Disseminação de Conteúdo Poluído em redes P2P

Cristiano Costa[1], Vanessa Soares[1], Fabricio Benevenuto[1], Marisa Vasconcelos[1], Jussara Almeida[1], Virgilio Almeida[1], Miranda Mowbray[2]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil

[2]HP Labs
Bristol, UK

{krusty, vanessa, fabricio, isa, jussara, virgilio}@dcc.ufmg.br

miranda.mowbray@hp.com

Abstract. *Recently, file sharing Peer-to-Peer (P2P) systems are experimenting a new form of malicious behaviour:* content pollution*. The dissemination of polluted content reduces content availability and consequently the confidence of user in such systems. This work points two strategies used to pollute P2P systems and investigates the dissemination of pollution by the use of these strategies in moderate networks (e.g. KaZaA) and non-moderate networks (e.g. BitTorrent). Our results show that only a high level of incentive to users delete their polluted files or the presence of a system moderator is able to stop the propagation of pollution in P2P systems.*

Resumo. *Recentemente, sistemas Par-a-Par (P2P) para compartilhamento de arquivos vêm experimentando uma nova forma de comportamento malicioso:* poluição de conteúdo*. A disseminação de conteúdo poluído reduz a disponibilidade dos arquivos e conseqüentemente a confiabilidade dos usuários no sistema. Este trabalho aponta duas estratégias utilizadas para poluir conteúdo e investiga a disseminação de conteúdo poluído com o uso destas estratégias em Redes Não Moderadas (ex: KaZaA) e em Redes Moderadas (ex: BitTorrent). Os resultados mostram que apenas nível de incentivo alto para que os usuários apaguem seus arquivos poluídos ou a presença de um moderador no sistema é capaz de conter a propagação de poluição em sistemas P2P.*

1. Introdução

Desde seu surgimento, os sistemas Par-a-Par (P2P) para compartilhamento de arquivos vêm crescendo rapidamente em número de usuários, arquivos compartilhados e em tráfego na Internet. De fato, a maior parte de todo o tráfego na Internet hoje em dia é causada por aplicações P2P [Saroiu et al. 2002].

A rápida evolução de sistemas P2P não vem causando impacto somente no tráfego da Internet, mas também, na indústria de música e vídeo. Estas indústrias experimentaram perdas de milhões de dólares em vendas de CDs e DVDs devido à pirataria promovida pelas redes P2P. Desde a primeira ação legal contra o Napster [Kurose and Ross 2005], existe uma verdadeira guerra entre a indústria de música e vídeo contra a pirataria e

consequentemente contra sistemas P2P. Com o surgimento de sistemas de P2P descentralizados como o Gnutella [Gnutella], KaZaA [KaZaA], eDonkey [eDonkey] e BitTorrent [BitTorrent], criou-se uma dificuldade para a interrupção desses sistemas, o que levou a indústria de música, sem muito sucesso, a processar os usuários destes sistemas.

Entretanto, um estudo recente [Liang et al. 2005] mostrou evidências de que a intervenção da indústria de música através da disseminação de conteúdo poluído tem obtido um sucesso significativo. Poluição consiste na disseminação de cópias de um arquivo específico (uma música ou filme), com o mesmo metadado de um arquivo não poluído (ex.: nome, artista), mas com conteúdo corrompido [Christin et al. 2005]. Utilizando esta estratégia, as companias de música visam tornar mais difícil para os usuários encontrarem uma cópia correta do arquivo procurado. Em [Liang et al. 2005] foi mostrado que mais de 50% dos arquivos encontrados através de buscas na rede FastTrack [Fasttrack] são poluídos.

Existem várias técnicas para disseminar conteúdo poluído em sistemas P2P. A mais conhecida é chamada de *inserção de versões falsas*. Quando um usuário procura por um arquivo, o sistema P2P retorna versões desse arquivo, onde cada versão possui um número de cópias. Esta técnica consiste em inserir uma versão falsa do arquivo na rede com a finalidade de dificultar um usuário encontrar uma versão correta do arquivo. Outro importante mecanismo para disseminação de conteúdo é a inserção de versões corrompidas na rede com o mesmo identificador de um arquivo já existente e correto. Essta técnica é chamada de *corrupção por chave*.

Existem vários tipos de sistemas P2P sendo utilizados hoje em dia e cada um destes sistemas possui um determinado nível de poluição dos arquivos. Neste trabalho foi avaliado o espalhamento de poluição em dois tipos de redes: moderadas e não moderadas. Nas redes moderadas existe a presença humana de um agente moderador capaz de interferir nos arquivos que estão sendo compartilhados na rede, como no caso do BitTorrent. As redes não moderadas constituem os demais tipos de redes e não possuem este agente centralizador. Além disso, foi avaliado qual o efeito no espalhamento de poluição pela rede quando os usuários do sistema recebem incentivos para apagarem seus arquivos poluídos. Foi mostrado em [Liang et al. 2005] que usuários não classificam seus arquivos. Nós acreditamos que os grandes responsáveis pelo espalhamento de conteúdo poluído em algumas redes são os próprios usuários, que não apagam arquivos poluídos após o download. Foi avaliado até quando estratégias de incentivos para que usuários apaguem seus arquivos poluídos podem ser eficientes para evitar poluição.

