Uma Arquitetura para Controle de Privacidade na Web

Thesis · December 2003

CITATIONS
4

READS
23

5 authors, including:

Lucila Ishitani
Pontifícia Universidade Católica de Minas Gerais
43 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Wagner Meira Jr.
Federal University of Minas Gerais
399 PUBLICATIONS **3,867** CITATIONS
SEE PROFILE

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Uma Arquitetura para Controle de Privacidade na Web

Lucila Ishitani

Prof. Virgílio Augusto F. Almeida (Orientador)
Prof. Wagner Meira Júnior (Co-orientador)

Tese submetida ao Colegiado do Curso de Pós-Graduação em Ciência da Computação da UFMG, como um requisito parcial para a obtenção do grau de Doutora em Ciência da Computação.

Dezembro de 2003

Aos meus pais, Shigueki e Haruê,
ao meu marido e grande companheiro, César, e aos meus preciosos filhos, Nádia, Daniel e Elisa.

Agradecimentos

À Deus, pela minha vida e pelo seu amor;

Ao Prof. Virgílio e ao Prof. Wagner, pela paciência e pelas brilhantes idéias que foram determinantes para a concretização deste trabalho;

Aos meus pais, Shigueki e Haruê, pelo amor e carinho e, principalmente, por fazerem de mim uma pessoa apta a esta conquista;

Ao meu marido, César, e aos meus filhos, Nádia, Daniel e Elisa, pela fonte inesgotável de amor, alegria e energia;

À Profa. Clarisse, pelas suas valiosas sugestões;

Aos membros da banca examinadora, pelas suas contribuições;

Aos meus amigos e colegas do DCC-UFMG e da PUC Minas, pelo apoio e pela torcida;

Aos professores e funcionários do DCC-UFMG, por terem me auxiliado a tornar possível este trabalho.

Resumo

Esta tese propõe uma arquitetura que permite ampliar o controle do usuário sobre o ambiente computacional, no que se refere à privacidade. A privacidade na Web é uma questão que tem levantado, atualmente, várias discussões. Primeiramente, porque muitos não sabem como uma invasão de privacidade pode ocorrer ou o que se deve fazer para proteger sua privacidade. Na verdade, nem mesmo o conceito de privacidade está claro, pois há uma sobreposição dos conceitos de privacidade e segurança que necessita ser esclarecido. Um outro ponto a ser discutido é o valor da privacidade para cada indivíduo e o quanto vale a pena abrir mão de um pouco desta, para poder usufruir de serviços variados na Web. Essa discussão ocorre, por exemplo, no conflito entre personalização e privacidade: por um lado, os usuários apreciam a idéia de receber serviços personalizados e não aprovam o fato de que suas ações estejam sendo gravadas, acompanhadas e analisadas; por outro lado, esse tipo de informação é fundamental para que possa haver personalização de serviços. A arquitetura proposta nesta tese dá ao usuário da Web melhores condições para compreensão dos seus riscos, no que concerne à privacidade e, simultaneamente, lhe oferece recursos para proteção de sua privacidade, através da anonimidade, sem lhe tirar o direito de ter acesso a serviços personalizados. Compõe também este trabalho uma visão conceitual de privacidade na Web: conceito e importância de privacidade; distinção entre privacidade e segurança; proteção de privacidade na Web.

Abstract

This thesis proposes an architecture that allows users to enhance their privacy control over the computational environment. Web privacy is a topic that is raising, nowadays, many discussions. Firstly because many people do not know how their privacy can be violated or what can be done to protect it. In general, people do not even know what privacy means, and there is an overlap of the concepts of privacy and security, that needs to be cleared. Another topic to be discussed is the value that each one gives to privacy and when it is worth to give up some privacy in order to profit from several different Web services. The value of privacy has generated many conflicts. Among them, we would like to show up the one that happens between privacy and personalization: by one side, users appreciate the idea of receiving personalized services and do not approve the collection, tracing and analysis of their actions; by the other side, personalization services need this type of information in order to profile their users. The architecture proposed in this thesis helps users to understand better how their privacy can be invaded and, at the same time, gives them a better control of their privacy, through anonymity, without preventing them from receiving personalized services. This thesis also includes a conceptual vision of Web privacy: concepts and importance of privacy; a distinction between privacy and security; Web privacy protection.

Conteúdo

Capítulo 1

Introdução

Apesar da popularização da Internet, muitas pessoas ainda evitam usufruir plenamente de seus serviços, por recearem ter sua privacidade invadida. Na verdade, esse receio é justificável, pois os avanços tecnológicos permitem que informações sobre os usuários sejam coletadas, armazenadas, monitoradas, analisadas e divulgadas com muita facilidade.

Para se ter uma idéia do crescimento do volume de dados armazenados, segundo um trabalho realizado por *Sweeney* [55], em 1983 havia, aproximadamente, 0,02 MB armazenados no mundo por pessoa e em 2000, 474 MB, por pessoa. No contexto da Internet, o volume de dados gravados é tão grande que, segundo estimativas, somente 7% deste consegue ser utilizado pelas empresas [27].

Um agravante para essa situação reside no fato de que muitos usuários não sabem que seus dados estão sendo coletados, ou, se o sabem, não têm idéia da quantidade coletada nem tampouco do objetivo desta. Dessa forma, cresce, na sociedade, a preocupação com relação à perda de privacidade.

Há tentativas de se buscar uma solução para o problema, através da disponibilização de ferramentas para proteção de privacidade de usuários e da divulgação da política de privacidade adotada por *site*. Entretanto, tanto quanto sabemos, nenhuma das soluções propostas obteve resultados satisfatórios. Segundo uma pesquisa conduzida por *Pew Internet & American Life Project* [23], a maior parte dos usuários da Web nunca usou alguma das ferramentas existentes para proteção de sua privacidade. Quanto à política de privacidade, muitas vezes ela é expressa através do uso de jargões que dificultam o entendimento por grande parte de usuários.

Pesquisas demonstram que usuários aceitam que seus dados sejam coletados e até mesmo estão dispostos a fornecer informações pessoais, se estas forem utilizadas em seu benefício, como é o caso de serviços personalizados [2].

Os serviços personalizados incluem a adaptação do conteúdo das páginas ao comportamento do usuário e aos seus interesses atuais. A personalização traz benefícios para ambos os lados de uma interação da Web: usuários e *sites*. Entretanto, o processo de coleta e análise de dados dos usuários, necessário para que a personalização possa ocorrer, pode caracterizar invasão de privacidade. A privacidade informacional pode ser caracterizada como o direito que os indivíduos têm de proteger sua capacidade de revelar, seletivamente, suas informações pessoais [48]. No caso da Web, o fato de muitas pessoas não saberem, ao certo, o quê ou quanto ou para quê seus dados são coletados deixa claro que a infra-estrutura atual da Web representa um sério risco à privacidade dos usuários de seus serviços.

Dessa forma, nos encontramos à frente do seguinte conflito: como disponibilizar dados para *sites* da Web, de forma que serviços personalizados possam ser oferecidos para os usuários, sem que ocorra invasão de sua privacidade? Há, basicamente, dois grupos de soluções propostas para este conflito: o primeiro se baseia na idéia do próprio usuário escolher quais dados deseja disponibilizar; o segundo, na disponibilização de dados agrupados de vários usuários, de forma que não seja possível associar dados a um indivíduo específico, protegendo, assim, a privacidade dos indivíduos que compõem o grupo. Neste trabalho, avaliamos o problema sob diversos ângulos e propomos uma solução para este conflito: a arquitetura MASKS (**M**anaging **A**nonymity while **S**haring **K**nowledge to **S**ervers) [28]. MASKS se baseia na idéia de minimizar a exposição do usuário através do uso de uma "barreira de anonimidade" que filtre as informações que fluem entre os usuários e os *sites* da Web. Essas "informações filtradas" protegerão a privacidade dos usuários, sem impedir que serviços personalizados possam ser oferecidos a eles.

