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, Peer-to-Peer (P2P) file sharing 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). The 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 um alto nível de incentivo 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 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 fonográfica. Estas indústrias experimentaram perdas de milhões de dólares em vendas de CDs e DVDs devido à distribuição não autorizada de conteúdo nas redes P2P. Desde a primeira ação legal contra o Napster [Kurose and Ross 2005], existe uma verdadeira guerra da indústria fonográfica contra

a pirataria e, conseqüentemente, contra os sistemas P2P. Com o surgimento de redes P2P descentralizados como o Gnutella [Gnutella], KaZaA [KaZaA], eDonkey [eDonkey] e BitTorrent [BitTorrent], a indústria fonográfica tornou-se incapaz de processar os sistemas P2P e tentou, sem muito sucesso, processar alguns usuários destes sistemas.

Entretanto, um estudo recente [Liang et al. 2005] mostrou evidências de que a intervenção da indústria fonográfica através da poluição de arquivos populares tem obtido um sucesso significativo. Poluição consiste na disseminação de cópias de um arquivo específico (uma música ou filme), com o mesmo metadado (ex.: nome, artista) de um arquivo não poluído, mas com conteúdo corrompido [Christin et al. 2005]. Utilizando esta estratégia, as companhias de música visam tornar mais difícil para os usuários encontrarem uma cópia não corrompida do arquivo procurado.

Existem duas principais técnicas para disseminar conteúdo poluído em sistemas P2P. A mais conhecida é chamada de *inserção de versões falsas*, que consiste na disseminação de versões corrompidas de um determinado arquivo na rede. 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 não corrompido. Esta técnica é chamada de *corrupção por chave*.

A poluição em redes P2P gera um grande desperdício de banda na Internet. Recentemente, [Liang et al. 2005] mostrou que mais de 50% dos arquivos populares encontrados através de buscas na rede FastTrack [Fasttrack] são poluídos. Como já foi dito anteriormente, os sistemas P2P são responsáveis pela maior parte do tráfego na Internet, o que torna o tráfego de conteúdo poluído um assunto que deve ser estudado. A poluição também faz com que usuários obtenham novamente outras versões na tentativa de encontrar um arquivo não poluído, o que contribui ainda mais para o aumento do tráfego.

Cada um dos vários tipos de sistemas P2P utilizados hoje em dia possui um determinado nível de poluição dos arquivos. Neste trabalho foi avaliado o espalhamento da 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. Além disso, foi avaliado qual o efeito no espalhamento da poluição na rede quando os usuários do sistema recebem incentivos para apagarem seus arquivos poluídos. A grande maioria dos usuários não apaga seus arquivos poluídos. Isso indica que o aumento do conteúdo poluído em algumas redes é causado pela negligência dos próprios usuários, que não apagam arquivos poluídos após o recebimento e os compartilham. Este trabalho avalia também a eficiência dos incentivos para que usuários apaguem seus arquivos poluídos para evitar a disseminação da poluição.

As demais seções deste trabalho estão assim organizadas. A seção 2 apresenta os 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, 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 apresenta as 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 obter arquivos na rede FastTrack. Com base nos dados coletados, eles mostram que o nível da poluição nessa rede é muito alto, principalmente para arquivos recentes e populares. [Christin et al. 2005] avalia a disponibilidade de conteúdo e o conteúdo corrompido em três sistemas P2P: KaZaA, eDonkey e Gnutella. Pouwelse et al. [Pouwelse et al. 2005] analisa a integridade do sistema BitTorrent/Suprnova ao inserir conteúdo poluído no sistema e mostram que 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 fonográfica para espalhar poluição. Além disso, eles consideram apenas inserção de versões falsas como método utilizado.

Outras ameaças a sistemas P2P são os *freeriders*, vermes e *spyware*. *Free riding* já foi estudado em diversos trabalhos e várias soluções foram propostas para incentivar o compartilhamento pelos usuários dos sistemas P2P [Adar and Huberman 2000, Golle et al. 2001, Bretzke and Vassileva 2003] . 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 a 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 a disseminação da 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.

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

Esta subseção apresenta uma descrição superficial das principais características das redes moderadas e não moderadas que são úteis para o entendimento deste trabalho. [Benevenuto et al. 2004] e [Pouwelse et al. 2005], descrevem redes moderadas e não moderadas com maior detalhamento.