As demais seções deste trabalho são organizadas da seguinte forma. Seção 2. apresenta trabalhos relacionados. A seção 3. discute os principais aspectos das redes P2P consideradas neste trabalho e descreve duas estratégias utilizadas para disseminar poluição nestas redes. A seção 4. apresenta a metodologia adotada, descreve o simulador desenvolvido e as métricas e parâmetros utilizados em nossos experimentos. Os principais resultados são apresentados na seção 5.. A seção 6. discute estratégias para conter o avanço da poluição em sistemas P2P e a seção 7. oferece conclusões e direções futuras para esse trabalho.

2. Trabalhos Relacionados

A Poluição em redes P2P é recente e por isso existem poucos trabalhos sobre esse fenômeno. O primeiro estudo [Liang et al. 2005] sobre poluição foi feito na rede FastTrack [Fasttrack]. Os autores desenvolveram uma ferramenta capaz de buscar e baixar arquivos na rede FastTrack. Com base nos dados coletados, eles mostram que o nível de de poluição nessa rede é muito alto, principalmente para arquivos recentes e populares. Christin et al. [Christin et al. 2005] avalia a disponibilidade de conteúdo e o conteúdo corrompido em três sistemas P2P: KaZaA [KaZaA], eDonkey [eDonkey] and Gnutella [Gnutella]. Pouwelse et al. [Pouwelse et al. 2005] analisa a integridade do sistema BitTorrent/Suprnova inserindo conteúdo poluído no sistema. Eles mostram que, neste tipo de sistema, a eficiência dos moderadores torna a disseminação de conteúdo poluído mais difícil.

Estes trabalhos foram responsáveis pelas primeiras análises sobre conteúdo poluído em redes P2P, e mostraram que poluição pode realmente ser prejudicial para estes sistemas. Entretanto, esses trabalhos não avaliam como o conteúdo poluído é disseminado e não exploram técnicas utilizadas pela indústria de música para espalhar poluição. Além disso, eles consideram apenas inserção de versões falsas como método utilizado. Apesar de não haverem evidências concretas de conteúdo poluído disseminado por corrupção por chave, várias discussões sobre poluição causada por este mecanismo podem ser encontradas na Internet. Por exemplo, uma companhia chamada Viralg afirma ser capaz de erradicar qualquer conteúdo desejado de um sistema P2P utilizando o mecanismo de corrupção de chave [Viralg].

Outras ameaças a sistemas P2P são vermes e spyware. Existem alguns trabalhos recentes que estudaram o espalhamento de vermes e de spyware entre usuários P2P. O espalhamento de spyware foi estudado utilizando logs de um ambiente universitário em [Saroiu et al. 2004]. Um modelo analítico para avaliação da propagação de conteúdo e ataques de vermes em P2P é proposto e validado por [Yu et al. 2005]. Entretanto, existe uma diferença significativa entre disseminação de poluição e a disseminação de vermes. A propagação de conteúdo poluído é causada pelos próprios usuários, enquanto que a propagação de vermes é desencadeada por si só, em um mecanismo recursivo.

3. Sistemas P2P

Esta seção discute brevemente os principais aspectos das redes P2P abordadas neste trabalho além de apresentar as duas principais técnicas para disseminação de conteúdo poluído.

3.1. Arquiteturas P2P

Nas sub-seções seguintes, nós apresentamos uma descrição superficial das principais características das redes moderadas e não moderadas que são úteis para o entendimento deste trabalho. Para uma descrição mais detalhada destas redes, nós direcionamos o leitor para as referências [Benevenuto et al. 2004] e [Pouwelse et al. 2005], que descrevem redes moderadas e não moderadas respectivamente.

3.1.1. Redes Não Moderadas

Existem vários tipos de topologias de redes que não utilizam um moderador capaz de controlar o conteúdo disponível na rede. Os sistemas P2P para compartilhamento de arquivos podem ser categorizados em três grupos, com base no mecanismo de localização de conteúdo. O primeiro deles, utilizado no Napster [Napster], é baseado em um servidor central com a localização de todos os arquivos compartilhados no sistema. Em sistemas descentralizados estruturados, tais como Chord [Stoica et al. 2001], Pastry [Castro et al. 2002], e CAN [Ratnasamy et al. 2001], os índices dos arquivos são armazenados nos nós participantes, que são organizados em uma estrutura bem definida utilizada para roteamento de consultas. O terceiro grupo consiste de sistemas descentralizados não estruturados, tais como Gnutella [Gnutella] e KaZaA [KaZaA].

Em particular, neste último grupo estão as arquiteturas hierárquicas utilizadas no KaZaA, que se destacam em termos de popularidade de usuários. A maior parte das aplicações de P2P não moderadas utilizam esta arquitetura. O funcionamento geral destes sistemas é feito de forma que cada nó esteja conectado a um determinado número de super-nós. Quando um usuário procura um arquivo, os super-nós deste nó recebem uma mensagem, podendo repassá-la ou não, dependendo da topologia e da arquitetura adotada. Os nós que recebem esta mensagem e possuem este arquivo respondem para o nó origem. Ao fim da busca o usuário escolhe qual versão do arquivo ele quer baixar. A descoberta de novos super-nós se dá através de mensagens de manutenção da rede enviadas periodicamente [Benevenuto et al. 2005].

3.1.2. Redes Moderadas

Dentre as redes moderadas existentes hoje em dia, as mais conhecidas são a do BitTorrent [BitTorrent] e do eDonkey [eDonkey]. O mecanismo utilizado pelo BitTorrent consiste na publicação de informações sobre o arquivo alvo. Os usuários que decidem baixar este arquivo, na verdade estarão baixando pedaços uns dos outros, seguindo uma política de "olho por olho, dente por dente". O moderador do arquivo consiste na pessoa que disponibiliza o metadado do arquivo (ex: nome, autor) e mantém a lista de usuários que estão baixando o arquivo. Este moderador pode apagar determinada lista de download caso ele perceba que aquele arquivo está poluído.