Este trabalho tem dois objetivos principais. O primeiro é analisar vários aspectos relacionados à privacidade na Web, o que inclui: o conceito de privacidade, a apresentação de métodos e técnicas para proteção e invasão de privacidade, uma proposta de classificação de camadas de proteção de privacidade e a distinção entre segurança e proteção de privacidade. O segundo objetivo é propor uma arquitetura que aumente o controle do usuário da

Web sobre a sua privacidade e que lhe permita equilibrar os desejos contraditórios de ter privacidade protegida e, ao mesmo tempo, poder ter acesso a serviços personalizados. A estratégia adotada permite a divulgação de informações para *sites*, para que o serviço de personalização possa ocorrer, sem que seja possível identificar a quem essas informações pertencem. A sua concretização se baseia em uma solução já amplamente conhecida: anonimidade.

Esta tese está organizada da seguinte forma: o capítulo 2 apresenta uma discussão sobre o conceito de privacidade, a importância de se proteger a privacidade das pessoas, o estado atual das leis e regulamentações relacionadas à proteção de privacidade na Web e as formas de invasão de privacidade na Web.

O capítulo 3 aborda as estratégias que podem ser utilizadas pelas pessoas para proteção de sua privacidade no mundo real e no mundo virtual. O capítulo também apresenta uma taxonomia para camadas de proteção de privacidade. E, por fim, traz uma diferenciação entre privacidade e segurança, termos estes que muitas vezes são utilizados como sinônimos, mas que demonstramos que podem ser dissociados.

O capítulo 4 apresenta o MASKS (**M**anaging **A**nonymity while **S**haring **K**nowledge to **S**ervers) - a arquitetura que estamos propondo como solução para os interesses conflitantes dos usuários: proteção de privacidade recebendo, simultaneamente, serviços personalizados.

O capítulo 5 descreve, com mais detalhes, um dos componentes do MASKS: o **P***rivacy and* **S***ecurity* **A***gent* (PSA). O PSA inclui: a interface com o usuário; o processamento de requisições dos usuários, antes de serem enviadas aos *sites* da Web; o processamento das respostas que chegam aos navegadores, antes de serem apresentadas aos usuários.

O capítulo 6 apresenta as características principais do componente do MASKS responsável pelo processo de anonimização: o *Masks Server*. Dentre as características abordadas, destacam-se o algoritmo de anonimização e as estratégias adotadas para que a arquitetura seja mais segura.

O capítulo 7 aborda a análise dos resultados de avaliações qualitativas e quantitativas do MASKS e as metodologias utilizadas para obtenção destes resultados.

O capítulo 8 apresenta as conclusões deste trabalho e trabalhos futuros a realizar.

Capítulo 2

Privacidade

2.1 Conceito de privacidade

Privacidade é um conceito abstrato cujo valor e extensão variam de pessoa para pessoa. Podemos comparar a visão que cada pessoa tem de sua privacidade a uma bolha que a envolve. Essa bolha que cada um determina como sendo o seu limite de privacidade terá tamanhos diferenciados para cada pessoa. O que uma pessoa considera invasão de privacidade, outra pessoa pode considerar como algo completamente normal e aceitável. *Elgesem* [20] recomenda abandonar a dicotomia rígida entre o que é privado e o que é público, pois, em geral, as situações privadas ocorrem dentro de um escopo maior que são as situações públicas. Por exemplo, uma mulher conversando de forma privada em um telefone público poderá ser vista por todas as pessoas que passarem por ela. Para *Elgesem*, a privacidade está fortemente conectada com a idéia de que existem algumas coisas que outras pessoas não deveriam ver ou saber.

Um conceito de privacidade amplamente difundido é o discutido por *Warren & Brandeis* no famoso artigo *The Right to Privacy* da *Harvard Law Review*, de 1890 [61]: "privacidade é o direito de estar sozinho". Nesse mesmo artigo, encontramos também a seguinte regra: "O direito à privacidade termina com a divulgação de fatos pelo indíviduo ou com o seu consentimento". A partir dessa regra, identificamos um cuidado que cada um deve ter em proteger sua privacidade pois, uma vez que alguém divulgue ou autorize a divulgação de um fato ou informação pessoal, não há como voltar atrás.

Com os avanços tecnológicos, as pessoas têm rapidamente perdido a sua privacidade e

essa situação tende a se agravar: filmagens em lojas e estacionamentos, eletrodomésticos conectados à Internet, bancos de dados armazenando grande volume de dados pessoais e ações (compras efetuadas, ligações telefônicas, depósitos e retiradas de contas bancárias, etc).

Nesse novo contexto, surge um novo sentido de privacidade, que a coloca como uma propriedade - a propriedade de ter controle sobre o seu fluxo de informação pessoal [20]. *Fried* [24] afirma que "privacidade não é simplesmente a ausência de informação a nosso respeito na cabeça de outros, mas também, o controle que temos sobre estas informações".

No contexto da Web, privacidade se refere à privacidade de informações [11]. Uma definição reconhecida para privacidade de informações é a apresentada por *Alan Westin* [62], em 1987, como sendo "a reinvindicação de indivíduos, grupos ou instituições de poderem determinar quando, como e quanto de suas informações podem ser divulgadas a outros". Um outro conceito que merece destaque é o proposto por *Wang et al.* [60], que nos coloca que "privacidade geralmente se refere a informações pessoais, e invasão de privacidade é geralmente interpretada como coleta, publicação ou outro uso não-autorizado de informações pessoais, como um resultado direto de transações". Chamamos a atenção para o fato de que quem determina se uma informação pode ser divulgada ou não é o usuário, através de autorização ou consentimento. Essa relação existente entre privacidade e consentimento do usuário pode gerar um outro problema resultante das desigualdades sociais: pessoas pobres tenderão a divulgar suas informações com mais freqüência do que as ricas, sempre que estiver em jogo um incentivo financeiro, como descontos ou até mesmo pagamento pelos seus dados. É importante ressaltar também que, anteriores ao problema da divulgação de informações pessoais, existem os problemas da coleta, análise e atualização de dados.

2.2 Privacidade: importância e conflitos

Segundo os resultados da pesquisa sobre usuários da Web, realizada pelo *GVU Center* [29], 88,1% dos usuários acham interessante a idéia de visitar um *site* anonimamente, 77,5% consideram que a privacidade é mais importante do que a conveniência, 71,4% pensam que as leis atuais não são suficientes para proteger a privacidade.

Neste momento, cabe aqui o seguinte questionamento: por que as pessoas consideram

a privacidade tão importante? Uma primeira justificativa seria o fato de que o grau de intimidade no relacionamento pessoal está relacionado com a quantidade e a qualidade da informação compartilhada com outros [20]. A disseminação de nossas informações pessoais, sem o nosso controle, acaba por tirar o nosso controle de relações pessoais.

Uma segunda justificativa seria o fato de que as pessoas devem se sentir à vontade para realizar seus projetos de vida: viajar quando, para onde e com quem quiserem; ler, falar e comprar o que quiserem; pensar, explorar novas idéias e agir da forma que quiserem, dentro dos limites da lei [57]. A privacidade ajuda as pessoas a manterem sua autonomia e individualidade. Se todas as ações forem monitoradas, a capacidade de formular novas idéias e opiniões pode ficar seriamente restrita.

Benn [7] defende a idéia de que a noção de respeito pelas pessoas é essencial para uma perfeita compreensão do valor da privacidade. Proteger a privacidade de alguém é proteger sua capacidade de desenvolver e realizar seus projetos da forma que quiser, por respeitar a forma de pensar do outro. E isto é essencial para o funcionamento de uma sociedade saudável.