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. O segundo grupo são os sistemas descentralizados estruturados, tais como Chord [Stoica et al. 2001], Pastry [Castro et al. 2002], e CAN [Ratnasamy et al. 2001], onde 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 em sistemas descentralizados não estruturados, tais como Gnutella e KaZaA.

Em particular, neste último grupo estão as arquiteturas hierárquicas que se destacam em termos de popularidade de usuários. A maior parte das aplicações de P2P não moderadas utiliza 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 a mensagem de consulta, 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 deseja obter. A descoberta de novos super-nós se dá através de mensagens de manutenção da rede enviadas periodicamente [Benevenuto et al. 2005].

3.1.1. Redes Moderadas

Dentre as redes moderadas existentes hoje em dia, as mais conhecidas são a do BitTorrent e a do eDonkey. O mecanismo utilizado pelo BitTorrent consiste na publicação de informações sobre o arquivo alvo. Os usuários que decidem obter este arquivo receberão pedaços uns dos outros, seguindo uma política de "olho por olho, dente por dente"(*tit-for-tat*), o que implica que os nós só irão ceder recursos para outros nós que têm tendência a cooperar. O moderador consiste na pessoa que disponibiliza o metadado do arquivo (ex: nome, autor) que contém informações suficientes para um usuário obter o arquivo. Este moderador pode apagar um metadado disponibilizado caso perceba que aquele arquivo está poluído.

A rede do eDonkey possui super-nós nos quais os nós se conectam para participar da rede. Estes super-nós geralmente não possuem arquivos compartilhados, mas possuem informações dos arquivos dos nós conectados a eles e, com isso, controlam as buscas dos nós. Várias aplicações que se conectam à rede do eDonkey permitem que o usuário digite o identificador do arquivo desejado ao invés de realizar uma busca pelo arquivo na rede. A partir deste mecanismo, surgiram alguns sítios na Internet que disponibilizam listas de arquivos não poluídos e suas respectivas chaves. Assim, o usuário procura um arquivo nesses sítios e obtém o identificador do arquivo. De posse dessa chave a aplicação eDonkey se conecta ao servidor central e, a partir dela, obtém a lista de usuários que possuem o arquivo. 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 disseminação da poluição utilizada nos sistemas P2P de compartilhamento de arquivos. Esta técnica consiste na disseminação na rede de versões poluídas de um arquivo com o intuito de dificultar a localização de uma versão não poluída do arquivo na rede. Os arquivos poluídos inseridos

na rede contém os mesmos metadados dos arquivos não poluídos. 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 obtenha uma cópia poluída e não a 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. Algumas empresas como Viralg [Viralg], RetSpan [Retspan] e OverPeer [Overpeer] oferecem serviços para eliminar qualquer conteúdo não autorizado distribuído em redes P2P. Para isso, 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 obtido de múltiplas fontes simultaneamente. Este identificador é gerado aplicando-se uma função de *hash* no conteúdo do arquivo. Cada sistema P2P utiliza um algoritmo diferente para gerar esse 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 um identificado com um ID distinto e com um certo número de cópias. Feito isso, o usuário escolhe uma versão e obtém pedaços de diferentes cópias recebidas de diferentes usuários. Se um pedaço recebido 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 necessariamente 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, 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 alterar os clientes P2P de forma que estes não computassem 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 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 do simulador implementado e a seção 4.2 apresenta as métricas e parâmetros.

4.1. Simulador de Redes P2P

Para avaliar a disseminação de conteúdo poluído, foi desenvolvido um simulador com duas partes: uma para redes moderadas e outra para redes não moderadas. Ambas partes do simulador compartilham as características descritas a seguir.

O simulador avalia a disseminação da 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 estão ativos. O número total de nós na rede é de 5.000 (2.500 nós ativos). No início da simulação, cada nó compartilha um único arquivo. O número total de arquivos distintos 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 obtém um arquivo a uma taxa de 1 arquivo a cada 5.000 unidades de simulação (que segue uma distribuição exponencial). Ao requisitar um arquivo, um nó ativo escolhe um dos 20 arquivos existentes no sistema. Em seguida, 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 dessas versões para obter 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 deste trabalho, a modelagem da topologia da rede sobreposta (*overlay*) é desnecessária. Sendo assim, foi assumido que os nós sempre encontram todas as versões e cópias do arquivo procurado. Então, o nó seleciona de quais fontes ele deseja obter o arquivo e este é recebido instantaneamente. Por simplicidade, não foi modelado o tempo de transferência dos arquivos.