A rede do eDonkey possui servidores centrais, nos quais os nós se conectam para participarem da rede. Este servidores não possuem arquivos compartilhados, mas possuem informações dos arquivos que os nós a eles conectados possuem e com isso, controlam as buscas dos nós. Várias aplicações que conectam na rede do eDonkey permitem que o usuário digite o identificador do arquivo desejado ao invés realizar uma busca pelo arquivo na rede. A partir deste mecanismo simples, surgiram alguns sítios na Internet que disponibilizam uma lista de arquivos não poluídos e suas respectivas chaves. O usuário procura o arquivo em um sítio confiável e então digita o identificador do arquivo na aplicação da rede eDonkey, que conecta ao servidor central e obtém a lista de usuário que possui o arquivo para download. Note que este mecanismo é muito semelhante ao da rede do BitTorrent.

3.2. Estratégias para Disseminação de Conteúdo Poluído

Esta seção descreve duas estratégias para disseminação de conteúdo poluído: a de *inserção de cópias falsas* e a de *corrupção por chave*.

3.2.1. Inserção de Versões Falsas

A inserção de versões falsas é uma técnica comum de sabotagem utilizada nos sistemas P2P de compartilhamento de arquivos. Esta técnica consiste na disseminação na rede de versões poluídas de um arquivo com o intuído de dificultar a localização de uma versão não poluída do arquivo na rede. Os arquivos poluídos inseridos na rede, contém os mesmos metadados (ex.: nome, autor) dos arquivos corretos. De forma geral, quando um usuário procura por um arquivo, a aplicação P2P agrupa as cópias em diferentes versões e apresenta as versões com o maior número de cópias para o usuário. Caso, o usuário faça o download de uma cópia poluída e não o apague imediatamente, esta cópia pode se espalhar, tornando cada vez mais difícil encontrar uma versão verdadeira.

Os ataques que utilizam esta técnica consistem em espalhar versões poluídas de algum conteúdo mesmo antes que ele se torne popular. Feito isso, os usuários que baixam essas cópias geralmente não as apagam [Liang et al. 2005] e encarregam de disseminar e tornar mais populares as versões poluídas. Algumas empresas como Viralg [Viralg], RetSpan [Retspan] e OverPeer [Overpeer] oferecem serviços para eliminar qualquer conteúdo não autorizado distribuído em redes P2P. Elas utilizam um grande número de máquinas compartilhando um grande número de versões poluídas do conteúdo alvo.

3.2.2. Corrupção por Chave

Em um sistema P2P, quando um usuário começa a compartilhar um arquivo na rede, é criado um identificador (ID) único que é associado àquele arquivo. Este ID permite às aplicações identificarem os arquivos que os usuários compartilham. Além disso, quando um usuário recebe o resultado de uma busca, o cliente P2P agrupa os resultados com o mesmo ID, para que o arquivo possa ser baixado por múltiplas fontes simulataneamente. Este identificador é gerado aplicando-se uma função de hash no conteúdo do arquivo. Cada sistema P2P utiliza um algoritmo diferente para geração do identificador.

Os sistemas P2P assumem que o ID gerado pela função hash é único. Entretanto, existe a possibilidade de haver dois arquivos diferentes com o mesmo identificador. Isso acontece principalmente porque alguns dos algoritmos mais comuns para gerar o ID são baseados em apenas partes do arquivo. Sendo assim, nós maliciosos podem alterar as partes do arquivo que não foram utilizadas pelo algoritmo para gerar o identificador, criando assim diferentes arquivos com o mesmo ID. Quando um usuário requisita um arquivo, ele recebe uma lista de versões do arquivo, cada uma identificada com um ID distinto e com um certo número de cópias. Feito isso o usuário escolhe uma versão e baixa pedaços de diferentes cópias vindas de diferentes usuários. Se um pedaço baixado corresponde a uma parte corrompida do arquivo, o arquivo inteiro ficará comprometido. O nome dessa técnica de poluição é Corrupção por chave.

É importante ressaltar que o mecanismo de corrupção por chave não necessaria-

mente quebra funções de hash como MD4, MD5 e SHA-1. Esse mecanismo se aproveita da forma que os sistemas P2P utilizam estas técnicas para criar o identificador dos arquivos. Como exemplo podemos citar a rede do FastTrack [Fasttrack] que utiliza o algoritmo chamado uuhash [UUHash]. Este algoritmo gera o ID a partir de algumas partes do conteúdo do arquivo. Consequentemente, um arquivo diferente com o mesmo ID pode ser criado alterando-se partes do arquivo que não são utilizadas como entrada para a função hash. Outra forma de corromper o arquivo seria alterando os clientes P2P de forma que este não computasse corretamente o ID de um arquivo, associando um arquivo corrompido ao ID do arquivo alvo.

4. Modelo de Avaliação

Esta seção descreve a metodologia usada em nesse estudo. Foi desenvolvido um simulador orientado por eventos capaz de reproduzir os principais aspectos da disseminação de conteúdo poluído nos dois tipos de redes analisadas. A seção 4.1. apresenta os principais aspectos dos simuladores implementados e a seção 4.2. apresenta as métricas e parâmetros da simulação.

