Impacto de Padrões de Comportamento Dinâmico e Malicioso na Eficácia de Máquinas de Busca Par-a-Par

**Fabiano Atalla, Daniel Miranda, Jussara Almeida,
Marcos André Gonçalves, Virgílio Almeida**

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brasil

{fabiano,danielcm,jussara,mgoncalv,virgilio}@dcc.ufmg.br

Abstract. *In an attempt to increase the spectrum of searchable information while attenuating scalability issues, Peer-to-Peer (P2P) networks have been viewed as an alternative way to design new Web search engines. However, the effectiveness of P2P Web searching may be severely limited by characteristics commonly observed in real P2P systems such as peer churn and malicious behavioral patterns. This paper analyzes the impact of these two aspects on the effectiveness of P2P Web searching. Our findings reveal that they can strongly affect the effectiveness of P2P Web searching, suggesting that the design of future P2P Web search engines will highly depend on new, application-specific reputation and incentive mechanisms.*

Resumo. *Na tentativa de ampliar o espectro de busca e atenuar problemas de escalabilidade, redes Par-a-Par (P2P) têm sido apontadas como alternativa para novas gerações de máquinas de busca na Web. No entanto, a eficácia da busca na Web em ambientes P2P pode ser gravemente limitada por características observadas em sistemas P2P reais, tais como a entrada e saída dinâmica de pares no sistema bem como padrões de comportamento malicioso exibidos pelos pares. Nossos resultados mostram que cada um desses aspectos pode afetar consideravelmente a eficácia da busca na Web em P2P, sugerindo que o desenvolvimento das futuras máquinas de busca P2P dependerá amplamente de novos mecanismos de reputação e incentivo que considerem aspectos específicos desse tipo de aplicação.*

1. Introdução

Sistemas Par-a-Par (P2P) têm se mostrado particularmente atrativos pelo seu potencial de escalabilidade e autonomia dos usuários, além da crescente popularidade em diversos tipos de aplicação, como voz sobre IP e compartilhamento de arquivos [Yu et al. 2005, Pouwelse et al. 2005]. Em especial, sistemas P2P têm sido apontados como alternativa para aplicações de máquinas de busca na Web [Bender et al. 2006, Wu et al. 2005, Cuenca-Acuna et al. 2003, Lu and Callan 2003]. De fato, com o crescimento explosivo da Web, tanto em número de usuários quanto de documentos, as máquinas de busca tradicionais não apenas precisam recuperar e indexar uma enorme coleção de páginas mas também devem processar centenas de milhões de requisições de usuários (i.e., consultas) por dia [Ntoulas and Cho 2007]. Além disso, fatores como atualização

de páginas, monopólio de informação e interesses comerciais [Bender et al. 2006, Doulkeridis et al. 2006] sugerem a necessidade de se explorar alternativas de arquitetura para busca na Web eficazes e eficientes.

Uma máquina de busca P2P é composta de computadores independentes e heterogêneos (i.e., pares), cada um contendo parte dos documentos da rede. Cada par envia consultas conforme interesse dos usuários e responde a consultas (locais e remotas) com uma lista ordenada de documentos relevantes presentes em sua coleção local. A aplicação de uma arquitetura distribuída em P2P para a busca na Web pode não apenas amenizar restrições de escalabilidade [Wu et al. 2005] como também explorar a parte da Web que máquinas de busca centralizadas não indexam – a chamada *Web invisível*, com cerca de bilhões de documentos [Lewandowski and Mayr 2007].

O desenvolvimento de máquinas de busca P2P é recente. Grande parte dos esforços foca em obter uma eficácia (i.e., recuperação de documentos relevantes) próxima da obtida em máquinas de busca centralizadas. Além disso, em geral assume-se um cenário ideal em que os pares nunca saem da rede ou agem maliciosamente ao responder uma consulta. No entanto, estudos de aplicações populares P2P de *compartilhamento de arquivos* analisaram a dinâmica da participação dos pares, ou *churn*, concluindo que esses geralmente são muito dinâmicos, entrando e saindo várias vezes da rede [Stutzbach and Rejaie 2006]. Nesse cenário instável, um usuário pode não obter os documentos mais relevantes para sua consulta se esses pertencerem a pares ausentes no momento em que for submetida. Logo, *churn* está diretamente relacionado à disponibilidade de conteúdo e pode afetar significativamente a eficácia da busca na Web em P2P e, em última instância, a qualidade do serviço provido por máquinas de busca P2P.

Mesmo utilizando a mesma rede P2P, acreditamos que o impacto de *churn* na eficácia de máquinas de busca P2P difere daquele visto em aplicações P2P de compartilhamento de arquivos. Geralmente, em aplicações de compartilhamento de arquivos, o usuário tem interesse em um objeto em particular; se qualquer par tiver o objeto, a busca será efetiva. Além disso, o objeto pode estar replicado em vários pares. Por outro lado, em uma máquina de busca P2P, o usuário busca informação sobre um tópico ou assunto. Nesse caso, diferentes documentos de diferentes pares podem ser relevantes para o usuário em diferentes níveis. A ausência, no momento em que a consulta é realizada, de um par com muitos documentos relevantes ou muitos pares com poucos documentos relevantes pode ter um grande impacto na qualidade da resposta. Mesmo em casos em que o usuário procura por um único documento na máquina de busca P2P, possivelmente esse documento estará presente em apenas um par, o que pode levar à falha da busca se o par estiver indisponível. Tais aspectos sugerem que o conjunto de documentos resultantes de uma busca, assim como o conjunto dos pares selecionados para processá-la, são diferentes para cada uma dessas aplicações de busca, isto é, a eficácia da busca pode ser estritamente dependente da aplicação. Vale destacar que esses aspectos são válidos mesmo se os pares tiverem exatamente a mesma dinâmica de entrada e saída da rede.