Entretanto, na sociedade moderna, a todo momento temos que disponibilizar informações pessoais para várias instituições diferentes. Essa disponibilização de informação tem um custo: um aumento no risco de ter sua privacidade invadida. Se o custo for pequeno, é claro que optamos por reduzir um pouco da nossa privacidade protegida, para podermos obter o que queremos. Mas, com o advento da tecnologia, a cada dia esse custo tem aumentado. De acordo com *Elgesem* [20], "na vida real, temos que aceitar algum nível de risco: é impossível reduzir o risco a zero, e há também um limite no preço que é razoável pagar para se reduzir esse risco". Esse preço varia de pessoa para pessoa e de situação para situação.

Um exemplo prático desse conflito é a personalização. Personalizar é adaptar o serviço oferecido a um cliente, de acordo com suas necessidades e preferências. No contexto da Web, serviços personalizados incluem a adaptação do conteúdo das páginas ao comportamento do usuário e aos seus interesses atuais. A personalização traz benefícios para ambos os lados de uma interação da Web: usuários e *sites*.

Kobsa [39] afirma que "clientes necessitam sentir que possuem um relacionamento pessoal e único com a empresa"e para confirmar essa idéia, ele apresenta o resultado de

uma pesquisa que mostra que *sites* que oferecem serviços personalizados conseguiram um aumento de 47% no número de novos clientes. Esse resultado também demonstra que a personalização traz benefícios para o *site*. O aumento no número de clientes tende a aumentar as vendas e, conseqüentemente, os lucros. Na verdade, já se sabe que *sites* que oferecem serviços personalizados conseguem, em relação aos *sites* não personalizados, uma taxa muito maior de conversão de visitantes para consumidores [58].

Entretanto, o processo de coleta e análise de dados dos usuários, necessário para que a personalização possa ocorrer, pode caracterizar invasão de privacidade. Por isso, o ideal seria que o usuário tivesse alguma forma de proteger sua privacidade sem ter que abrir mão de serviços personalizados.

2.3 Legislações e regulamentações

Os governos de vários países discutem leis que regulamentem a proteção de privacidade e punam aqueles que desrespeitem essas leis. A *Electronic Privacy Information Center*[1] (EPIC) e a *Privacy International*[2] elaboram, conjuntamente, um relatório anual abordando as legislações e os avanços em vários países do mundo, no aspecto da proteção de privacidade de dados. O relatório do ano de 2003[3] inclui avaliações de 56 países. Segundo este relatório, a situação atual é a seguinte:

- Os países europeus e da Oceania se destacam pelo conjunto de ações em defesa da privacidade de dados. Na Europa, o *Council of Europe* aprovou a Convenção 108[4] (*Convention for the protection of individuals with regards to automatic processing of personal data*), em 1981, que protege os dados pessoais manipulados tanto pelo setor público, quanto pelo privado, limitando a coleta, armazenamento e transmissão destes dados.

- No continente africano, somente a África do Sul iniciou, em 2002, a elaboração de um projeto de lei em defesa da privacidade. Até junho de 2003, não havia, ainda, nenhum documento disponível para análise.

[1] http://www.epic.org
[2] http://www.privacyinternational.org
[3] Privacy and Human Rights 2003: An International Survey of Privacy Laws and Developments - http://www.privacyinternational.org/survey/phr2003
[4] http://www.conventions.coe.int/Treaty/en/Treaties/Html/108.htm

- No continente americano, destacam-se as ações do Canadá, Argentina e Chile.
- Na Ásia, destacam-se Israel, Japão, Hong Kong e Taiwan. Com exceção da Tailândia, Índia e Coréia do Sul, os demais países não possuem nenhum tipo de legislação para proteção de privacidade de dados e nem tampouco iniciaram um processo nessa direção.

País	Ações
Argentina	O artigo 43 da Constituição dá aos indivíduos o direito de saberem o conteúdo e o objetivo de todos os dados arquivados, a eles associados. Em novembro de 2000, foi aprovada a "Lei para Proteção de Dados Pessoais". A Argentina é o primeiro país da América Latina a obter a aprovação da União Européia, com relação à proteção de dados.
Brasil	O artigo 5 da Constituição de 1988 dá a todos os cidadãos o direito da privacidade. Em 1999, foi proposta uma lei que descreve os crimes de informação, que incluem a coleta, processamento e distribuição de informação.
Chile	O Chile foi o primeiro país latino-americano a aprovar uma lei de proteção de dados. Esta lei, de 1999, cobre os direitos das pessoas, quanto ao acesso, correção e controle de dados pessoais.
Peru	A Constituição de 1993 determina o direito à privacidade e à proteção de dados. Em 2002, o Ministro da Justiça criou uma comissão especial para escrever um novo documento que detalhe a proteção de dados. Contudo, não houve progressos nesta área.

Tabela 2.1: Ações para proteção de dados na América do Sul

Apresentamos, nas tabelas 2.1, 2.2, 2.3, 2.4 e 2.5 um resumo da situação de alguns países dos vários continentes do mundo, onde já foram implantadas ações para proteção da privacidade de dados.

Com o intuito de proteger a privacidade dos usuários da Web, surgiram propostas para regularizar a proteção de privacidade, das quais duas se destacam: a primeira é da *Organization for Economic Co-operation and Development*[5] (OECD) e a segunda, da *Federal Trade Commission*[6] (FTC).

O conjunto de princípios estabelecidos em 1980 pela OECD especificam de que forma os dados pessoais devem ser protegidos. Apresentamos, sucintamente, os oito princípios [9]:

[5]http://www.oecd.org
[6]http://www.ftc.gov

País	Ações
Canadá	A privacidade de seus cidadãos está protegida por dois decretos: o "Decreto Federal de Privacidade", de 1982, e o "Decreto de Informações Pessoais e Documentos Eletrônicos", de 2001. O decreto de 1982 regula a coleta, o uso e a divulgação de dados pessoais por órgãos do governo. O decreto de 2001 estabelece dez princípios que as organizações devem respeitar, com relação à coleta, o uso, a divulgação e o armazenamento de dados pessoais.
Estados Unidos da América	Na Constituição do país, não há nenhum direito explícito à privacidade. O "Decreto de Privacidade", de 1974, restringe a coleta, o uso e a disseminação de informações por agências federais. Não há leis que regulem a proteção da privacidade para o setor privado. Desde janeiro de 2001, já foram apresentados ao Congresso mais de duzentos documentos que tratam da proteção de privacidade.
México	Na Constituição do México, não é possível encontrar uma lei que trate diretamente de proteção de dados. Apesar de ser membro da OECD, ainda não adotou suas diretivas.

Tabela 2.2: Ações para proteção de dados na América do Norte e Central

1. Princípio do Limite de Coleta: deve haver limite à coleta de dados pessoais e, quando essa ocorrer, deve ser feita através de meios legais e, quando apropriada, com o conhecimento e o consentimento do "proprietário" dos dados.

2. Princípio da Qualidade dos Dados: dados pessoais devem ser relevantes para os objetivos onde serão utilizados e devem ser precisos, completos e mantidos atualizados.

3. Princípio da Especificação de Objetivo: os objetivos da coleta de dados devem ser especificados antes da coleta e o uso desses dados deve estar restrito a esses objetivos.

4. Princípio da Limitação de Uso: os dados pessoais não podem ser divulgados, disponibilizados ou usados para outros propósitos além dos especificados exceto: a) quando há consentimento do "proprietário"dos dados ou b) por uma autoridade da lei.

5. Princípio da Segurança (*Security Safeguards*): dados pessoais devem estar protegidos por mecanismos de segurança razoáveis.

6. Princípio da Transparência (*Openness*): deve haver uma política geral de divulgação sobre práticas e políticas com respeito a dados pessoais.

