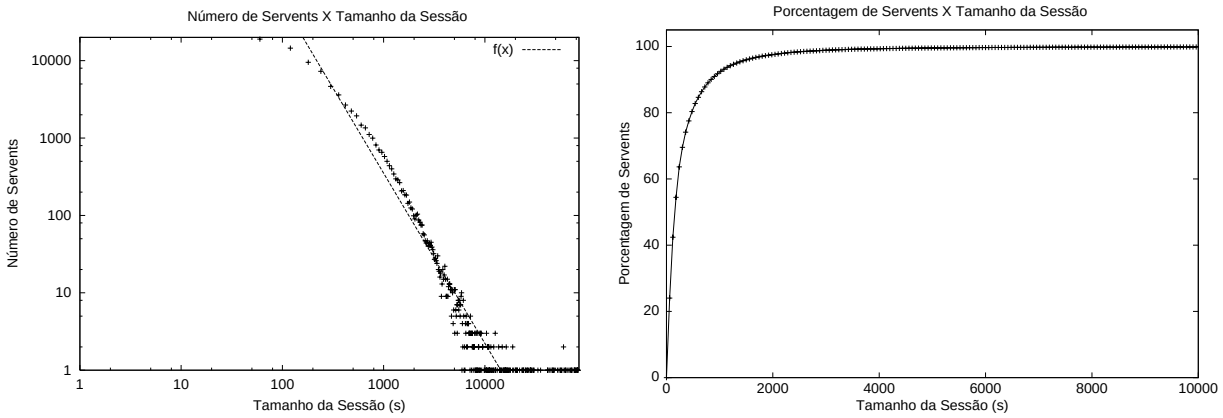


Figura 6: (a) Disponibilidade dos servents da gNet e (b) Distribuição percentual acumulada da disponibilidade dos servents

Esses números complementam estudos anteriores em que se observou que poucos servents gNet são realmente provedores de arquivos [AH00], mostrando que os servents também são predominantemente clientes no que concerne à disponibilidade para compartilhar tais recursos.

3.3.2.2 Volume de requisições recebidas Tratando o perfil servidor dos servents da rede Gnutella, temos 4 tipos de mensagens recebidas por eles: PING, QUERY, DOWNLOAD e PUSH.

Este item trata o primeiro grupo de experimentos, em que apresentamos análises dos PINGs e QUERYs recebidos. Em função de limitações nos experimentos realizados[4], não podemos analisar requisições de DOWNLOAD e PUSH.

Assim, o objetivo deste experimento é quantizar os números de QUERYs e PINGs recebidos por um servent ao longo do tempo, assim como a relação entre esses números, de forma a caracterizar o tráfego de mensagens recebidas pelo lado servidor de um servent. Para tanto, foram registrados todos os PINGs e QUERYs recebidos pelo nosso servent durante 24 horas.

Foram recebidos durante o experimento um total de 72.065 PINGs e 2.992.391 QUERYs, com médias de 0,8 e 34,5 requisições por segundo, respectivamente. As taxas máximas observadas foram de 2,9 PINGs e 100,7 QUERYs por segundo.

Observa-se a ocorrência de rajadas tanto de PINGs como de QUERYSs ao longo do período do experimento. Portanto, o tráfego dessas mensagens não é constante, apresentando ainda alterações bruscas ao longo do experimento.

Outra característica do tráfego é que o número de QUERYs é muito maior do que o de PINGs.

O valor médio da razão entre o número de QUERYs e PINGs recebidos ao longo do período do experimento é 49,9, com desvio padrão de 39%. O valor máximo observado foi de 159,9 e o mínimo de 16,5.

3.3.2.3 Volume de requisições atendidas Complementar ao grupo de experimentos anterior, os dois experimentos a seguir analisam requisições atendidas pelos servents.

[4]Não podemos disponibilizar arquivos em função do tráfego que isso acarretaria ao link da instituição de ensino a que estamos ligados.

O objetivo do primeiro experimento é observar e quantizar a capacidade que os servents apresentam de atender às consultas feitas na gNet. Para tanto, enviamos QUERYs para todos os servents da nossa base de endereços (120.535) durante 24 horas. As consultas foram preparadas a partir da distribuição de temas observada nas consultas recebidas pelo nosso servent (vide seção 3.2.1). Os QUERYHITs recebidos (741.860) foram armazenados, sendo contabilizados os números de respostas enviadas por cada servent. De posse desses dados, construímos o histograma do número de consultas respondidas pelos servents.