Trabalhos recentes [Gyongyi and Garcia-Molina 2005, Fetterly et al. 2004] destacam que as máquinas de busca tradicionais têm sido desafiadas a se proteger contra a ação maliciosa de alguns usuários – em particular, de *Web spammers*, que visam melhorar a avaliação de certas páginas Web pela máquina de busca. Em máquinas de busca P2P descentralizadas, o impacto de *Web spamming* na qualidade das respostas pode ser

ainda maior: em um cenário em que a informação está distribuída em pares autônomos, usuários geralmente não têm acesso direto às coleções de documentos dos pares. Assim, ao invés de tentar melhorar a avaliação indiretamente, por exemplo alterando o conteúdo do documento ou informações de *link*, um par *Web spammer* pode alterar diretamente a posição do documento em sua resposta a uma consulta, de forma a forçar a promoção de certas páginas Web.

Portanto, *churn* e comportamento malicioso geram um ambiente inseguro e instável sobre o qual as máquinas de busca P2P deverão residir. O impacto desses aspectos na eficácia da busca ainda não foi de todo analisado. Ressaltamos ainda que a robustez da aplicação em relação a esses fatores refletirá a qualidade do serviço na visão do usuário e, em última instância, a confiabilidade em relação à aplicação. Além disso, a ação maliciosa e oportunista por parte de alguns usuários pode acarretar um aumento no consumo de recursos não só de processamento mas também de rede, devido ao tráfego de documentos que não refletem o interesse do usuário.

Nessa direção, apresentamos uma análise mais realista da busca na Web em P2P, quantificando, via simulação, o impacto desses aspectos em sua eficácia. O ambiente de simulação é composto por milhares de pares independentemente entrando e saindo da rede. Além disso, avaliamos o impacto da inserção na rede P2P de pares *Web spammers*, que respondem às consultas com conteúdo não solicitado. Em nossa avaliação usamos cargas reais compostas por documentos e consultas de duas coleções, WBR [Calado 1999] e TREC-8 [Voorhees and Harman 1999], também referidas como c=$\{wbr, trec\}$ ao longo do texto.

Nossos resultados mostram que a eficácia de busca na Web em P2P sofre significativamente com a instabilidade e insegurança dos pares, mesmo em uma rede completamente colaborativa (i.e., muita informação difundida na rede acerca dos documentos existentes e sua relevância). Quando pares estão disponíveis, em média, 75% do tempo de simulação, a degradação na Precisão Média (AP), i.e., a fração de documentos relevantes presente na resposta a uma consulta, comparada à solução centralizada, é superior a 24% para 20% das consultas. Além disso, mesmo agregando os resultados da AP sobre todas as consultas (MAP), a degradação chega a 22%. Analogamente, na maior coleção, se apenas 0,05% dos pares agem maliciosamente, a degradação pode ser superior a 26% para 20% das consultas. Também mostramos que em ambientes menos colaborativos, o impacto pode ser ainda mais intenso: para 20% das consultas, a degradação é de pelo menos 73% e 75%, para os mesmos níveis de instabilidade e insegurança, respectivamente. Assim, confirmamos que tanto *churn* quanto comportamento malicioso representam sérios desafios no desenvolvimento de máquinas de busca em P2P.

O restante do artigo está estruturado da seguinte forma: a Seção 2 discute trabalhos anteriores. A Seção 3 descreve nosso modelo de simulação, assim como as cargas e métricas utilizadas em nossa avaliação. Discutimos nossos principais resultados na Seção 4. A Seção 5 apresenta conclusões e trabalhos futuros.

2. Trabalhos Relacionados

[Rhea et al. 2004] mostram que o *churn*, ou entrada e saída dinâmica de pares, pode afetar consideravelmente o projeto e a avaliação de sistemas P2P. Por exemplo, os autores mostram que as atuais implementações de DHT não lidam bem com altas ta-

xas de *churn*. Por sua vez, [Stutzbach and Rejaie 2006] estudaram *churn* em diferentes aplicações P2P e verificaram que grande parte das sessões dos usuários (tempo online) duravam apenas alguns minutos, enquanto outras duravam dias ou mesmo semanas. Além disso, diversos trabalhos procuram entender e caracterizar padrões de comportamento malicioso em redes P2P, como conluio, ataques Sybil e poluição de conteúdo [Marti and Garcia-Molina 2006, Douceur 2002, Liang et al. 2005]. Por exemplo, [Costa et al. 2006] mostram que a poluição de conteúdo pode reduzir substancialmente a confiabilidade em aplicações P2P de *compartilhamento de arquivos*.

Diversos trabalhos propõem modelos de aplicação P2P para *máquinas de busca na Web* considerando diferentes tipos de arquitetura. Um exemplo de modelo de máquina de busca em redes P2P estruturadas é o Minerva [Bender et al. 2006], em que pares publicam estatísticas sobre sua coleção local em um diretório global, porém fisicamente distribuído (DHT). Exemplos de modelo de máquina de busca em redes P2P não estruturadas utilizam redes semânticas [Doulkeridis et al. 2006], em que pares com interesses similares são agrupados, e sistemas como 6Search [Wu et al. 2005], em que pares executam um algoritmo de roteamento auto-adaptativo que reconfigura a topologia da rede de forma que consultas possam ser satisfeitas pelos melhores pares (i.e., os que possuam os documentos mais relevantes à consulta).

No entanto, *churn* e padrões de comportamento malicioso ainda foram pouco explorados em aplicações P2P para máquinas de busca na Web. A maioria das máquinas de busca P2P assume um ambiente confiável e colaborativo, não cobrindo devidamente aspectos de segurança e autonomia dos pares. Em especial, a busca por conteúdo sugere que os efeitos de *churn* podem ser bem particulares nesse tipo de aplicação. De fato, [Stutzbach and Rejaie 2006] mostram que o *churn* é largamente dependente do tipo de aplicação P2P. Além disso, ambientes P2P podem abrir espaço para padrões de comportamento malicioso adotados em máquinas de busca centralizadas. Por exemplo, com o *Web spamming*, usuários maliciosos (e oportunistas) tentam melhorar a avaliação de documentos que teriam pouca ou nenhuma relevância para as consultas dos usuários. [Gyongyi and Garcia-Molina 2005] mostram algumas técnicas para realizar *Web spamming* enquanto [Fetterly et al. 2004] propõem um mecanismo para identificar spam através de análise estatística.