País	Ações
Israel	A "Lei de Proteção de Privacidade" regula o processamento de informações pessoais em bancos de dados, especificando um conjunto de atividades proibidas, relacionadas ao objetivo, uso e segurança de dados coletados.
Japão	Em 1988, foi aprovado o "Decreto para Proteção de Dados Pessoais Processados por Computador e Armazenados por Órgãos Administrativos". Esse decreto impõe regras para segurança, acesso e atualização de dados. Em 1998, o Ministério de Comércio Internacional e Indústria criou uma entidade para supervisionar empresas, com relação ao respeito e a proteção de dados pessoais dos consumidores.
Rússia	A "Lei sobre Informação, Informatização e Proteção de Informação" considera todo dado pessoal como informação confidencial e, por isso, proíbe coleta, armazenamento, uso e distribuição de dados de um indivíduo, sem sua autorização explícita. Entretanto, a lista de dados a serem protegidos deveria estar estipulada por uma lei federal que ainda não foi aprovada. Por isso, é comum observar, na Rússia, coleta e distribuição ilegal de dados.
Hong Kong	Em 1996, foi aprovada uma legislação sobre "Privacidade de Dados Pessoais", que regula a informação, coleta, uso, armazenamento e acesso a dados pessoais, de forma muito semelhante à Convenção 108 da Europa.
Taiwan	Em 1995, foi aprovada a "Lei para Proteção de Dados Pessoais Processados em Computador" que dá, às pessoas, o direito de acessar e corrigir seus dados e de determinar quando não querem que seu dados sejam coletados e processados.

Tabela 2.3: Ações para proteção de dados na Ásia

7. Princípio da Participação Individual: um indivíduo deve ter o direito de obter e pesquisar dados relativos a si mesmo.

8. Princípio da Responsabilidade: um controlador de dados deve ser responsável por cumprir todos os princípios acima.

As práticas de informação justas (*Fair information practices*) da FTC são praticamente um resumo dos oito princípios apresentados acima [21]:

1. Informação (*Notice*) - *sites* da Web devem informar aos usuários o que coletam, como e para quê, se terceiros têm acesso a informações coletadas e de que forma disponibilizam aos usuários os três serviços apresentados a seguir: escolha, acesso e segurança.

País	Ações
Alemanha	A Alemanha possui uma das leis de proteção de dados mais rigorosas da União Européia. A "Lei Federal de Proteção de Dados", cuja última revisão foi em 2002, cobre a coleta, processamento e uso de dados pessoais, coletados por órgãos públicos e privados. A "Comissão Federal de Proteção de Dados" é uma agência federal que cuida do cumprimento desta lei.
Espanha	O "Decreto Espanhol para Proteção de Dados", aprovado em 1992, cobre dados manipulados pelos setores público e privado. A lei estabelece que os cidadãos têm o direito de corrigir, apagar e saber quais dados a seu respeito estão armazenados. A "Agência de Proteção de Dados" foi criada para registrar e investigar casos de violação da lei. Em 2002, foi aprovada a "Lei de Serviços e Comércio Eletrônico da Sociedade de Informação" que, dentre outras punições, fecha sites envolvidos em atividades ilegais.
França	Em 1978, foi aprovado o "Decreto de Proteção de Dados", que cobre dados armazenados por agências públicas e privadas. Aquele que quiser processar dados de outros deve se registrar e obter permissão para isso, junto à "Comissão Nacional de Informática". Indivíduos têm o direito de acesso, atualização e remoção de dados pessoais.
Portugal	A Constituição portuguesa cobre extensivamente o direito à privacidade e à proteção de dados. Segundo a lei, todo cidadão tem o direito de saber quais são os dados armazenados a seu respeito e os objetivos da coleta. Em 1998, foi aprovado o "Decreto de Proteção de Dados Pessoais" que limita a coleta, o uso e a disseminação de informações pessoais. A fiscalização do cumprimento deste decreto está sob responsabilidade da "Comissão Nacional de Proteção de Dados".

Tabela 2.4: Ações para proteção de dados na Europa

2. Escolha (*Choice*) - *sites* da Web devem oferecer aos usuários a opção de escolher como suas informações pessoais podem ser utilizadas além dos objetivos para os quais foram fornecidas. Por exemplo, se permitem que as informações disponibilizadas para realizar transações possam ser utilizadas para o envio de propagandas.

3. Acesso (*Access*) - *sites* da Web devem oferecer aos usuários acesso às suas informações pessoais coletadas, dando-lhes, inclusive, a oportunidade de estarem atualizando, corrigindo ou apagando essas informações.

4. Segurança (*Security*) - *sites* da Web devem proteger, com segurança, as informações coletadas sobre os usuários.

País	Ações
Austrália	O principal estatuto federal é o "Decreto de Privacidade" de 1988, que possui onze princípios que se aplicam às atividades de setores públicos e privados. Este decreto criou o "Comissário Federal de Privacidade" que é o responsável pela fiscalização do cumprimento da lei.
Nova Zelândia	Em 1993, foi aprovado o "Decreto de Privacidade da Nova Zelândia", que regula a coleta, uso e disseminação de informações pessoais pelos setores públicos e privados. Antes mesmo da aprovação desse decreto, em 1991, foi criada a "Repartição do Comissário de Privacidade", cuja função principal é a monitoração do cumprimento da legislação. O país está, atualmente, em negociações para adequar suas leis às diretivas da União Européia.

Tabela 2.5: Ações para proteção de dados na Oceania

Podemos observar que ambas as propostas se baseiam na idéias de que a privacidade está relacionada com a noção de consentimento do usuário e todas tentam cobrir todas as formas de uso de dados: coleta, processamento, armazenamento, manutenção e divulgação.

2.4 Invasão de privacidade na Web

Elgesem [20] distingue duas formas de invasão de privacidade. A primeira consiste na disseminação de informação pessoal sem o consentimento de seu proprietário. A segunda forma diz respeito ao uso de informações pessoais para tomar decisões relacionadas ao indivíduo. O problema dessa segunda forma de invasão de privacidade reside no fato de que se estas decisões forem tomadas com base em informações irrelevantes ou incorretas, então o indivíduo pode ser prejudicado. Às vezes, mesmo quando o conjunto de informações disponíveis for correto e relevante, ainda assim as conclusões obtidas a partir de um conjunto de informações podem estar incorretas ([8, 20, 57]). Por exemplo, alguém pode fazer uma pesquisa sobre o tema AIDS e, posteriormente, ter um emprego ou um plano de seguro de vida ou saúde negado, porque a empresa envolvida concluiu que a pessoa é aidética. Essas duas formas de invasão de privacidade sempre ocorreram, mas a tecnologia de informação permite disseminar e processar um conjunto de informações com muito mais eficiência.

A Web aumenta os riscos de invasão de privacidade, pois facilita a coleta, monitoramento e análise de informações sem que os usuários sequer percebam que isso esteja ocorrendo. E os dados coletados podem ser guardados por vários anos para serem usados em algum momento do futuro.

Uma pesquisa realizada na *Humboldt Universität zu Berlin* identificou um comportamento contraditório por parte dos usuários. Quando não conectados à Web, eles se dizem muito preocupados com a proteção de sua privacidade, contudo, uma vez conectados, parecem esquecer todas as suas preocupações e estão dispostos a revelar informações pessoais [53].

Talvez o problema seja a ignorância, conforme exposto por *Esther Dyson* [18]: "Algumas pessoas não estão muito preocupadas e, por isso, não tomam nenhum cuidado. Outras estão preocupadas demais e são paranóicas. Ninguém sabe o que é conhecido e o que não é".

A Web oferece muitas opções para que a invasão de privacidade possa ocorrer. Servidores Web podem armazenar registros contendo dados sobre: quais páginas um usuário visitou, quanto tempo permaneceu em cada página, o comportamento de navegação, e-mails recebidos e enviados. E todo esse conjunto de dados coletados pode ser correlacionado a outros, de tal forma que se torna possível descobrir informações que os usuários supunham estarem protegidas. Por exemplo, nos EUA, existem aproximadamente dez pessoas com um mesmo código de endereçamento postal (CEP), que possuam uma mesma data de aniversário. Portanto, uma página da Web que solicite a data de aniversário, o CEP e a idade de um usuário praticamente também estará tendo acesso ao nome e endereço do usuário [25]. De forma análoga, a combinação de data de aniversário, sexo e CEP identificam univocamente 87% da população americana [56].