Foi assumido que quando um nó recebe um arquivo e este não é imediatamente apagado, ele é compartilhado, ou seja, o arquivo fica disponível. Além disso, os nós podem obter mais de um arquivo diferente e repetir a mesma busca, caso já não tenham este arquivo ou o tenham apagado. O intervalo de tempo em que cada nó fica ativo e inativo seguem uma distribuição exponencial com média de 10.000 passos de simulação.

4.2. Métricas e Parâmetros

Com o intuito de avaliar a disseminação da poluição, foi definida a métrica *disseminação de conteúdo poluído*, que representa 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 de rede. Os parâmetros escolhidos são os que têm maior impacto na propagação da poluição na rede. Os parâmetros para as redes moderadas foram:

- **Incentivo para apagar (**IA**)**: corresponde à probabilidade de um usuário apagar o arquivo poluído imediatamente após o término de seu recebimento. 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 neste trabalho 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 IDs de arquivos foi definido VH, como a porcentagem do arquivo que pode ser corrompido sem alterar seu ID. Por exemplo, se o algoritmo uuhash for utilizado para gerar os IDs de arquivos de 5 MB e 600 MB, os VHs 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 obtido. 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.

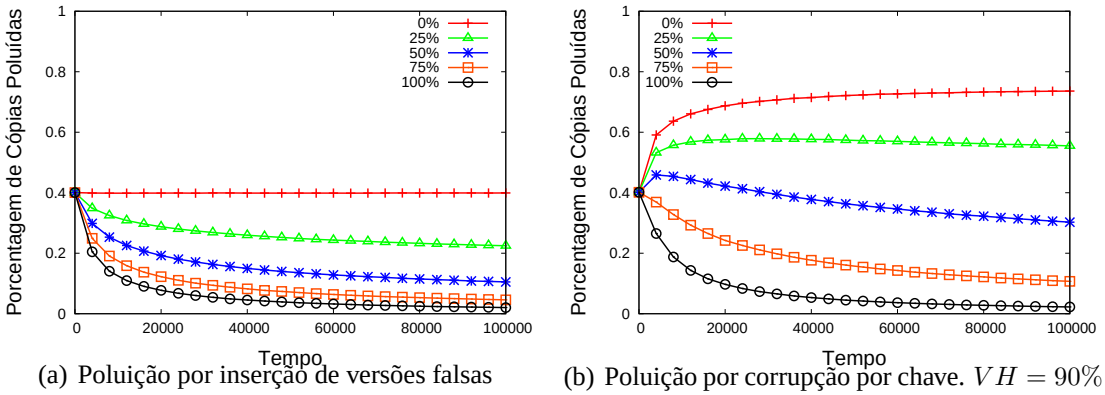
Para a avaliação de disseminação de conteúdo poluído em redes moderadas foram utilizados os seguintes parâmetros:

- **Incentivo para reportar (IR)**: Em redes moderadas, a poluição pode ser evitada quando os usuários comunicam ao moderador que o arquivo recebido 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(TRM)**: 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 o 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, TRM, como sendo o tempo que o moderador do sistema demora 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 por unidades de tempo de simulação. 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 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 possuíam 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 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 disseminação de cópias poluídas utilizando a estratégia de inserção de versões falsas é apresentada na figura 1(a). Cada curva mostra o número de cópias poluídas para



(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. $NF = 10$

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 ($IA = 0$), a porcentagem de cópias corrompidas na rede se mantém. 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 de 25% ($IA = 0,25$), a porcentagem de cópias poluídas na rede decresce de 44% 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 recebe o 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, para alguns valores de incentivo, inicialmente o número de arquivos poluídos aumenta com o tempo se a poluição for disseminada pelo mecanismo de corrupção por chave. Isso ocorre porque o *download* é feito de 10 fontes diferentes e com isso, a probabilidade de se obter um pedaço corrompido aumenta. Somente quando os usuários recebem um incentivo maior para apagar ($IA > 0,5$) é que a disseminação de conteúdo poluído começa a diminuir.