3. Modelo de Simulação

Avaliamos o impacto de *churn* e comportamento malicioso em máquinas de busca P2P via simulação. De forma a capturar as características essenciais de máquinas de busca P2P, nosso modelo possui duas camadas: a rede P2P e a máquina de busca executada sobre ela. Note que o modelo foca na *eficácia*, e não no desempenho, de máquinas de busca P2P. As próximas seções descrevem as principais considerações e componentes do modelo. A Tabela 1 sumariza os parâmetros de simulação.

Tabela 1. Parâmetros do Modelo de Simulação

Parâmetro	Descrição
n_c	# de pares na coleção de teste c (c=wbr, $trec$)
τ_c	Limiar de similaridade para c (c=wbr, $trec$)
t_q	Tempo médio entre chegadas das consultas
λ_{on}, ρ_{on}	Parâmetros da distribuição Weibull para os tempos online dos pares
λ_{off}, ρ_{off}	Parâmetros da distribuição Weibull para os tempos offline dos pares
M_c	% de pares maliciosos para c (c=wbr, $trec$)

3.1. Modelo da Rede P2P

Como focamos no impacto de *churn* e comportamento malicioso, nosso modelo de rede P2P não modela uma arquitetura específica (e.g., P2P estruturada ou não). Dessa forma, assumimos que *(1)* pares online são sempre visíveis entre si, *(2)* conteúdo online é sempre localizado e *(3)* roteamento de consultas e respostas nunca falha. Além disso, como a qualidade das respostas do sistema, e não o desempenho, é a métrica de interesse, desconsideramos os tempos de consulta e de resposta.

3.1.1. *Churn*

Churn em máquinas de busca P2P ainda não foi caracterizado na literatura, embora vários trabalhos tenham investigado *churn* em aplicações P2P de *compartilhamento de arquivos*. Em particular, uma caracterização recente [Stutzbach and Rejaie 2006] de três sistemas P2P populares mostra que a duração da sessão (i.e., tempo online) dos pares é bem modelada por distribuições Weibull[1] com parâmetros de forma $\rho \approx 0.44$ e de escala $\lambda \approx 35.20$ unidades de tempo, em média. Embora relevante, esse estudo não apresenta dados similares para os tempos offline, i.e., o tempo entre duas sessões consecutivas de um mesmo par, outro aspecto de *churn* de suma importância para a eficácia da busca em P2P.

Devido à falta de caracterização de *churn* em aplicações de busca na Web em P2P assim como sua caracterização completa em outros tipos de aplicação em P2P, optamos por modelar os tempos online e offline dos pares por distribuições Weibull com parâmetros de forma ρ_{on} e ρ_{off}, e parâmetros de escala λ_{on} e λ_{off}, respectivamente.

3.1.2. Comportamento Malicioso

Consideramos cenários em que pares maliciosos, em particular *Web spammers*, utilizam máquinas de busca P2P para publicar conteúdo não solicitado. Uma possível estratégia seria atribuir a um documento com conteúdo não solicitado um alto valor de similaridade com a consulta, independentemente de sua relevância. Como conseqüência, a máquina de busca do usuário que realizar a consulta, ao agregar as respostas de cada par, atribuirá ao documento uma relevância maior do que ele realmente deveria ter [Gyongyi and Garcia-Molina 2005].

Em nosso modelo, adicionamos à rede uma fração M_c de pares *Web spammers* cujos documentos não fazem parte da coleção original. Esses pares fazem conluio [Marti and Garcia-Molina 2006], isto é, colaboram entre si de forma a realizar um ataque mais efetivo, respondendo a cada consulta com um único documento *spam*. Assumimos que todo *Web spammer* evita divulgar que seus documentos possuem similaridades muito altas com a consulta; do contrário, seus documentos poderiam ocupar sempre as primeiras posições na lista de resposta obtida pelo usuário, tornando-o alvo fácil de mecanismos de filtragem e detecção. Consideramos então um cenário realista em que, para evitar que seja detectado e ainda garantir que seu conteúdo seja publicado, todo *Web spammer* busca antes "aprender" o comportamento da rede: ao invés de responder diretamente a uma consulta com conteúdo não solicitado, encaminha a consulta à rede e observa as similaridades dos documentos obtidos [2]. Dessa forma, quando (se) a mesma consulta

[1] $f(x) = \frac{\rho x^{\rho-1}}{\lambda^{\rho}} e^{-(x/\lambda)^{\rho}}$

[2] Um *Web spammer* submete a mesma consulta um número razoável de vezes de tempos em tempos com o intuito de recuperar todos os documentos relevantes presentes na rede.

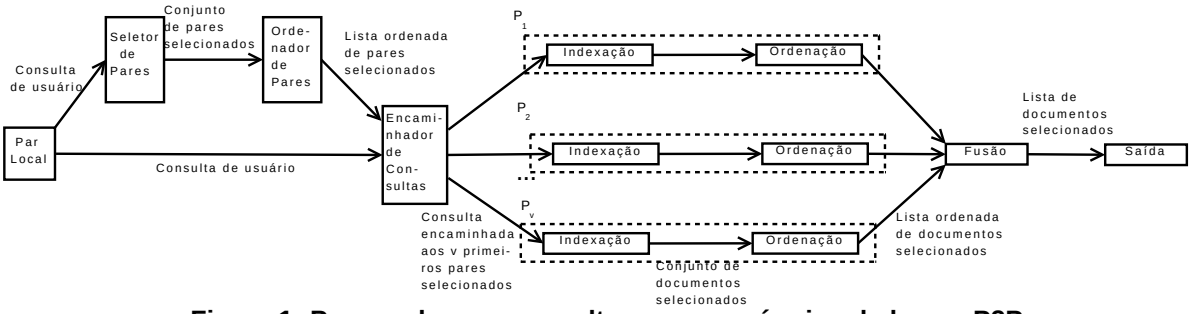


Figura 1. Passos de uma consulta em uma máquina de busca P2P

for recebida novamente, o *Web spammer* divulgará que seu documento possui uma similaridade uniformemente distribuída entre τ_c (ver 3.2.2) e a maior similaridade observada. Em nossos experimentos, iniciamos a simulação após o período de aprendizagem. Finalmente, no modelo, os *Web spammers* sempre divulgam que possuem documentos que satisfazem a qualquer consulta.