Às vezes, essa vigilância dos dados dos usuários (*dataveillance*) tem por objetivo lhes trazer benefícios. Por exemplo, é importante que companhias telefônicas e administradoras de cartões de crédito analisem o comportamento de seus usuários para que seja mais fácil detectar e prevenir erros e fraudes. Entretanto, o que realmente caracteriza todas essas ações como invasão de privacidade é o fato de que os usuários, em geral, desconhecem o que está acontecendo, ou seja, tudo é feito sem o seu consentimento prévio.

Um outro ponto a observar é a relação existente entre privacidade e a dependência do

consentimento de alguém para que informações sejam divulgadas, consentimento este que irá variar de pessoa para pessoa, de acordo com a visão de bolha apresentada na seção 2.1.

Levando isso em conta, *Wang* [60] não descreve os tipos de invasão de privacidade, que dependem de uma visão pessoal, mas sim, as preocupações que um usuário deve ter com relação à sua privacidade na Web:

- Acesso impróprio: acesso direto ao computador do usuário, sem permissão ou aviso prévio.
- Coleta imprópria: coleta de dados do usuário, sem permissão ou aviso prévio.
- Monitoramento impróprio: monitoramento das atividades do usuário, sem permissão ou aviso prévio. Isso pode ser feito, por exemplo, usando *cookies*.
- Análise imprópria: análise dos dados do usuário, sem permissão ou aviso prévio e derivação de conclusões a partir dessa análise. Essas conclusões incluem as preferências e o comportamento do usuário ao fazer compras.
- Transferência imprópria: transferência de dados do usuário, sem permissão ou aviso prévio. Por exemplo, há companhias que vendem, publicam ou compartilham dados de seus clientes.
- Transmissão não desejada: transmissão de informações a consumidores em potencial, sem permissão ou aviso prévio.
- Armazenamento impróprio: armazenamento de dados de uma forma não segura, por exemplo, permitindo que um cliente acesse dados de outros clientes, ou que dados sejam alterados sem autorização.

Analisaremos, a seguir, algumas possíveis formas de invasão de privacidade na Web.

2.4.1 Divulgação de informações por navegadores

Navegadores (*browsers*) são programas cujo objetivo principal é exibir conteúdo disponível na Internet. É através deles que usuários se comunicam remotamente com o servidores, onde a informação é armazenada. O problema é que os navegadores em geral

enviam, aos servidores, mais informações do que o necessário para estabelecer uma comunicação: a data e a hora da requisição; o tipo de navegador utilizado; a página que o usuário estava consultando; o sistema operacional instalado. Mas o mais grave são as informações que a própria URL (*Uniform Resource Locator*) carrega. Por exemplo, a URL *www.google.com/search?q=AIDS+treatments-&btnG=Google+Search* disponibiliza dois tipos de informações: a primeira é que o usuário está fazendo uma pesquisa usando o *Google*; a segunda, é a seqüência de caracteres que segue o delimitador '?' e mostra a expressão consultada (*query string*). No caso, a seqüência nos mostra que o usuário está interessado em tratamentos contra a AIDS. *Martin Jr. et al.* [45] nos relatam que, em experimentos realizados, descobriram várias expressões contendo o nome do usuário, e-mail, endereço residencial, número de telefone, número de vôo, e assim por adiante.

2.4.2 Cookies

Um *cookie* é um pequeno arquivo de texto, geralmente gravado na própria máquina do usuário, contendo informações trocadas entre um servidor Web e um usuário, através do navegador. O objetivo é gravar dados, ações e preferências do usuário, para solucionar a característica de ausência de estado do protocolo HTTP. A presença de *cookies* é importante principalmente no contexto de comércio eletrônico, para saber, por exemplo, o que está no carrinho de compras de um usuário.

Cookies ameaçam a privacidade porque, na maioria dos casos, armazenam dados sem o consentimento do usuário. Além disso, não é raro que esses dados sejam distribuídos e disponibilizados sem o conhecimento do usuário.

Os navegadores só armazenam *cookies* recebidos de servidores já visitados. O problema é que pode acontecer do navegador visitar um servidor sem que o usuário saiba e ter um *cookie* desse servidor armazenado em sua máquina. Esse *cookies* são conhecidos por *cookies de terceiros*. Um navegador pode receber *cookies de terceiros* quando, por exemplo, carrega uma página de um *site* que possui imagens de outro *site* que, por sua vez, envia um *cookie* junto com as imagens. Deve-se estar atento à aceitação de *cookies de terceiros*, pois estes permitem compartilhar informações de vários *sites*, facilitando uma melhor análise do perfil do usuário.

Os navegadores mais conhecidos oferecem, ao usuário, a opção de desligar *cookies*. No

entanto, essa solução nem sempre funciona, pois algumas páginas estão com o conteúdo tão vinculado ao uso de *cookies* que o usuário só conseguirá ter acesso a elas se aceitar *cookies*. Além disso, alguns navegadores não permitem desabilitar o envio de *cookies* que já foram aceitos. Para fazer isso, o usuário deverá apagar seus *cookies* explicitamente.

Um estudo mostrou que os usuários rejeitam menos de 1% dos *cookies*, em mais de um bilhão de páginas acessadas[7]. Segundo *Kristol* [41], este resultado pode ter várias justificativas, dentre as quais destacamos: usuários não sabem o que é um *cookie*; eles sabem o que é um *cookie* e para que serve, mas não estão preocupados; eles não sabem como desabilitar *cookies*; eles assumem que as entidades que irão coletar informação irão protegê-la; eles assumem que o Governo impedirá o uso inadequado de suas informações pessoais. Esse conjunto de justificativas nos mostra a necessidade de "alfabetização" do usuário da Web, como instrumento de proteção de sua privacidade.

Por fim, cabe ressaltar que *cookies* não são necessariamente ferramentas para invasão de privacidade. Podem ser usado como tal, mas também podem ser utilizados para melhorar a interação entre as aplicações Web e seus usuários.

2.4.3 Web bugs

Web bugs são pequenas imagens inseridas em páginas Web ou em mensagens de correio eletrônico, para monitorar usuários da Web. Em geral, um *Web bug* é uma imagem do tipo *Graphics Interchange Format*(GIF), transparente, de tamanho 1 pixel x 1 pixel. Por suas características, também são conhecidos como *clear GIFs*, *1-by-1 GIFs* ou *invisible GIFs*. Como são invisíveis aos olhos de um usuário comum, para visualizá-los é necessário ver o código HTML da página ou da mensagem que os contém. Como os usuários não ficam lendo o código HTML das páginas acessadas, acabam por carregar *Web bugs*, juntamente com o restante do conteúdo de uma página Web, sem o saberem.

De acordo com *Curtin et al.* [15], o número de *Web bugs*, no ano de 1999, havia mais que duplicado em relação ao ano anterior. Em 2000, foram encontrados mais de 4 milhões de *Web bugs*. Esse mecanismo está presente em vários *sites* que os usuários estão acostumados a utilizar, sem preocupação, como, por exemplo: *netscape.com*, *geocities.com*, *yahoo.com*, *google.com*, *aol.com*, *amazon.com*. Apesar disso, pouco se fala sobre

[7] http://www.websidestory.com/cgi-bin/wss.cgi?corporate&news&press_2_124

o assunto.

Ao carregar, sem saber, um *Web bug* que em geral pertence a um *site* distinto daquele com o qual o usuário está interagindo diretamente, um usuário estará disponibilizando várias informações, tais como: o tipo do navegador que carregou o *Web bug*, a hora que a imagem foi carregada, o endereço IP da máquina que carregou o *Web bug*, a URL do *site* que está sendo visitado pelo usuário. Através das informações disponibilizadas, torna-se possível obter vários resultados, por exemplo: o número de vezes que uma determinada propaganda foi mostrada, as páginas visitadas por um usuário, perfil dos usuários, relação entre propagandas visitadas e compras efetuadas.