4.1. Simuladores de Redes P2P

Para avaliar a disseminação de conteúdo poluído, foram desenvolvidos dois simuladores: um para redes moderadas e outro para redes não moderadas. Ambos os simuladores compartilham as características descritas a seguir.

Os simuladores avaliam a disseminação de poluição de vários arquivos diferentes compartilhados em uma rede P2P. A simulação começa com um número constante de nós ativos (*online*), no caso especifico de nossas simulações 50% do total de nós. O número total de nós na rede é de 5.000 (2.500 nós ativos). No inicio da simulação cada nó têm armazenado 1 arquivo em seu disco. O número total de arquivos diferentes no sistema foi definido como 20 e cada arquivo possui 100 versões diferentes, ou seja, 2.000 arquivos com IDs únicos no total. Tanto a popularidade dos arquivos quanto das versões segue uma distribuição Zipf com coeficiente α igual a 1.

Cada nó ativo baixa um arquivo numa taxa de 1 arquivo a cada 5.000 unidades de simulação (seguindo uma distribuição exponencial). Para requisitar um arquivo, um nó ativo escolhe um dos 20 arquivos existentes no sistema. Depois de requisitar um arquivo, o nó recebe uma lista de versões daquele arquivo e o número de usuários (fontes) que possuem cópias de cada versão. O nó então escolhe uma das versões para baixar de acordo com a popularidade desta versão. Por exemplo, se 50% das cópias disponíveis são de uma única versão, então a probabilidade do nó escolher esta versão é 0,5. Como os mecanismos de buscas não são o escopo destre trabalho, a modelagem da topologia da rede sobreposta é desnecessária. Sendo assim, foi assumido que os nós sempre acham todas as versões e cópias do arquivo procurado. O nó, então, seleciona de quais fontes ele deseja obter o arquivo e este é recebido instâneamente. Por simplicidade, não foi modelado o tempo de transferência dos arquivos.

Foi assumido que quando um nó baixa um arquivo e este não é imediatamente apagado, ele é compartilhado, ou seja, ele pode ser baixado por outros nós. Além disso, os nós podem baixar mais de um arquivo diferente e repetir a mesma busca, caso já não tenham o arquivo ou tenham apagado o arquivo baixado anteriormente. O tempo em que

cada nó fica ativo e inativo segue uma distribuição exponencial com média de 10.000 passos de simulação.

4.2. Métricas e Parâmetros

Com o intuito de avaliar a disseminação de poluição, foi definida a métrica *disseminação de conteúdo poluído* como sendo a porcentagem de cópias poluídas dos nós ativos no sistema. Foi avaliada a disseminação de conteúdo poluído nas redes moderadas e não moderadas e, para isso, foram criados parâmetros específicos e apropriados para cada um dos tipos rede. Os parâmetros para as redes moderadas foram os seguintes:

- **Incentivo para apagar** (IA): corresponde à probabilidade de um usuário apagar o arquivo poluído imediatamente após o término de seu download. Esta métrica consiste em todos os mecanismos dos sistemas P2P sem moderadores que dão incentivos para o usuário apagar de sua máquina um arquivo poluído compartilhado. É importante ressaltar que não foi criado e nem avaliado nenhum mecanismo específico para acabar com a poluição.
- **Vulnerabilidade do Hash** (VH): para avaliar a classe inteira de algoritmos de geração de ID de arquivos foi definido VH, como a porcentagem do arquivo que pode ser corrompido sem alterar seu ID. Por exemplo, se algoritmo uuhash [UUHash] for utilizado para gerar os ID de arquivos de 5 MB e 600 MB, os VH destes arquivos seriam 88% e 99,5% respectivamente.
- **Número de fontes de download** (NF): consiste no número máximo de fontes simultâneas que um arquivo pode ser baixado. Foi avaliada a disseminação de conteúdo para diferentes valores de NF para a estratégia de corrupção por chave. Pode-se notar, que NF não possui impacto para a técnica de inserção de versões falsas, já que todas as cópias de uma versão estão poluídas ou não.

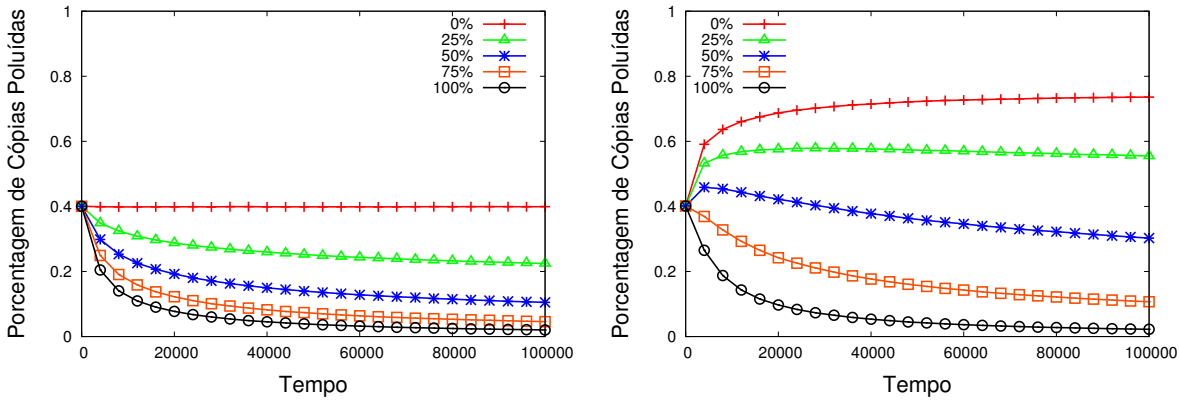
Para a avaliação de disseminação de conteúdo poluído em redes moderadas foram utilizados os seguintes parâmetros:

- **Incentivo para reportar** (IR): Em redes moderadas, a poluição pode ser evitada quando os usuários comunicam ao moderador que o arquivo baixado está corrompido. Então o moderador é encarregado de apagar todas as referências daquele arquivo e evitar que o conteúdo poluído se espalhe. Chamamos de Incentivo para reportar, IR, a probabilidade de um usuário reportar um arquivo poluído para o moderador do sistema.
- **Tempo de reação do moderador**(TDM): A presença humana como moderador de um sistema P2P claramente possui problemas de escalabilidade. Gerenciar um sistema com muitos usuários pode se tornar uma tarefa complexa e as ações do moderador contra conteúdo poluído podem não ser imediatas. Além disso, quando os usuários reportam ao moderador do sistema que determinado arquivo está poluído, o moderador ainda precisa verificar se o conteúdo reportado está realmente corrompido. Desta forma, definimos o tempo de reação do moderador, TDM, como sendo o tempo que o moderador do sistema leva para apagar um arquivo poluído reportado por um usuário.

5. Resultados

Esta seção apresenta os resultados relativos à disseminação de conteúdo poluído em redes não moderadas e em redes moderadas. Os gráficos apresentados mostram a

disseminação de conteúdo poluído. O eixo Y indica a porcentagem de cópias no sistema que estão poluídas e o eixo X indica o tempo de simulação percorrido. São apresentados os resultados para um sistema que possui 5.000 nós e com 40% das cópias poluídas. Os resultados foram qualitativamente semelhantes para os experimentos em que o número inicial de nós e a porcentagem inicial de cópias poluídas possuiam valores diferentes. A porcentagem de cópias poluídas utilizada na simulação é típica de sistemas reais [Liang et al. 2005]. A simulação termina após 100.000 passos de simulação. Cada resultado da simulação é a média de 20 execuções de um mesmo experimento. Com nível de confiança de 95%, os resultados diferem da média em no máximo 10%.



(a) Poluição por inserção de versões falsas

(b) Poluição por corrupção por chave. $VH = 90\%$

Figura 1. Disseminação de cópias poluídas variando o incentivo para apagar (IA) para a rede não moderada. $NF = 10$

A disseminação de cópias poluídas para a estratégia de inserção de versões falsas é apresentada na figura 1(a). Cada curva mostra o número de cópias poluídas para diferentes valores de incentivo para apagar (IA). Como esperado, mesmo para valores pequenos de IA, a porcentagem de cópias poluídas na rede decresce com o passar do tempo. Pode-se notar, que quando os usuários não apagam seus arquivos poluídos, a porcentagem de cópias corrompidas na rede se mantêm. Isto acontece porque a probabilidade de um usuário entrar no sistema e requisitar uma versão poluída é igual à porcentagem de arquivos que estão corrompidos, mantendo a taxa de cópias poluídas e não poluídas para cada arquivo. Quando o incentivo para apagar aumenta, a porcentagem de conteúdo poluído no sistema diminui significativamente. Se a probabilidade dos usuários apagarem seus conteúdos poluídos imediatamente após o download for 0,25 ($IA = 0,25$), a porcentagem de cópias poluídas na rede decresce para 22,5% quando a simulação termina.

A figura 1(b) mostra o gráfico correspondente para o mecanismo de corrupção por chave. Neste cenário, inicialmente, todas as versões possuem a mesma porcentagem de cópias poluídas, de forma que versões mais populares possuem mais cópias poluídas. Foi assumido que um usuário baixa conteúdo de 10 fontes simultaneamente ($NF = 10$) e que inicialmente um arquivo poluído contém 90% de seu conteúdo corrompido pelo mecanismo de corrupção por chave (VH = 90%). Pode-se observar que, mesmo quando os usuários possuem um incentivo para apagar ($IA > 0$), o número de arquivos poluídos

aumenta com o tempo, se a poluição for disseminada pela técnica de corrupção por chave. Somente quando os usuários recebem um incentivo maior para apagar, mais de 50% dos arquivos poluídos, é que a disseminação de conteúdo poluído no sistema começa a diminuir. Isso ocorre porque o download é feito de 10 fontes diferentes, e todas as versões possuem cópias poluídas.

Comparando os mecanismos de inserção de versões falsas e corrupção por chave, podemos notar a eficácia do segundo mecanismo. A eficiência em espalhar poluição deste mecanismo ocorre porque, em geral, o download é feito de várias fontes ($NF = 10$) e todas as versões possuem cópias poluídas. Por exemplo, se apenas uma das dez fontes de download fornecer uma cópia poluída e o usuário baixar um pedaço poluído dessa fonte, o arquivo obtido estará comprometido. A poluição, causada pelo mecanismo de inserção de versões falsas, pode ser drasticamente reduzida aumentando-se o incentivo para apagar, enquanto que para o mecanismo de corrupção por chave o efeito de aumentar o incentivo de apagar é menos efetivo. Pode-se observar nas figuras 1(a) e 1(b) que para o mecanismo de inserção de versões falsas a porcentagem de poluição na rede diminui 44% ao final da simulação para IA igual a 50%, enquanto que esta porcentagem aumenta 38% utilizando a técnica de corrupção por chave para o mesmo IA.