3.2. Modelo de Busca P2P

Nosso modelo considera que cada par tem sua própria coleção indexada de documentos. Quando um usuário conecta-se à rede P2P por algum par, pode realizar consultas em outros pares, pesquisando remotamente por documentos com conteúdo relevante à consulta submetida. Em troca, esse par deve processar consultas de outros pares, retornando informação sobre documentos locais. Além disso, o conteúdo desses documentos também estará disponível para acesso.

3.2.1. Passos de Busca

A Figura 1 ilustra o modelo de busca adotado. Após o par receber uma consulta submetida pelo usuário, os pares mais promissores para responder à consulta são geralmente obtidos pelo *Seletor de Pares*, que obtém os pares que satisfazem os critérios de busca, e pelo *Ordenador de pares*, que ordena os pares que satisfazem os critérios de busca por sua similaridade com a consulta. Em seguida, a consulta é *encaminhada* para os pares selecionados (local e/ou remotos). Cada par então busca os documentos mais relevantes para a consulta em sua coleção local e responde com uma lista ordenada de documentos (ver 3.2.2). Finalmente, os resultados de cada par são fundidos em uma única lista no par requisitante e apresentada ao seu usuário como um conjunto de *links* ordenados por relevância.

O Seletor e o Ordenador de Pares podem aprender o conteúdo de cada par submetendo periodicamente consultas sintéticas à rede P2P e observando os documentos retornados. Essa técnica é conhecida como *amostragem baseada em consultas* [Callan and Connell 2001]. No momento em que o usuário submeter uma consulta, assumimos que os pares mais promissores já tenham sido obtidos via amostragem. Consideramos também que um par é promissor se tiver pelo menos um termo da consulta.

3.2.2. Processador de consultas

Cada par possui seu próprio processador de consultas baseado no *modelo de espaço vetorial* [Baeza-Yates and Ribeiro-Neto 1999]. Nesse modelo, a consulta q e o documento j são representados por vetores w-dimensionais, onde w é o total de termos da coleção e cada componente dos vetores corresponde ao peso de um termo no documento (ou consulta) que o contém. Os pesos usados nesse modelo são baseados em estatísticas TF-IDF.

TF-IDF é uma abordagem tradicional usada para medir a importância de termos em documentos. TF é o componente do peso que representa a frequência do termo no documento e IDF está relacionado ao número de documentos na coleção contendo o termo.

O componente TF de um termo i em um documento j geralmente é dado pela fórmula $tf\,(i,j) = \frac{freq(i,j)}{max(freq(l,j))}$, onde $freq\,(i,j)$ é a frequência (i.e., número de vezes) em que i ocorre em j e $max\,(freq\,(l,j))$ é a maior frequência dentre qualquer termo l de j. O IDF de um termo i é dado por $idf\,(i) = log\,(D/d_i)$, onde D é o número total de documentos da coleção e d_i é o número de documentos da coleção com o termo i. Portanto, o peso de um termo i em um documento j é dado por $W_{ij} = tf\,(i,j) \times idf\,(i)$.

O modelo vetorial avalia a *similaridade* entre a consulta do usuário, q, e o documento j a partir do cosseno do ângulo entre os vetores. Assim, quanto maior a similaridade, melhor será a posição do documento na resposta. Em nosso modelo, assumimos que documentos com similaridade abaixo de um limiar τ_c não são retornados para o usuário.

3.2.3. Níveis de Colaboração

Quando um par possui muito conhecimento sobre a relevância de seus documentos na rede P2P (um ambiente altamente colaborativo), espera-se que a ordem de seus documentos na lista de resposta aproxime-se daquela obtida em um ambiente centralizado. No entanto, em uma rede P2P real não estruturada, os pares tendem a se comportar de forma autônoma e independente, podendo ser custoso para um par manter-se atualizado sobre a relevância de seus documentos na rede P2P. Assim, de forma a capturar o impacto da instabilidade e insegurança em redes P2P com diferentes níveis de colaboração, limitamos nossa análise em níveis extremos de conhecimento dos pares sobre os documentos da rede.

De um lado, nosso nível de total conhecimento, ou *limite superior*, assume que os pares divulgam estatísticas sobre sua coleção para todo par na rede, de forma que cada par mantém-se atualizado sobre estatísticas de documentos na rede sem depender de serviços centralizados. O algoritmo de *Gossiping* [Demers et al. 1987] é uma solução proposta na literatura para construir e manter tal ambiente. Nesse nível, processamento de consultas e fusão dos resultados, assim como seleção e classificação de pares, são executados como se cada par tivesse pleno acesso às estatísticas sobre a coleção de documentos da rede.

De outro lado, no nível de pouco conhecimento, ou *limite inferior*, além de utilizar seu próprio processador de consultas, cada par processa consultas utilizando apenas conhecimento local. Em outras palavras, nesse ambiente P2P não se espera que os pares colaborem entre si trocando estatísticas sobre documentos. Nesse nível, selecionamos e classificamos os pares da mesma maneira que no *limite superior*, mas restringimos os processadores de consulta a usar apenas conhecimento sobre sua coleção local e, na fusão dos resultados, apenas as similaridades dos documentos informados pelos pares pesquisados.

3.3. Coleções de Teste

Utilizamos duas coleções de teste: WBR e TREC-8 *ad-hoc* (ou TREC, no decorrer do texto). A WBR [Calado 1999] foi desenvolvida a partir de um conjunto de documentos coletados na Web brasileira pelo TodoBR. A TREC [Voorhees and Harman 1999] é formada por documentos de alguns jornais e conferências do governo dos EUA, distribuídos para desenvolvimento e testes de sistemas de recuperação de informação.

A Tabela 2 sumariza as coleções avaliadas. Cada coleção possui, além de do-

Tabela 2. Coleções

Coleção (c)	# Documentos (D_c)	# Consultas	# max. documentos relevantes por consulta
WBR	5.751.296	50	96
TREC-8	528.155	50	347

cumentos, um conjunto de consultas de usuário. A WBR possui as 50 consultas mais populares realizadas no TodoBR. A TREC possui 50 consultas manualmente elaboradas por especialistas. Além disso, os documentos possuem julgamento de relevância por consulta. O julgamento de relevância em ambas coleções foi feito pelo método de *pooling*. Nesse método, são utilizados diversos algoritmos de recuperação de informação. Cada algoritmo classifica um número de documentos como relevantes, que são julgados manualmente por especialistas. Documentos que não foram recuperados por nenhum algoritmo não são julgados e são considerados não relevantes. Finalmente, assumimos que os pares submetem consultas de forma independente. Optamos então por modelar o tempo entre consultas com uma distribuição exponencial com média t_q. Em nosso modelo, consultas são realizadas por pares escolhidos de forma aleatória.