Alguns defendem a idéia de que *Web bugs* permitem que empresas melhorem a qualidade de seus serviços, com os dados e estatísticas disponibilizados. Acrescentam, também, que eles são tão pequenos que com certeza não perturbam ninguém. Cabe aqui, repetir a pergunta presente no *site* da empresa *Bugnosis*[8]: "até que ponto estar quieto é só para evitar que o usuário seja perturbado e até que ponto estar quieto é uma tentativa para evitar ser detectado?"

2.4.4 Código móvel

Com o intuito de aumentar a funcionalidade de navegadores, foram desenvolvidas novas tecnologias para "baixar" arquivos de programas e executá-los automaticamente. Esses tipos de programas são comumente denominados de *código móvel*. Dentre as tecnologias existentes para geração de código móvel destacam-se: *ActiveX*, *Java*, *Javascript*, *Flash*.

O problema dos códigos móveis é que podem ser utilizados para fins negativos: programas que apagam o conteúdo do disco do usuário, disseminam vírus de computadores, pesquisam e transmitem informações armazenadas no disco de usuários que, muitas vezes, acreditavam estar anônimos [25]. Em especial, esta última aplicação de códigos móveis representa um sério risco à privacidade de usuários da Web.

2.4.5 Ataques a cache

Há um tipo de ataque que pode ser feito contra o cache do navegador e que torna possível determinar se um visitante de um *site* visitou ou não um outro *site* da Web [22]. O ataque

[8] http://www.bugnosis.com

pode ser feito sem o conhecimento/consentimento do usuário e do *site* visitado e, por isso, caracteriza invasão de privacidade.

Basicamente, o método de ataque consiste em medir o tempo que um navegador gasta para carregar um determinado conteúdo. A cache do navegador armazena o conteúdo de um conjunto de páginas acessadas recentemente. O tempo para acessar o conteúdo de uma cache é, em média, aproximadamente 50% a 80% menor do que o tempo necessário para acessar o mesmo conteúdo diretamente de um servidor Web [22]. Assim, se o tempo para carregar um determinado conteúdo for pequeno, pode-se deduzir que esse conteúdo está presente na cache e, portanto, a página a que esse conteúdo pertence já foi acessada pelo usuário.

Segundo *Felten & Schneider* [22], não há solução para este tipo de ataque, pois:

- para evitar o ataque, seria necessário desligar a opção de cache, o que acarretaria em grande perda de desempenho;
- há diferentes técnicas para forçar um navegador a carregar um conteúdo específico e, portanto, torna-se difícil evitar todas elas;
- uma outra opção para evitar o ataque seria aumentar aleatoriamente o tempo de acesso a uma página, mesmo que ela esteja presente na cache, o que também é inviável, pois o resultado seria similar ao processo de desligar a cache.

2.5 Escopo da tese

2.5.1 Problema central

Atualmente, os administradores de *sites* têm um grande interesse em encontrar características de seus usuários, quanto a preferências e uso. Esse tipo de informação permite-lhes melhorar o projeto dos serviços oferecidos pelo *site*, bem como possibilita a identificação de um usuário a cada vez que visitar o *site*, com o objetivo de personalizar o *site* às suas características e interesses, tornando a interação mais agradável [54].

Um dos desafios do tema privacidade na Web é resolver o conflito da personalização sem invasão de privacidade. Em outras palavras, como oferecer um serviço mais eficiente

e direcionado às características do usuário, sem realizar uma mineração dos dados (*data mining*), ou seja, sem armazenar, analisar e monitorar os dados e atividades do usuário?

Ann Cavoukian [9] apresentou uma estimativa de que aproximadamente metade das 1000 maiores companhias do mundo usam da mineração de dados. Como exemplo de empresas que analisam seus bancos de dados para predizer tendências e comportamentos futuros citamos: *Blockbuster*, *American Express* e *MasterCard*.

Analisando os oito princípios da OECD, apresentados na seção 2.3, e levando em consideração o trabalho realizado por *Cavoukian* [9] e discutido por *Thearling* [59], podemos concluir que a mineração de dados vai contra todos esses princípios:

1. Princípio do Limite de Coleta: a mineração de dados faz uma coleta de dados sem limites e sem autorização prévia do usuário.

2. Princípio da Qualidade dos Dados: como, em geral, os usuários não sabem que a mineração de dados está ocorrendo, nem tampouco os dados que estão sendo coletados, não há como mantê-los atualizados.

3. Princípio da Especificação de Objetivo: segundo *Ann Cavoukian* [9], no contexto de mineração de dados, talvez esse princípio seja o mais difícil de ser cumprido. "O minerador de dados não sabe, não pode saber, quais dados pessoais serão importantes e quais relacionamentos irão surgir. Assim, identificar um objetivo principal, no início do processo, e depois restringir o uso dos dados a esse objetivo é uma antítese da prática da mineração de dados". Há um risco das empresas colocarem que o principal objetivo da coleta de dados é a "mineração de dados". Esse objetivo não pode ser aceito pela sociedade, por não respeitar o Princípio da Especificação de Objetivo, já que a "mineração de dados" pode incluir qualquer tipo de processamento e análise de dados.

4. Princípio da Limitação de Uso: as técnicas de mineração de dados permitem que dados coletados para um objetivo sejam utilizados para outros objetivos secundários. Na verdade, a mineração de dados está associada a uma coleta de dados para uso futuro, cujo objetivo não é conhecido no momento desta.

5. Princípio da Segurança: não há garantias de que os dados coletados estejam armazenados de forma segura.

6. Princípio da Transparência: esse princípio traz que as pessoas devem estar cientes de como os seus dados estão sendo usados e armazenados. Entretanto, a mineração de dados por si só não é uma tarefa transparente. Além disso, grande parte dos usuários nem sabem que as suas informações pessoais estão sendo usadas em atividades de mineração de dados e nem tampouco pode-se esperar isso deles. Portanto, para que esse princípio seja respeitado, é necessário desenvolver ambientes que ajam em defesa dos consumidores, caso eles assim o desejem.

7. Princípio da Participação Individual: como conseqüência da falta de transparência da mineração de dados, não há como o usuário requisitar acesso a suas informações.

8. Princípio da Responsabilidade: por conseqüência, como todos os demais princípios são desrespeitados, esse também o será.

É imprescindível oferecer aos consumidores formas de alertá-los sobre o que realmente está ocorrendo durante interação com um *site*. Ou então, uma das soluções propostas por *Cavoukian* [9] é oferecer ao usuário uma das três opções abaixo:

1. Não permitir nenhuma mineração de dados.

2. Permitir mineração de dados somente para uso interno (*site* com o qual o usuário interagiu).

3. Permitir mineração de dados para usos interno e externo (terceiros).

O problema dessa proposta é que há o risco dos usuários sempre optarem por não permitir nenhuma mineração de dados, o que, de certa forma, também seria negativo, pois, sem dados, não haveria a possibilidade de avaliar e incrementar o projeto de *sites*, nem tampouco de oferecer serviços personalizados às características dos usuários. Assim, outras propostas devem ser buscadas e implementadas.

2.5.2 Trabalho desenvolvido

A privacidade é um tema que envolve conceitos políticos, filosóficos, éticos e tecnológicos. Por se tratar de um trabalho da área de Computação, esta tese focaliza os aspectos técnicos de proteção de privacidade: por um lado, de que forma a tecnologia de informação facilita

a invasão de privacidade e, por outro lado, como pode ser utilizada para proteção da privacidade dos usuários da Web.

Apesar da arquitetura proposta se basear no uso da anonimidade, não será abordada a anonimidade de emails. E, embora a proposta da tese seja útil no contexto do comércio eletrônico, não estudaremos protocolos de pagamento, como *Digicash*[9] e *FirstVirtual*[10], que já estão devidamente descritos e detalhados em outros trabalhos.