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 10 fontes de *download* fornecer uma cópia poluída e o usuário receber 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 da poluição na rede diminui 44% ao final da

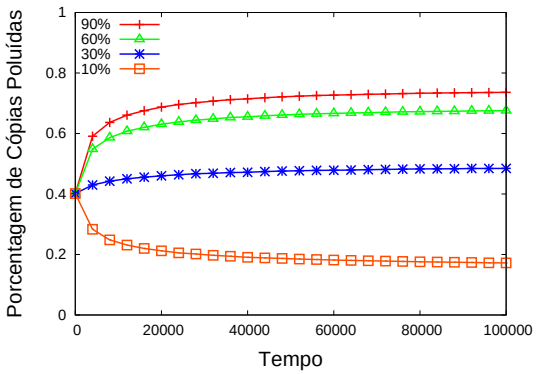


Figura 2. Disseminação de cópias poluídas pelo mecanismo de corrupção por chave variando a vulnerabilidade do *hash* (VH). $IA = 0$ e $NF = 10$

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 os 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 *downloads* 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% do arquivo poluído ($VH = 0,3$), a porcentagem de cópias poluídas aumenta 20% ao final da simulação. Nota-se que aumentando a vunerabildade do *hash* de 10% para 30% o nível da poluição aumenta cerca de 150% no fim da simulação.

A figura 3 mostra como a disseminação da poluição, com a técnica de poluição por corrupção de chave, é afetada com o aumento do número de fontes (NF). Como esperado, quando maior o número de fontes das quais o arquivo é obtido, maior se torna a probabilidade de se receber um pedaço de uma fonte poluída. Porém, é necessario muito poucas fontes para que a corrupção por chave alcance altos níveis de poluição. Ao fim da simulação, a diferença entre 5 fontes e 100 é somente 15%.

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

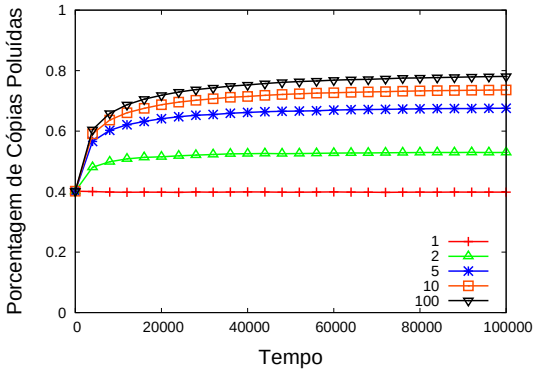
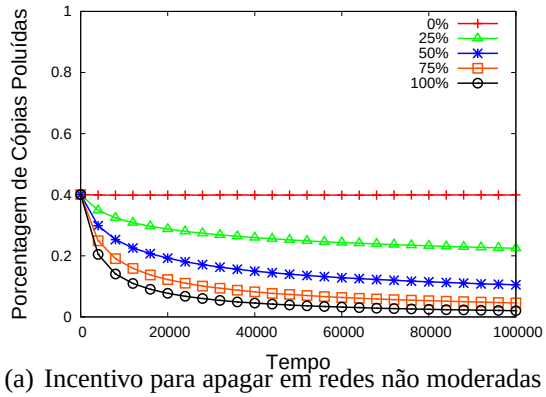
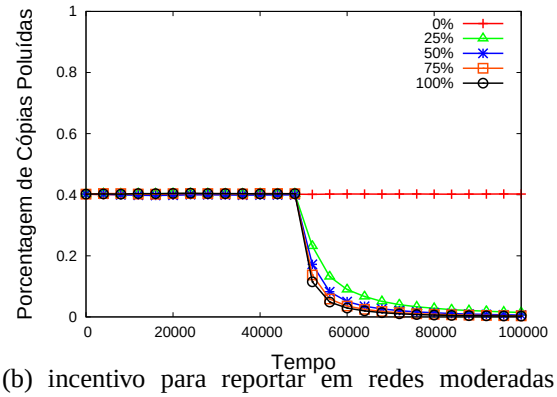


Figura 3. Disseminação de cópias poluídas pelo mecanismo de corrupção por chave variando o número de fontes de download simultâneas (NF). $IA = 0$ e $VH = 90\%$



(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

50.000 unidades de simulação para o 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 recebem algum conteúdo poluído reportam), o moderador consegue excluir da rede todos os arquivos poluídos ate o final da simulação (Figura 4-b). Pode-se notar que o incentivo para 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 centralizadora, o moderador, que tem o poder de apagar todas as cópias de uma determinada versão.