3.4. Métricas de Qualidade de Busca na Web

Utilizamos a Precisão Média (AP, *Average Precision*) para quantificar a qualidade da busca. Dado o conjunto R dos documentos relevantes da coleção por consulta, a lista dos k documentos mais similares à consulta de acordo com a máquina de busca, o julgamento binário de relevância x_i para o i-ésimo documento da lista de documentos e a precisão $p_m = \frac{1}{m}\sum_{j=1}^{m} x_j$ do m-ésimo documento da lista, define-se formalmente [Kishida 2005]

$$AP = \frac{1}{|R|}\sum_{i=1}^{k} x_i p_i \tag{1}$$

Isto é, AP fornece a fração de documentos relevantes, ponderados pela posição na lista de documentos recuperados. AP é largamente utilizada para medir a eficácia de algoritmos de recuperação de informação, incorporando fatores como precisão e revocação sem desprezar a ordem dos documentos na resposta. Os valores de AP sobre cada consulta são agregados em outra métrica, a média aritmética da AP, ou MAP [Kishida 2005].

Essa seção descreveu o modelo adotado para avaliação do impacto de *churn* e comportamento malicioso em máquinas de busca P2P. O modelo foi inspirado em caracterizações de sistemas P2P reais da literatura [Stutzbach and Rejaie 2006] e implementado a partir de um processador de consultas real e de coleções de teste reconhecidas. Logo, acreditamos que o modelo captura os principais aspectos relacionados à eficácia de máquinas de busca P2P. O simulador proposto foi validado em um cenário com um único par sempre disponível e que nunca age maliciosamente (i.e., M=0, λ_{off}=0, λ_{on}=∞), o que corresponde a um ambiente centralizado seguro. Conforme esperado, os resultados são equivalentes aos encontrados pelo processador de consultas adotado e qualitativamente semelhantes aos de trabalhos anteriores que utilizaram o mesmo processador de consultas e as mesmas coleções de teste [Almeida et al. 2007].

4. Resultados da Simulação

Esta seção mostra os resultados quantitativos mais relevantes do impacto de *churn* e comportamento malicioso na eficácia de busca na Web em P2P. Implementamos o modelo de

simulação em C++ por questões de desempenho e escalabilidade. Utilizamos a biblioteca de simulação SimPack [Fishwick 1992] para controle de eventos e um processador de consultas baseado em modelo vetorial adotado na literatura [Almeida et al. 2007].

Consideramos cenários com diversos níveis de *churn* e tamanhos da população de pares maliciosos, analisando o impacto para ambas coleções de teste, WBR e TREC. Para cada configuração analisada, os resultados são apresentados em uma distribuição acumulada dos valores de AP de consultas individuais e por valores agregados (MAP). Além disso, cada AP resulta de uma média de 50 simulações e os valores de AP são agregados com um desvio padrão de no máximo 14% (8%) da média para a TREC (WBR).

Tomamos como configuração de linha de base (LB) um cenário com um único par, que está sempre disponível e nunca age maliciosamente; esse cenário é equivalente a uma máquina de busca centralizada. Destacamos que no *limite superior* (ver 3.2.3) nosso modelo P2P também apresenta resultados equivalentes aos de uma máquina de busca centralizada. Isso acontece porque os valores de similaridade entre a consulta e os documentos que um par possui são calculados usando os mesmos algoritmos e as mesmas estatísticas (globais) da máquina de busca centralizada. Por outro lado, as similaridades obtidas no *limite inferior* podem ser bem diferentes em redes P2P com alta heterogeneidade, pois as similaridades dependem essencialmente de como os documentos da coleção estão distribuídos entre os pares (ver discussão a seguir). Nossos resultados mostram que redes com diferentes níveis de colaboração entre pares podem sofrer de forma diferente com a instabilidade dos pares e insegurança da rede.

Devido a diferenças significativas entre as coleções (ver Tabela 2), atribuímos a cada coleção c uma distribuição de documentos diferente. Na WBR, cada documento está vinculado a uma URL, o que nos permite fazer uma distribuição direta entre os pares na linha de base: cada uma das n_{wbr} máquinas hospedeiras (i.e., *hosts*) é associada aleatoriamente a um único par e as páginas Web da máquina hospedeira são associadas aos documentos do par correspondente. Por outro lado, como a TREC não possui característica Web, distribuímos uniformemente os documentos pelos pares, de forma que cada par contenha em torno de D_{trec}/n_{trec} documentos, uma amostra da coleção com um nível de confiança de 99% e acurácia em torno de 5%.

Para normalizar a lista final de resultados no cenário LB para o mesmo tamanho em ambas coleções, usamos um limiar de similaridade τ_c diferente para cada coleção. Em particular, τ_c produz uma lista de resultados com média em torno de 1.000 documentos, o tamanho máximo daquela produzida pelo Google [Gopalakrishnan et al. 2006].

A Tabela 3 mostra os parâmetros usados na configuração do cenário LB.

Tabela 3. Configuração do cenário de linha de base (LB)

n_{trec}	n_{wbr}	τ_{trec}	τ_{wbr}	t_q
880	110.912	0,10	0,32	100

4.1. Impacto de *Churn*

Nesta seção quantificamos o impacto de diferentes níveis de *churn* na eficácia da busca na Web em P2P. Focamos os cenários em que nenhum par age maliciosamente, i.e., $M_c = 0$.

Para avaliar o impacto de diferentes níveis de *churn*, fixamos a distribuição dos tempos online dos pares com uma média de μ_{on}=$91,84$ unidades de tempo, dada pelos parâmetros Weibull ρ_{on} e λ_{on} iguais a 0,44 e 35,20 respectivamente (ver 3.1.1).