É importante também esclarecer que não faz parte do escopo desta tese a proposta de políticas de privacidade, nem tampouco a discussão de aspectos de segurança de redes. A arquitetura proposta pressupõe que os recursos e tecnologias de segurança de redes existentes já estão devidamente implantados.

Esta tese aborda os tópicos descritos a seguir:

Privacidade São apresentados conceitos de privacidade e uma discussão sucinta sobre a importância, para a sociedade, de haver recursos para proteção da privacidade dos indivíduos. Também são descritas formas de invasão de privacidade na Web e as principais regulamentações relacionadas à proteção de privacidade na Web.

Proteção de privacidade São listadas as estratégias que podem ser utilizadas pelas pessoas para proteção de sua privacidade no mundo real e no mundo virtual. No mundo virtual, como uma única estratégia não é suficiente para proteger o usuário, este deverá se utilizar de uma combinação de vários recursos, simultaneamente, de acordo com a taxonomia apresentada, para camadas de proteção de privacidade (Seção 3.3).

Distinção entre privacidade e segurança Os conceitos de privacidade e segurança são discutidos de forma a mostrar as diferenças entre estes dois conceitos, sob quais aspectos a privacidade depende da segurança e em quais situações uma está dissociada da outra.

Proposta de solução para o conflito entre privacidade e personalização A necessidade de proteção de privacidade acaba por afetar o acesso do usuário a facilidades e serviços que lhe são úteis, como é o caso dos serviços personalizados. Para solução deste

[9] http://www.digicash.com
[10] http://www.fv.com

conflito é proposta uma arquitetura que permite ao usuário ter um melhor controle de sua privacidade, sem deixar de ter acesso a serviços personalizados.

Critérios e metodologia para avaliação da arquitetura proposta A arquitetura proposta foi avaliada qualitativa e quantitativamente através de critérios e metodologia que podem ser aplicados a outros trabalhos na área.

Projetos futuros A arquitetura proposta neste trabalho pode ser aperfeiçoada em busca de novas aplicações. É interessante, também, que novas metodologias de avaliação da arquitetura sejam desenvolvidas, para que se torne possível uma análise mais profunda da qualidade dos dados disponibilizados.

Capítulo 3

Proteção de Privacidade

3.1 Proteção de privacidade no mundo real

A preocupação com a proteção de privacidade não é um tema novo e vários métodos foram e são utilizados pelas pessoas com o intuito de se preservarem. Dentre esses métodos, estão: criptografia, anonimato, uso de pseudônimos e uso de máscaras.

3.1.1 Criptografia

A criptografia é a utilização de algum método matemático para proteção de informação. A idéia básica é cifrar ou transformar uma mensagem de tal forma que a torne ininteligível a todos, exceto àqueles que possuam a chave de deciframento que permite recuperar a mensagem original. Em outras palavras, a criptografia previne ou dificulta o acesso não-autorizado a informações.

O primeiro uso comprovado de criptografia foi por volta do ano 1900 A.C., no Egito. Historicamente, a criptografia foi utilizada para fins militares, mas, nas últimas décadas, o seu uso se estendeu a outras áreas: informações governamentais, eleições, comércio eletrônico e transações financeiras, dentre outras [25].

3.1.2 Anonimato

O anonimato, ou ocultamento do nome do autor de uma ação ou obra, representa uma forma antiga de agir ou produzir obras, com a proteção da privacidade da identidade do

autor da ação ou obra.

O anonimato pode ser usado tanto para objetivos socialmente lícitos quanto para ilícitos. Dentre os objetivos lícitos, destacamos: testemunho e denúncia de crimes; participação em grupos de ajuda, como os Alcoólicos Anônimos. Em outras palavras, o anonimato é uma forma de dar a um indivíduo, maior liberdade de expressão e ação.

Quanto aos usos ilícitos do anonimato, podemos citar: envio de cartas com conteúdo ameaçador, fraudes, ações criminosas e terroristas.

Pseudônimos

Ao contrário do anonimato, onde o objetivo é não se identificar, o pseudônimo é um identificador que pode ser utilizado por um indivíduo mais de uma vez. E, se for descoberto a verdadeira identidade associada a um pseudônimo, todo o conjunto de ações e obras realizadas no passado, sob um determinado pseudônimo, poderão ser automaticamente transferidas para o indivíduo que o utilizava. Os objetivos para o uso de pseudônimos são variados. Escritores, jornalistas e artistas podem utilizar desse recurso para se manifestarem e, ao mesmo tempo, evitarem perseguição política. Mulheres podem preferir usar um pseudônimo masculino para evitar discriminação de sexo. Concursos artísticos constumam solicitar que as pessoas escolham pseudônimos para garantir uma maior transparência do processo de avaliação.

3.1.3 Máscaras

As pessoas podem utilizar dois tipos de máscaras: as físicas e as psicológicas.

As máscaras físicas têm por objetivo o anonimato e são utilizadas tanto em momentos de diversão (festas), como para objetivos criminosos. É comum criminosos utilizarem esse recurso para não serem reconhecidos.

As máscaras psicológicas ou *personae* foram definidas por *Carl Gustav Jung* [30] como sendo a tentativa de um indivíduo de se esconder ou de camuflar a personalidade real, em resposta às convenções da sociedade e às "suas próprias necessidades arquetípicas interna. O propósito da máscara é produzir uma impressão definida nos outros e, muitas vezes, embora não obrigatoriamente, dissimula a natureza real do indivíduo". Dessa forma a *persona* representa o conjunto de características que cada um apresenta ao mundo, em

oposição às suas características reais. Em outras palavras, algumas características poderão ficar escondidas por detrás da máscara.

Uma mesma pessoa pode utilizar várias máscaras e, normalmente, os indivíduos fazem uso de máscaras diferenciadas para o trabalho, a vida familiar e a vida social.

3.2 Proteção de privacidade em ambientes eletrônicos

Da mesma forma que a tecnologia pode ser utilizada para automatizar o processo de coleta e análise de dados, ela também pode ser utilizada para aumentar o controle dos usuários sobre suas informações pessoais.

Na Web, podem ser aplicadas as mesmas técnicas utilizadas no mundo real e citadas na seção 3.1. Apresentamos, a seguir, de que forma elas e outras técnicas adicionais se aplicam.

3.2.1 Criptografia

Segundo *Wang* [60], as ferramentas de encriptação são as mais utilizadas e as que obtiveram mais sucesso com relação à proteção da privacidade de usuários da Internet. A vantagem dessas ferramentas é impedir que um terceiro compreenda o conteúdo de mensagens transmitidas entre dois outros indivíduos. Conseqüentemente, se um terceiro não é capaz de entender uma mensagem, não haverá interesse em coletar e armazenar essas informações.

Entretanto, esse método não é totalmente eficiente contra a mineração de dados, pois mesmo sem ser possível saber o conteúdo de uma mensagem, ainda é possível saber o endereço IP do cliente e servidor, o comprimento dos dados permutados, a hora em que uma comunicação foi realizada e a freqüência das transmissões. Por isso, ele deve ser utilizado em conjunto com outras opções de tecnologia para proteção de privacidade.

Dentre os programas e protocolos de criptografia existentes, destacam-se: PGP (Pretty Good Privacy)[1], S/MIME (Secure/Multipurpose Internet Mail Extensions)[2], SSL (Secure

[1] http://www.pgp.com
[2] http://www.ietf.org/html.charters/smime-charter.html

Socket Layer)[3], SET (Secure Electronic Transactions)[4] e SSH (Secure Shell)[5].

3.2.2 Agente de privacidade

Um tipo de ferramenta para proteção de privacidade seriam os programas que mantêm os usuários informados acerca do seu grau de exposição e dos riscos que correm de terem sua privacidade invadida.

Dentro dessa abordagem, *Ackerman & Cranor* [1] propuseram os *Privacy Critics* (críticos de privacidade), que são um tipo de agente inteligente que auxilia usuários a protegerem suas informações privadas, através de sugestões e *feedbacks*.