A análise desses dados mostra uma grande concentração de servents respondendo a poucas consultas durante o experimento: 87% não responderam mais do que 1 consulta e 94% não responderam mais do que 15. Por outro lado, há 1 servent que enviou 24.946 respostas (o segundo maior enviou 13.636 respostas). Resultado semelhante já foi observado em [AH00].

O objetivo do segundo experimento é verificar a existência de relação entre o número de consultas respondidas por um servent da gNet e a quantidade de arquivos compartilhados pelo mesmo. Para tanto, utilizamos a base de dados do experimento anterior, para obter o número de consultas respondidas por cada servent, combinada aos dados coletados para análise do número de arquivos compartilhados por um servent (vide seção 3.3.1).

Neste experimento foram contactados cerca de 90.282 servents dos quais 54% não respondiam a nenhuma consulta. Tem-se ainda que cerca de 35% dos servents não compartilham absolutamente nada e que 28% não respondem nem compartilham nada, assim temos então que 61% do total dos servents não é útil para nossa análise. Foi verificado também que 7% dos servents tinham dados se contradizendo, pois não compartilhavam nenhum arquivo e mesmo assim respondiam a consultas e que 26% possuiam arquivos compartilhados porém não respondiam a nenhuma consulta. Assim os dados utilizados para construção do histograma foram os 39% dos servents os quais possuíam arquivos compartilhados e respondiam a consultas. Esta característica dos servents da gNet foi estudada em detalhes em [AH00].

Utilizamos neste experimento apenas os dados dos 39% restantes dos servents que compartilhavam arquivos e respondiam a consultas. De posse desses dados, agrupamos os servents em função do número de arquivos compartilhados. Assim, para cada número de arquivos compartilhados, calculamos o número médio de consultas respondidas pelos servents que pertencem a esse grupo. Os gráficos da Figura 7 mostram os resultados obtidos.

Visualmente, percebe-se no gráfico (a) da Figura 7 a existência de uma relação entre as duas grandezas, mas aparentemente ela não é linear. Para verificar qual a relação, eliminamos inicialmente os pontos superiores a 1000 arquivos compartilhados, que correspondem a menos de 5% do espaço amostral, por apresentarem excesso de ruídos. Fizemos então o ajuste da curva observada a uma função do tipo $f(x) = ax^b$, assumindo que há uma relação de proporcionalidade, mas não necessariamente linear.

Os valores para a e b encontrados foram de 0,29 e 0,65, respectivamente, e a curva obtida foi mostrada no gráfico (b) da Figura 7. Assim, existe uma relação de proporcionalidade, mas o número de consultas respondidas cresce com menor velocidade do que o número de arquivos compartilhados pelos servents da gNet.

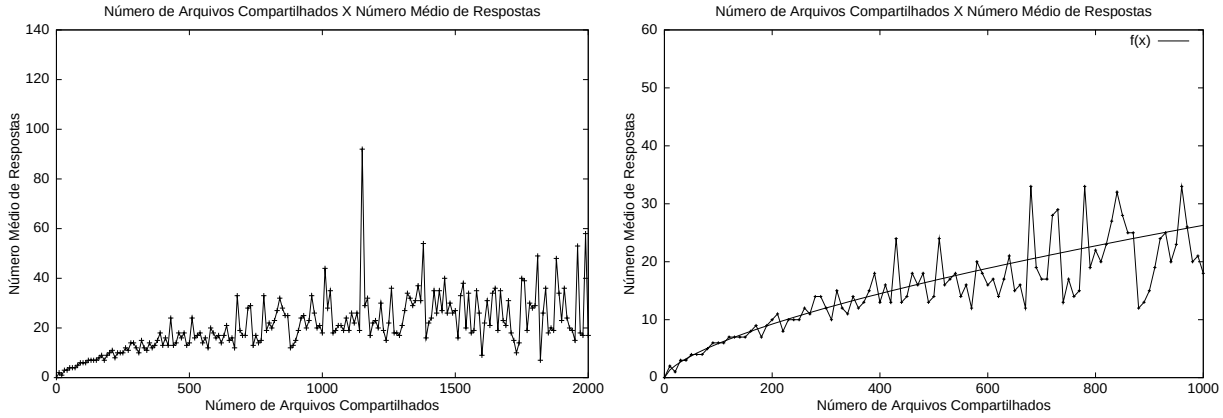


Figura 7: Número de consultas respondidas em função do número de arquivos compartilhados

4 Conclusões e Trabalhos Futuros

Neste artigo, apresentamos uma metodologia de caracterização de carga de redes P2P. A metodologia foi aplicada à gNet através da inserção de um agente coletor de dados na rede. Este agente foi programado para coletar os dados necessários para os experimentos de caracterização. Com os experimentos realizados, foi possível chegar a algumas conclusões relevantes em relação ao comportamento da gNet.