Além disso, fixamos ρ_{off}=ρ_{on} enquanto variamos λ_{off}, produzindo diferentes tempos médios offline μ_{off}. Finalmente, de forma a capturar a estabilidade do sistema, define-se [Menasce and Almeida 2001] *disponibilidade do par*, A, como sendo o percentual do tempo (simulado) que o par participa da rede, em média. Isto é:

$$A = \frac{\mu_{on}}{\mu_{on} + \mu_{off}} \times 100 \quad (2)$$

A Tabela 4 sumariza os parâmetros de *churn* avaliados; A=100% é equivalente à configuração da linha de base (se M=0%) e no cenário A=0% AP é sempre zero.

Tabela 4. Disponibilidade dos pares para diferentes cenários de *churn* (ρ_{off}=ρ_{on}=0,44, λ_{on}=35,20)

A	λ_{off}	μ_{off}
0	∞	∞
25	105,60	275,52
50	35,20	91,84
75	11,73	30,61
100	0	0

Tabela 5. Impacto de *churn* sobre MAP para os limites superior e inferior de colaboração (MAP entre parênteses)

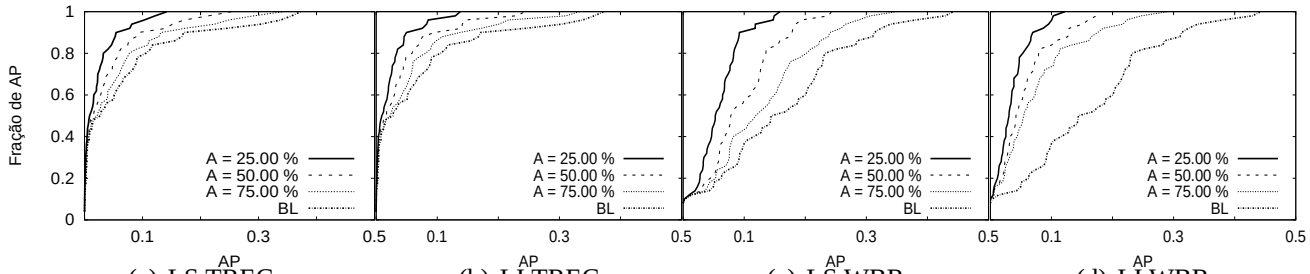
	TREC		WBR	
A %	**LS %**	**LI %**	**LS %**	**LI %**
LB	0 (0,069)		0 (0,166)	
75	22 (0,054)	23 (0,053)	22 (0,129)	54 (0,076)
50	43 (0,039)	45 (0,038)	43 (0,094)	66 (0,056)
25	67 (0,023)	67 (0,023)	65 (0,058)	78 (0,036)

Os resultados principais estão sumarizados na Tabela 5. A Tabela mostra o percentual de degradação para os valores de MAP, quando comparados à linha de base (LB), para todos os cenários analisados, em ambos limites superior (LS) e inferior (LI) de colaboração entre os pares. A Tabela ainda mostra os valores de LB e os valores de MAP entre parênteses. Mesmo quando os pares têm conhecimento global dos documentos da rede (LS) e a maioria dos pares está online em média (A=75%), o MAP reduz em torno de 22% em ambas coleções. No entanto, em ambientes menos colaborativos (LI) o impacto pode ser ainda mais intenso na coleção WBR, pois a degradação do cenário LB chega a 54% também para A=75%. Quando documentos estão distribuídos por máquina hospedeira, as coleções dos pares podem ser muito específicas (em torno de um mesmo tópico), o que faz com a qualidade local dos resultados tenda a ser bastante distinta da qualidade da coleção global. Por outro lado, os resultados mostram que redes em que documentos estão distribuídos uniformemente (TREC) são muito menos vulneráveis à falta de colaboração. Para a mesma fração de pares online, os limites superior e inferior são bastante similares quando comparados à WBR, sugerindo que a distribuição uniforme garante aos pares uma amostra representativa dos documentos da rede.

Embora o impacto de *churn* nos valores de MAP (agregado) seja significativo, também é importante avaliar a degradação das consultas individualmente. A Figura 2 mostra a distribuição acumulada da qualidade (i.e., métrica AP) de resultados individuais de consultas, para vários níveis de *churn* incluindo o cenário LB. De fato, as curvas mostram uma grande variabilidade nos valores individuais de AP.

As curvas ainda mostram que o impacto de *churn* aumenta com a métrica AP. Isso acontece porque altos valores de AP são consequência do grande número de documentos relevantes presentes na rede (em alguns pares) e informados na resposta a uma consulta. Quanto maior esse número, maior é a chance de que documentos presentes em pares indisponíveis no momento sejam relevantes para a consulta.

O impacto de *churn* pode ser ainda maior do que sugerido pelos resultados agregados (MAP). De fato, comparado ao cenário LB, para A=75% no limite superior, 20% das consultas sofrem uma degradação acima de 24% na métrica AP em ambas coleções.



(a) LS TREC (b) LI TREC (c) LS WBR (d) LI WBR

Figura 2. Distribuição acumulada da métrica AP em função do comportamento dos pares (*Churn*, $M = 0\%$)

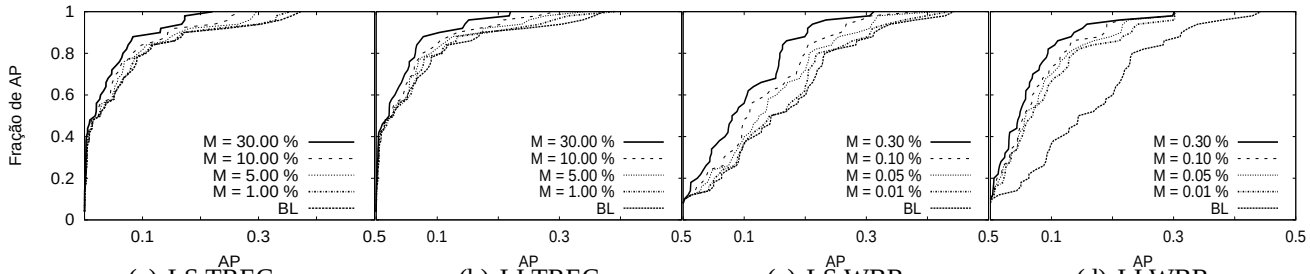
As consequências de cenários menos colaborativos são ainda mais severas na eficácia da busca na Web P2P: para 20% das consultas, o impacto é maior que 29% e 73%, para as coleções TREC e WBR, respectivamente (A=75%).