A figura 2 mostra resultados variando a vulnerabilidade do hash (VH). Somente para essa avaliação assumiu-se que nenhum usuário apaga seus arquivos poluídos ($IA = 0$) e que o número de fontes de download simultâneos é fixado em 10 ($NF = 10$). As curvas mostram que mesmo para um valor pequeno de vulnerabilidade do hash, há um forte impacto no número de cópias poluídas no sistema com o passar do tempo. Por exemplo, com 30% de arquivos poluídos ($VH = 30\%$), a porcentagem de cópias poluídas chega a 48% ao final da simulação. Isto acontece porque quando a vulnerabilidade do hash aumenta, a probabilidade de baixar um pedaço corrompido de uma fonte que fornece conteúdo poluído também aumenta.

A figura 3 mostra como a disseminação de poluíção, com a técnica de de poluição por corrupção de chave, é afetada com o aumento do número de fontes (NF) das quais o arquivo é obtido. Como esperado, quando maior o número de fontes das quais o arquivo é baixado, maior se torna a probabilidade de se baixar um pedaço de uma fonte poluída. Pode-se notar que quando o download é feito de uma única fonte, o mecanismo de corrupção por chave mantém constante a proporção de cópias poluídas, da mesma forma que o mecanismo de inserção de versões falsas.

A figura 4 mostra a comparação entre o incentivo para apagar em redes não moderadas e o incentivo para reportar em redes moderadas. Nas redes não moderadas o incentivo para apagar de 25% diminui a porcentagem de cópias poluídas em 45% ao fim da simulação (Figura 4-a). Para as redes moderadas foi assumido um intervalo fixo de 50.000 unidades de simulação desde momento em que o moderador recebe a requisição do usuário e apaga o arquivo poluído. Com um incentivo de 0,25 (25% dos usuários que baixam um conteúdo poluído reportam), o moderador consegue excluir da rede todos os arquivos poluídos no fim da simulação (Figura 4-b). Pode-se notar que, o incentivo de reportar nas redes moderadas tem um desempenho melhor do que o incentivo para apagar nas redes não moderadas. Isto acontece pois nas redes moderadas existe uma entidade centralizada, o moderador, que tem o poder de apagar todas as cópias de uma determinada versão.

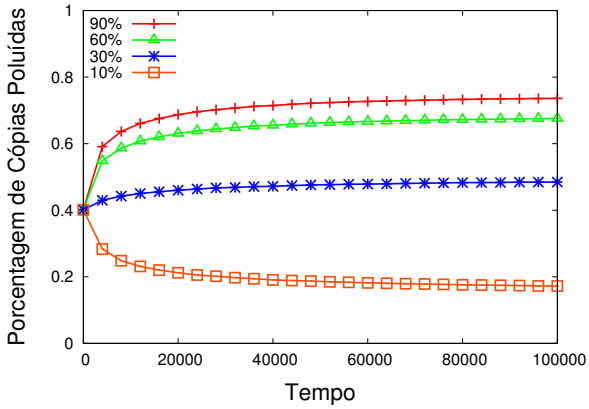


Figura 2. Disseminação de cópias poluídas pelo mecanismo de corrupção por chave variando a vulnerabilidade do hash (VH). $IA = 0$ e $NF = 10$

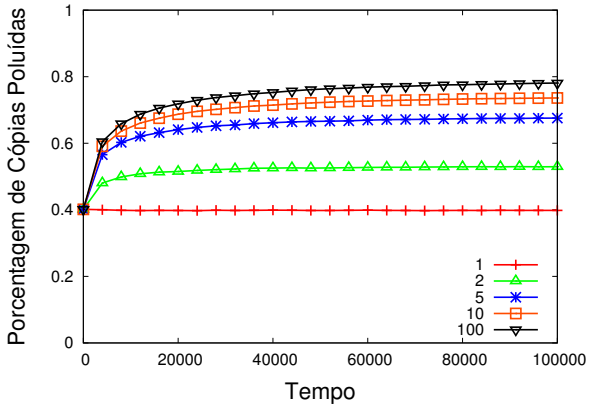


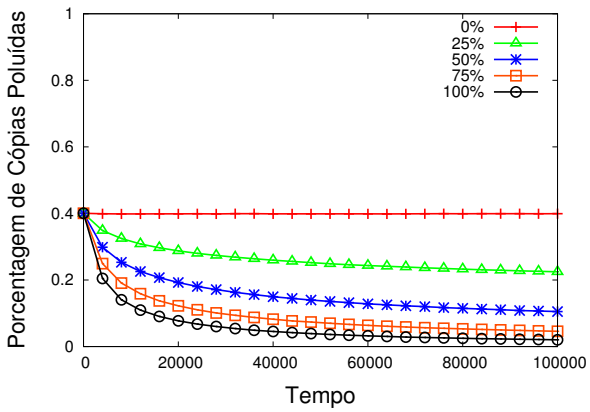
Figura 3. Disseminação de cópias poluídas pelo mecanismo de corrupção por chave variando o número de fontes de download simultâneas (NF). $IA = 0$ and $VH = 90\%$

A Figura 5 mostra a porcentagem de cópias poluídas variando-se o intervalo que o moderador demora para apagar um versão poluída reportada pelo usuário (este intervalo é modelado como um número fixo). Neste experimento, 25% dos clientes que baixam conteúdo poluído reportam. Estes resultados mostram que, uma vez que a poluição é reportada, a eficiência para reduzir a poluição depende somente do tempo gasto pelo moderador. Quando o moderador começa a verificar e atender aos usuários que reportaram a poluição, a rede é limpa em um curto período de tempo.