Na caracterização do servent como cliente, pudemos observar a Lei de Zipf no *ranking* de freqüência de palavras pesquisadas. Observamos também uma distribuição Log Normal na latência de comunicação entre os servents.

Na caracterização do servent como servidor pudemos observar uma relação linear entre o número de arquivos compartilhados e o tamanho dos mesmos, encontrando um tamanho médio por arquivo de 4Mbytes. Pudemos verificar a Lei de Zipf na distribuição das palavras presentes nos nomes de arquivos compartilhados na gNet e uma *Power Law* na disponibilidade dos servents na gNet. Observamos um tráfego intenso de QUERYs, cerca de 50 vezes maior que o de PINGs. Ainda na caracterização como servidor identificamos uma proporcionalidade não-linear entre o número de consultas respondidas e o número de arquivos compartilhados, em que a taxa de crescimento do número de consultas respondidas é inferior à de arquivos compartilhados pelos servents.

Até o momento, este trabalho de caracterização nos permitiu identificar distribuições estatísticas em uma série de características da gNet. Continuando nosso trabalho, estamos aprofundando o estudo de algumas características da gNet, incluindo, por exemplo, análises de reputação dos servents, e investigando possíveis alterações que poderiam melhorar a rede, de acordo com as conclusões das análises realizadas. Pretendemos ainda aplicar a metodologia apresentada neste artigo a outras redes P2P. No caso de outras redes de compartilhamento de arquivos, pretendemos realizar uma comparação com os resultados obtidos na gNet. Podemos ainda utilizar esta metodologia para realizar estudos de redes P2P que compartilham outros tipos de recursos, como processador e memória.

Referências

[Ada00] Lada A. Adamic. Zipf, power-laws, and pareto - a ranking tutorial. Technical report, Xerox Palo Alto Research Center, 2000.

[AH00] Eytan Adar and Bernardo A. Huberman. Free riding on gnutella. *First Monday*, 5(10), October 2000.

[CGM00] Brian Cooper and Hector Garcia-Molina. Peer to peer data trading to preserve information. Technical report, Stanford Database Group, November 2000.

[Cli00] Clip2.com. Gnutella: To the bandwidth barrier and beyond. Technical report, Clip2.com, November 2000.

[CSWH00] Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong. Freenet: A distributed anonymous information storage and retrieval system. In *Proceedings of the ICSI Workshop on Design Issues in Anonymity and Unobservability, Berkeley, CA*. International Computer Science Institute, July 2000.

[HC99] Chi-Chung Hui and Samuel T. Chanson. Improved strategies for dynamic load balancing. *IEEE Concurrency*, 7(3), July-Setember 1999.

[Kup00] Jerome Kuptz. Independence array - gnutella: Unstoppable by design. Technical report, Wired News - http://www.wirednews.com, October 2000.

[Mat] Mathworld. http://www.mathworld.wolfram.com.

[Ora00] Andy Oram. Gnutella and freenet represent true technological innovation. Technical report, The O'Reilly Network - http://www.oreillynet.com, December 2000.

[Rit00] Jordan Ritter. Why gnutella can't scale. no really. Technical report, 2000.

[Shi00] Clay Shirky. What is p2p... and what isn't? Technical report, The O'Reilly Network - http://www.oreillynet.com, November 2000.

[Sri01] Kunwadee Sripanidkulchai. The popularity of gnutella queries and its implications on scalability. Technical report, Carnegie Mellon University, February 2001.

[Sul00] Danny Sullivan. More than just music search. Technical report, The Search Engine Report - http://searchenginewatch.com/sereport, June 2000.

[Var01] Peter Varhol. Collaboration in the peer network environment. Technical report, Lotus Developer Network - http://www.lotus.com/home.nsf/welcome/developernetwork, 2001.

[Zip49] G. K. Zipf. *Human Behavior and the Principle of Least Effort - An Introduction to Human Ecology*. Addison-Wesley, Cambridge, Massachussets, 1949.