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 recebem 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.

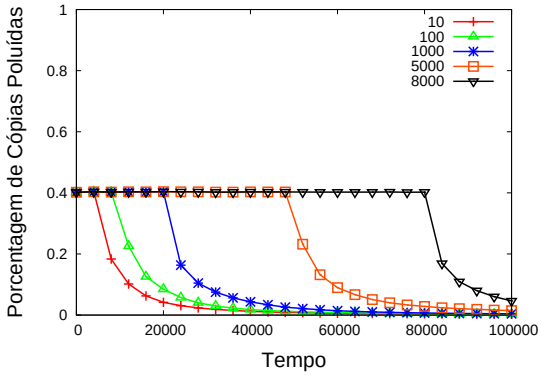


Figura 5. Disseminação de cópias poluídas em redes moderadas variando o tempo que o moderador leva para deletar (intervalos fixos e incentivo para reportar $= 0,25$)

6. Mecanismos de Combate à Poluição

Neste trabalho mostramos que a poluição pode ser bastante prejudicial para um 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 ao 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 observar na seção 5, o 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, é imediatamente apagado. Dentre os projetos que implementam esta idéia destaca-se o projeto Sig2dat [tool for FastTrack network], que disponibiliza uma ferramenta que calcula o identificador de arquivos da rede FastTrack. Desde seu surgimento, vários sítios vêm disponibilizando listas "negras" ou "brancas" de arquivos e seus respectivos identificadores para que os usuários possam se orientar antes de obter um arquivo. Porém, para que esse mecanismo funcione, é necessário que o arquivo seja recebido por completo, 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 recebimento. Neste trabalho, mostramos que quando a maior parte dos usuários apaga 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. Para conter a poluição existe o Credence [Walsh and Sirer 2005], que é um mecanismo de reputação que classifica os arquivos distribuídos na rede. Sempre que um nó deseja obter um arquivo ele pergunta para outros nós confiáveis se o arquivo desejado está poluído ou não. A partir das respostas obtidas, o nó 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 nós 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.

O incentivo para apagar estudado neste trabalho é uma modelagem abstrata para o combate à poluição de forma distribuída. Logo, os nossos resultados mostram que se o nível da poluição cai lentamente para um dado incentivo para apagar, significa dificuldade para aplicar mecanismos distribuídos baseados em reputação. Isso provavelmente leva-

ria à escolha de mecanismos centralizados para o combate da poluição ou mecanismos distribuídos que faça com que os usuários alcancem altos níveis de incentivo.

7. Conclusão e Trabalhos Futuros

Este trabalho analisa a disseminação de conteúdo poluído 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 dissemina a poluição mais rapidamente do que a poluição por inserção de versões falsas. Mesmo quando usuários possuem um alto nível de incentivo para apagar o conteúdo poluído, o mecanismo de corrupção por chave se mostra eficiente para disseminar a poluição. Comparando redes moderadas com redes não moderadas, os resultados mostraram que a presença de um moderador na rede pode ser eficaz para conter o espalhamento da poluição. Entretanto, a participação do moderador 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 poluído em sistemas P2P. Mais especificamente, pretendemos criar um novo mecanismo de reputação para combater todos os tipos de poluição. Pretendemos também comparar este novo mecanismo com outros mecanismos de reputação existentes, como o Credence [Walsh and Sirer 2005].

8. Agradecimentos

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

Referências

[Adar and Huberman 2000] Adar, E. and Huberman, B. A. (2000). Free riding on gnutella. *First Monday*.

[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/.

[Bretzke and Vassileva 2003] Bretzke, H. and Vassileva, J. (2003). Motivating cooperation on peer to peer networks. *User Modeling*, pages 218–227.

[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.

[Golle et al. 2001] Golle, P., Leyton-Brown, K., Mironov, I., and Lillibridge, M. (2001). Incentives for sharing in peer-to-peer networks. *Lecture Notes in Computer Science*, 2232:75+.

[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.