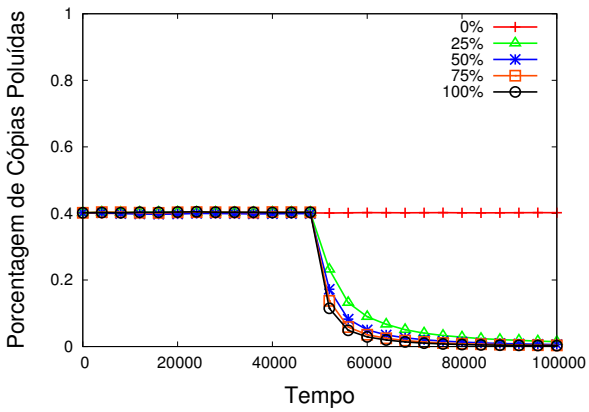
6. Mecanismos de Combate à Poluição

Neste trabalho mostramos que poluição pode ser bastante prejudicial para uma sistema P2P e como os incentivos ajudam a combater esse mal. Uma pergunta que surge ao analisar nossos resultados é o que pode ser feito para conter a poluição? Esta seção discute mecanismos que podem ser utilizados para o combate à poluição em redes P2P.

Nas redes moderadas o problema é minimizado devido controle centralizado do sistema e à presença de um moderador capaz de gerenciar de forma eficiente os arquivos distribuídos. Já que existe ação humana envolvida, a escalabilidade desse método ainda precisa ser avaliada com um maior cuidado. Porém, como pudemos ver na seção 5., o



(a) Incentivo para apagar em redes não moderadas



(b) incentivo para reportar em redes moderadas (tempo do moderador apagar fixado em 50.000 unidades de tempo)

Figura 4. Disseminação de cópias poluídas para o mecanismo de Inserção de Cópias Falsas variando os incentivos

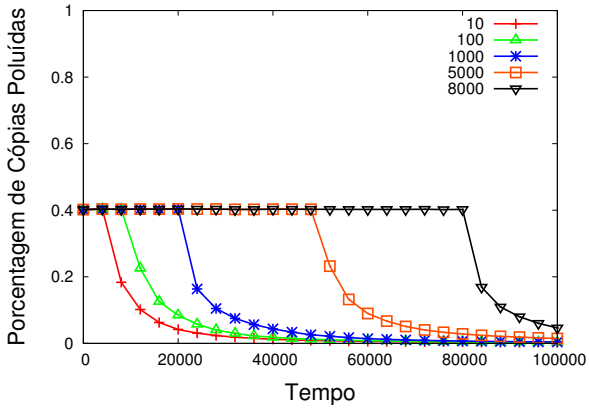


Figura 5. Disseminação de cópias poluídas em redes moderadas variando o tempo que o moderador leva para deletar (intervalos fixos e incentivo para reportar $= 0,25$)

maior problema está nas redes não moderadas, nas quais a disseminação de poluição é mais dificilmente evitada do que nas redes moderadas.

Uma idéia para combater a poluição nesses sistemas seria ter disponível um banco de dados confiável contendo todos os IDs dos arquivos existentes no sistema (ou IDs de pedaços de arquivos). Sempre que um usuário realizasse um download, a integridade do arquivo recebido poderia ser verificada com a base de dados. Caso o arquivo esteja corrompido, este é apagado imediatamente. Dentre algumas tentativas nesta direção podemos destacar o projeto Sig2dat [tool for FastTrack network], que disponibilizou uma ferramenta que calcula o identificador de arquivos da rede FastTrack. Desde então, vários sites vêm disponibilizando listas "negras" ou "brancas" de arquivos e seus respectivos identificadores para que os usuários possam se orientar antes de realizar o download. Porém, para que esse mecanismo funcione é necessário que o download do arquivo seja

realizado, o que acabaria desperdiçando banda da rede.

Uma outra forma de conter a poluição é fazer com que os próprios usuários apaguem seus arquivos poluídos logo após o download. Neste trabalho, mostramos que quando a maior parte dos usuários apagam seus arquivos poluídos ocorre uma redução no nível de poluição do sistema. Entretanto, criar um sistema P2P que consiga incentivar o usuário a apagar arquivos poluídos ainda é um desafio.

Outros mecanismos com uma maior chance de sucesso são os sistemas de reputação. Mecanismos de reputação classificam um agente do sistema, para que outros agentes possam escolher como se relacionar com ele de acordo com a sua reputação. Alguns trabalhos de sistemas de reputação para conter *free-riders* em redes P2P já foram propostos[**?**]. Para conter poluição existe o Credence [Walsh and Sirer 2005] que é um mecanismos que reputa os arquivos distribuídos na rede. Sempre que um nodo deseja baixar um arquivo ele pergunta para outros nodos confiáveis se o arquivo desejado está poluído ou não. A partir das respostas obtidas, o nodo decide se o download será realizado. Entretanto, como o Credence reputa arquivos, a sua eficiência provavelmente ficaria comprometida no caso de poluição por corrupção por chave. Além disso, esse mecanismo não pune os nodos maliciosos que disseminam a poluição, o que permite que eles continuem a usufruir dos recursos da rede. E, novamente, a participação do usuário é essencial para identificar se um arquivo está corrompido e para determinar se o nó fonte é ou não um agente poluidor.