Em suma, *churn* pode ter grande impacto na eficácia da busca na Web em P2P, mesmo em ambientes razoavelmente estáveis (A=75%) e redes completamente colaborativas (LS). Os resultados também sugerem que os efeitos de *churn* podem ser ainda maiores em redes menos colaborativas (e mais reais), em que se espera que os pares processem consultas usando quase que apenas conhecimento local e indexem sua própria coleção. Além disso, diferentemente de sistemas de compartilhamento de arquivos, não se espera que usuários voluntariamente repliquem documentos quando estiverem utilizando uma máquina de busca P2P. Assim, uma vez que *churn* é um aspecto intrínseco a redes P2P e logo difícil de se evitar, acreditamos que a viabilidade de máquinas de busca P2P dependerá fortemente da adoção de sofisticadas estratégias de replicação. Por exemplo, como em sistemas de compartilhamento de arquivos P2P, poderia-se introduzir algum tipo de *incentivo* para os usuários baixarem e indexarem as páginas visitadas.

4.2. Impacto do Comportamento Malicioso

Analisamos o impacto do comportamento malicioso na eficácia da busca na Web em P2P para diferentes números de *Web spammers*, expressos por uma fração M_c de pares adicionados à rede. Devido à (grande) diferença no tamanho das coleções, para fins de comparação adicionamos proporcionalmente a cada coleção diferentes números de pares *Web spammers*. Como conseqüência, a fração de documentos *spam* presentes na lista de documentos retornados é aproximadamente a mesma para ambas coleções. Além disso, de forma a capturar apenas o impacto do comportamento malicioso na eficácia da busca, assumimos que os pares estão sempre disponíveis ao longo da simulação (A=100%).

De forma análoga à Seção 4.1, a Tabela 6 mostra, para cada nível de colaboração, a porcentagem do impacto de *Web spamming* quando comparado à linha de base (LB), assim como os valores absolutos de LB e MAP. Concluímos que o impacto do comportamento malicioso pode ser significativamente alto à medida que M_c aumenta, especialmente para a maior coleção (WBR). Mesmo em um cenário completamente colaborativo (LS) e bastante seguro (M_{trec}=5% e M_{wbr}=0,05%), as consultas apresentam uma degradação agregada (MAP) de 14% e 15% para cada coleção. Além disso, os resultados reforçam que máquinas de busca P2P em que documentos estão uniformemente distribuídos entre os pares (TREC) não parecem ser sensíveis a baixos níveis de colaboração entre pares. No entanto, se cada par indexar documentos a sua maneira, a



(a) LS TREC (b) LI TREC (c) LS WBR (d) LI WBR

Figura 3. Distribuição acumulada da métrica AP em função do comportamento dos pares (Malicioso, $A = 100\%$)

degradação na qualidade dos resultados pode ser muito maior. De fato, para a coleção WBR, a degradação no MAP chega a 49%.

Novamente, a análise das consultas individuais revela degradações ainda maiores na eficácia da busca. A Figura 3 mostra os valores de AP para os mesmos números de pares maliciosos. No cenário completamente colaborativo (LS), M_{trec}=5% e M_{wbr}=$0,05\%$ (em que a lista de documentos retornados apresenta um aumento de aproximadamente 5% com documentos *spam*), uma fração significativa de consultas (20%) sofre degradação maior que 25% e 26% para a TREC e WBR, respectivamente. Além disso, a degradação na AP no cenário menos colaborativo é superior a 75% para 20% das consultas.

Em suma, *Web spamming* é um sério desafio para máquinas de busca P2P devido principalmente a dois aspectos: primeiro, nossos resultados sugerem que seu impacto é significativo mesmo em redes P2P com grande colaboração entre pares (LS). Em redes com pouca colaboração (LI), mostramos que os ataques podem ser ainda mais severos na qualidade da busca. Segundo, a detecção de *Web spamming* pode ser muito mais difícil nas máquinas de busca P2P do que nas tradicionais, principalmente pelo fato de que os pares não possuem acesso direto às coleções de outros pares. Logo, a eficácia de máquinas de busca P2P dependerá fortemente da adoção de sofisticados mecanismos de reputação.

Tabela 6. Impacto de comportamento malicioso sobre MAP nos limites superior e inferior de colaboração (MAP entre parênteses)

	TREC			WBR	
$M_{trec}\%$	**LS %**	**LI %**	$M_{wbr}\%$	**LS %**	**LI %**
LB	0 (0,069)		LB	0 (0,166)	
1	3 (0,067)	6 (0,065)	0,01	4 (0,159)	44 (0,093)
5	14 (0,059)	17 (0,057)	0,05	15 (0,141)	49 (0,084)
10	25 (0,052)	26 (0,051)	0,10	23 (0,127)	54 (0,077)
30	42 (0,040)	43 (0,039)	0,30	40 (0,100)	61 (0,064)

5. Conclusões e Trabalhos Futuros

Com o crescente interesse na adoção de sistemas P2P como arcabouço para as futuras máquinas de busca na Web, entender os efeitos do dinamismo e do comportamento malicioso dos pares na eficácia da busca mostra-se essencial para avaliação da viabilidade desse novo tipo de aplicação. Nessa direção, analisamos o impacto de cada um desses fatores na qualidade das respostas obtidas na busca na Web em P2P com diferentes níveis de colaboração e autonomia dos pares. Nossos resultados indicam que ambos fatores revelam-se sérios desafios para o desenvolvimento de máquinas de busca P2P reais.

Desenvolvemos um estudo inicial acerca dos limites de máquinas de busca P2P diante do dinamismo e comportamento malicioso dos pares. Preferimos então não avaliar replicação ou outras estratégias que minimizem o impacto na eficácia da busca de forma a evitar ruídos nessa análise inicial. No entanto, experimentos preliminares já indicam benefícios da replicação; uma análise mais cuidadosa será realizada em trabalhos futuros.