7. Conclusão e Trabalhos Futuros

Este trabalho analisa a disseminação de conteúdo em redes P2P. Foram consideradas duas estratégias para poluir sistemas P2P, com e sem moderadores: inserção de versões falsas e corrupção por chave. Mostramos que a corrupção por chave espalha poluição mais rapidamente do que a poluição por inserção de versões falsas. Mesmo quando usuários possuem um nível alto de incentivo para apagar conteúdo poluído, o mecanismo de corrupção por chave se mostra uma eficiente forma para disseminar poluição. Comparando redes moderadas com redes não moderadas, nossos resultados mostraram que a presença de um moderador na rede pode ser uma boa forma de conter o espalhamento de poluição. Entretanto, a presença humana centralizadora em um sistema com muitos usuários claramente pode apresentar problemas de escalabilidade.

Direções para trabalhos futuros incluem o estudo e análise de mecanismos capazes de conter a disseminação de conteúdo em sistemas P2P. Mais especificamente pretendemos criar um novo mecanismo de reputação para combater todos os tipos de poluição e investigar como a disseminação de poluição é afetada por outros mecanismos como o Credence [Walsh and Sirer 2005].

8. Agradecimentos

Este trabalho foi desenvolvido em colaboração com a HP Brasil R&D.

Referências

[Benevenuto et al. 2004] Benevenuto, F., Junior, J. I., and Almeida, J. (2004). Quantitative evaluation of unstructured peer-to-peer architectures. In *Proc. of IEEE First International WorkShop on Hot Topics in Peer-to-Peer Systems (Hot-P2P'04)*, Volendam, The Netherlands.

[Benevenuto et al. 2005] Benevenuto, F., Júnior, J. I., and Almeida, J. (2005). Avaliação de mecanismos avançados de recuperação de conteúdo em sistemas p2p. In *Anais do 23 Simpósio Brasileiro de Redes de Computadores, SBRC2005*, Fortaleza, Brasil.

[BitTorrent] BitTorrent. http://bitconjurer.org/bittorrent/.

[Castro et al. 2002] Castro, M., Druschel, P., Hu, Y., and Rowstron, A. (2002). Exploiting Network Proximity in Distributed Hash Tables. In *Proc. International Workshop on Future Directions in Distributed Computing*, Bertinoro, Italy.

[Christin et al. 2005] Christin, N., Weigend, A. S., and Chuang, J. (2005). Content availability, pollution and poisoning in file sharing peer-to-peer networks. In *Proc. of ACM E-Commerce Conference*, Vancouver, Canada.

[eDonkey] eDonkey. http://www.edonkey2000.com/.

[Fasttrack] Fasttrack. http://www.fasttrack.com.

[Gnutella] Gnutella. http://www.gnutella.com.

[KaZaA] KaZaA. http://www.kazaa.com.

[Kurose and Ross 2005] Kurose, J. and Ross, K. (2005). *Computer Networking: A Top-Down Approach Featuring the Internet*. Addison-Wesley.

[Liang et al. 2005] Liang, J., Kumar, R., Xi, Y., and Ross, K. W. (2005). Pollution in p2p file sharing systems. In *Proc. of IEEE Infocom*, Miami, FL, USA.

[Napster] Napster. http://www.napster.com.

[Overpeer] Overpeer. http://www.overpeer.com.

[Pouwelse et al. 2005] Pouwelse, J., Garbacki, P., Epema, D., and Sips, H. (2005). The bittorrent p2p file-sharing system: Measurements and analysisng. In *Proc. of IPTPS*, Ithaca, NY, USA.

[Ratnasamy et al. 2001] Ratnasamy, S., Francis, P., Handley, M., Karp, R., and Shenker, S. (2001). A Scalable Content-Addressable Network. In *Proc. ACM SIGCOM*, San Diego, CA, USA.

[Retspan] Retspan. http://www.retspan.info.

[Saroiu et al. 2004] Saroiu, S., Gribble, S. D., and Levy, H. M. (2004). Measurement and analysis of spyware in a university environment. In *Proc. of the 1st Symposium on Networked Systems Design and Implementation (NSDI)*, San Francisco, CA.

[Saroiu et al. 2002] Saroiu, S., Gummadi, P., and Gribble, S. (2002). A Measurement Study of Peer-to-Peer File Sharing Systems. In *Proc. Multimedia Computing and Networking 2002 (MMCN '02)*, San Jose, CA, USA.

[Stoica et al. 2001] Stoica, I., Morris, R., Karger, D., Kaashoek, M., and Balakrishnan, H. (2001). Chord: A Scalable Peer-to-peer Lookup Service for Internet. In *Proc. ACM SIGCOMM*, San Diego, CA, USA.

[tool for FastTrack network] tool for FastTrack network, S. http://www.geocities.com/vlaibb/tools.html.

[UUHash] UUHash. http://en.wikipedia.org/wiki/UUHash.

[Viralg] Viralg. http://www.viralg.com.

[Walsh and Sirer 2005] Walsh, K. and Sirer, E. G. (2005). Fighting peer-to-peer spam and decoys with object reputation. In *Proc. of the Third Workshop on the Economics of Peer-to-Peer Systems (p2pecon)*, Philadelphia, PA, USA.

[Yu et al. 2005] Yu, W., Boyer, C., Chellappan, S., and Xuan, D. (2005). Peer-to-peer System-based Active Worm Attacks: Modeling and Analysis. In *Proc. of IEEE International Conference on Communications (ICC 2005)*, Seoul, Korea.