Enfatizamos a necessidade de mecanismos de incentivo e reputação específicos para aplicações de busca na Web em P2P. Tais mecanismos devem considerar que, diferentemente de aplicações P2P de compartilhamento de arquivos, não se espera que usuários baixem e compartilhem páginas Web de outros pares, mas que apenas naveguem por elas. Além disso, diferentemente de máquinas de busca centralizadas, não se espera que usuários acessem livremente a coleção dos pares, mas apenas os documentos retornados à consulta. Também deve haver incentivo para que o usuário mantenha a aplicação em execução e para que as cópias locais estejam sempre acessíveis para a rede P2P.

Além da análise de novos mecanismos de incentivo e reputação, pretendemos investigar os efeitos de ataques Sybil e tipos mais sofisticados de conluio em máquinas de busca P2P. Em adição, não apenas a eficácia, mas também o desempenho é importante de ser avaliado, principalmente em redes de grande escala.

6. Agradecimentos

Esse trabalho foi parcialmente financiado pelo projetos GERINDO (CNPq/CT-INFO 55.2087/2002-5) e CTI2 (CNPq/Edital Universal 47.9564/2006-0).

Referências

Almeida, H., Goncalves, M., Cristo, M., and Calado, P. (2007). A combined component approach for finding collection-adapted ranking functions based on genetic programming. In *Proc. 30th Int'l. ACM SIGIR Conf.*, Amsterdã, Países Baixos.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Bender, M., Michel, S., Triantafillou, P., Weikum, G., and Zimmer, C. (2006). P2P content search: give the Web back to the people. In *5th IPTPS WS*, Santa Bárbara, EUA.

Calado, P. (1999). The WBR-99 collection: description of the WBR-99 Web collection data-structures and file formats. In *LAb. Treating INformation*, Belo Horizonte, Brasil.

Callan, J. and Connell, M. (2001). Query-based sampling of text databases. In *ACM TOIS*, Nova Iorque, EUA.

Costa, C., Soares, V., Benevenuto, F., Vasconcelos, M., Almeida, J., Almeida, V., and Mowbray, M. (2006). Disseminação de Conteúdo Poluído em Redes P2P. In *Proc. 24th SBRC Symp.*, Curitiba, Brasil.

Cuenca-Acuna, F. M., Peery, C., Martin, R. P., and Nguyen, T. D. (2003). PlanetP: using gossiping to build content addressable peer-to-peer information sharing communities. In *Proc. 12th IEEE HPDC Symp.*, Seattle, EUA.

Demers, A., Greene, D., Hauser, C., Irish, W., Larson, J., Shenker, S., Sturgis, H., Swinehart, D., and Terry, D. (1987). Epidemic algorithms for replicated database maintenance. In *Proc. 6th ACM PODC Conf.*, Vancouver, Canadá.

Douceur, J. R. (2002). The sybil attack. In *Proc. 1st IPTPS WS*, Cambridge, EUA.

Doulkeridis, C., Norvag, K., and Vazirgiannis, M. (2006). The SOWES approach to P2P Web search using semantic overlays. In *Proc. 15th WWW Conf.*, Edinburgh, Escócia.

Fetterly, D., Manasse, M., and Najork, M. (2004). Spam, damn spam, and statistics: using statistical analysis to locate spam Web pages. In *Proc. 7th WebDB WS*, Paris, França.

Fishwick, P. A. (1992). SimPack: Getting Started with Simulation Programming in C and C++. In *24th WSC Conf.*, Arlingtion, VA, USA.

Gopalakrishnan, V., Bhattacharjee, B., and Keleher, P. (2006). Distributing Google. In *2nd IEEE NetDB WS*.

Gyongyi, Z. and Garcia-Molina, H. (2005). Web spam taxonomy. In *Proc. 1st AIRWeb WS*, Chiba, Japão.

Kishida, K. (2005). Property of average precision and its generalization: an examination of evaluation indicator for information retrieval experiments. Technical report, NII, Tóquio, Japão.

Lewandowski, D. and Mayr, P. (2007). Exploring the academic invisible web. In *Library Hi Tech*.

Liang, J., Kumar, R., Xi, Y., and Ross, K. W. (2005). Pollution in P2P file sharing systems. In *Proc. 24th IEEE Infocom Conf.*, Miami, EUA.

Lu, J. and Callan, J. (2003). Content-based retrieval in hybrid peer-to-peer networks. In *Proc. 12th CIKM Conf.*, Nova Orleans, EUA.

Marti, S. and Garcia-Molina, H. (2006). Taxonomy of trust: categorizing P2P reputation systems. In *Computer Networks*, Stanford, EUA.

Menasce, D. A. and Almeida, V. A. F. (2001). *Capacity Planning for Web Services: metrics, models, and methods*. Prentice Hall.

Ntoulas, A. and Cho, J. (2007). Pruning policies for two-tiered inverted index with correctness guarantee. In *Proc. 30th Int'l. ACM SIGIR Conf.*, Amsterdã, Países Baixos.

Pouwelse, J. A., Garbacki, P., Epema, D. H. J., and Sips, H. J. (2005). The Bittorrent P2P File-Sharing System: Measurements and Analysis. In *4th Int'l. IPTPS WS*, Ithaca, EUA.

Rhea, S. C., Geels, D., Roscoe, T., and Kubiatowicz, J. (2004). Handling Churn in a DHT. In *Proc. USENIX Conf.*, Boston, EUA.

Stutzbach, D. and Rejaie, R. (2006). Understanding churn in peer-to-peer networks. In *Proc. 6th ACM SIGCOMM IMC Conf.*, Rio de Janeiro, Brasil.

Voorhees, E. M. and Harman, D. (1999). Overview of the eighth text retrieval conference (TREC-8). In *National Institute of Standards and Technology*, Gaithersburg, EUA.

Wu, L., Akavipat, R., and Menczer, F. (2005). 6S: distributing crawling and searching across Web peers. In *Proc. WTAS*, Calgary, Canadá.

Yu, W., Chellappan, S., and Xuan, D. (2005). P2P/Grid-based overlay architecture to support VoIP services in large-scale IP networks. In *FGCS J.*