Críticos possuem duas características importantes:

1. Eles fornecem *feedback* aos usuários, mas não necessariamente agem em seu nome. Um exemplo muito conhecido é o Assistente do *Microsoft Office*.

2. Um ambiente de críticos pode ter centenas de outros críticos independentes, trabalhando com diversos tipos de tarefas. Os usuários têm a liberdade de "ligar/desligar" esses críticos.

Mais especificamente, os críticos de privacidade dão ao usuário um maior controle de suas informações privadas, no sentido de que terão maior consciência do que possa estar ocorrendo com seus dados. De acordo com os resultados preliminares obtidos, as pessoas apreciam saber que existe algo "tomando conta" de sua privacidade, sem interferir demais em suas ações e sem tomar automaticamente atitudes em seu nome, sem o seu conhecimento, da forma como outros tipos de agentes normalmente o fazem.

3.2.3 Filtros

Filtros são ferramentas que seletivamente bloqueiam emails, páginas Web, *cookies*, propagandas, *JavaScript* e outros conteúdos. Filtros criam dificuldades para que a invasão de privacidade ocorra, mas não eliminam o risco, pois algumas informações do usuário

[3]http://home.netscape.com/eng/ssl3
[4]http://www.visa.com/set, http://www.mastercard.com/set
[5]http://www.ssh.com

ainda continuam expostas, como, por exemplo, seu endereço IP, a hora e a duração da interação com um *site*, sua localização geográfica.

Filtros são muito utilizados por pais que desejam evitar que seus filhos forneçam informações pessoais para estranhos ou que acessem conteúdos impróprios para sua idade. Como exemplos de filtros, citamos as ferramentas *CyberSitter*[6] e PGP[7].

3.2.4 Anonimidade

Uma estratégia útil para proteger privacidade é anonimidade. No caso da Web, o nome que se quer proteger é o endereço IP da máquina do usuário. Há várias razões para se querer proteger o endereço IP [25]: os endereços IP podem conter informações pessoais (por exemplo, a localização geográfica do usuário) e, da mesma forma que *cookies*, podem ser utilizados para correlacionar atividades através de diferentes *sites*. O endereço IP também pode ser usado para recuperar transações supostamente "anônimas" para revelar a identidade real de um usuário.

Uma solução para esse problema seria navegar a partir de um terminal público, como os terminais de bibliotecas públicas, escolas e cibercafés. Uma outra solução seria utilizar ferramentas de anonimidade.

Há dois tipos de anonimidade [27]: pseudo-anonimidade e anonimidade de uma única vez. A diferença está no fato de que os pseudônimos são persistentes, ou seja, os usuários mantêm uma determinada identificação em várias interações, identificação esta que, logicamente, não pode estar conectada à identidade do usuário. Na anonimidade de uma única vez, uma identificação de usuário só é válida durante uma interação. Assim, quando se usa pseudônimos, apesar de não ser possível associá-lo ao verdadeiro nome, é possível associar um conjunto de mensagens a um único usuário. No caso da anonimidade de uma única vez, nenhuma ligação entre mensagens e usuários pode ser feita.

O grande problema relacionado a anonimidade de uma única vez é a falta de possibilidade de oferecer serviços personalizados, como por exemplo, receber recomendações e adquirir privilégios por ser um cliente fiel. O uso de pseudônimos alivia esse problema, mas torna mais fácil descobrir o verdadeiro autor das mensagens.

[6] http://www.cybersitter.com
[7] http://www.pgp.com

Uma falha da anonimidade é que não consegue proteger a anonimidade de um usuário se o conteúdo da transação revelar sua identidade ao servidor Web. Esta situação ocorre, por exemplo, quando o usuário envia a um *site* um formulário preenchido, contendo dados pessoais como o seu nome e *e-mail*. Também não haverá proteção se o conteúdo de uma página for executável e abrir conexões diretas entre o navegador e o servidor Web, como no caso de *applets* Java.

Várias ferramentas de anonimidade se baseiam no uso de *proxies*: coloca-se um terceiro - o *proxy* - para submeter requisições Web em nome dos usuários. Como todas as requisições são submetidas pelo *proxy*, o único endereço IP revelado aos *sites* é o do *proxy*. Como os usuários desse serviço não são anônimos ao *proxy*, esse tipo de sistema é vulnerável a alguém que tenha controle ou acesso ao *proxy*, pois nesse caso é possível monitorar os remetentes e os destinatários de todas as comunicações. Além disso, se o *proxy* falha, não é possível continuar a navegação anônima pela Web. Exemplos de ferramentas que utilizam esta tecnologia: *Anonymizer*[8] e *HideIP*[9].

Dentre as várias ferramentas e tecnologias de anonimidade, ressaltamos as seguintes:

- *Anonymizer*[10] - um *proxy* Web que filtra todas as identificações do navegador. Isso permite que os usuários "surfem" pela Web anonimamente, sem revelar suas identidades ao servidor.

- *Onion Routing*[11] - se baseia em uma rede de *mixes*. Cada *mix* é um roteador responsável por esconder o caminho de uma mensagem através da rede ou, em outras palavras, impede que o destinatário de uma mensagem descubra quem foi o remetente. Esse processo ocorre através do uso de criptografia e de outras técnicas que tenham por objetivo impedir que um espião possa associar mensagens que chegam em um *mix*, com as que saem: transmitir mensagens em uma ordem diferente da ordem de chegada, gerar mensagens de mesmo tamanho e, no caso de um tráfego escasso de mensagens, gerar aleatoriamente mensagens extras para outros componentes da rede.

[8] http://www.anonymizer.com
[9] http://www.hideip.com
[10] http://www.anonymizer.com
[11] http://www.onion-router.net

Onion (cebola) *Routing* tem esse nome, porque o usuário dessa tecnologia cria uma estrutura de dados em camadas, chamada *onion*, que determina os algoritmos e as chaves de ciframento que serão usadas durante o transporte dos dados ao destinatário final. A cada parada (*onion-router*) da rota, uma camada de ciframento é removida, de acordo com as informações contidas no *onion*. A mensagem chega ao destinatário, na forma original, contendo somente o endereço IP do último *onion-router* do caminho. A vantagem dessa tecnologia é que não requer um terceiro centralizando o envio de mensagens.

- *Crowds* [49] - esse método se baseia na idéia de que uma pessoa pode ficar anônima, quando no meio de uma multidão. Para executar uma transação Web, um usuário deve primeiro entrar em um grupo (*crowd*) de usuários. A requisição do usuário será transmitida primeiramente a um membro aleatório do grupo. Esse membro pode submeter a requisição diretamente para o servidor final ou encaminhá-la para outro membro do grupo e assim por diante. Portanto, quando uma submissão é realizada, ela é feita por um membro aleatório do grupo, tornando difícil identificar o real "disparador" da requisição. Uma vantagem dessa tecnologia é que, da mesma forma que as redes de *mixes*, não depende de terceiros para manter a anonimidade de um usuário.

  A grande diferença entre uma rede de *mixes* e o *Crowds*, é que, no primeiro, o usuário determina um caminho a ser percorrido e, no segundo, esse caminho é definido à medida que uma mensagem for transmitida entre membros do grupo. A vantagem do *Crowds* é a maior facilidade em se adaptar a mudanças na rede.

Garvish & Gerdes [26] também destacam três aspectos de anonimidade que devem ser considerados:

- Anonimidade ambiental - fatores externos ao sistema de anonimidade que devem ser observados durante a sua operação. Esses fatores incluem: número de participantes, conhecimento de dados prévios dos participantes. Por exemplo, não faz sentido um sistema de anonimidade, seguro e bem projetado, que só seja utilizado por uma única pessoa cuja identidade seja de alguma forma conhecida externamente ao sistema.

- Anonimidade baseada em conteúdo - esse tipo de anonimidade existe quando não é possível encontrar pistas sobre a identidade real do usuário, pelo conteúdo que está sendo disponibilizado. Exemplos de pistas: nome, endereço, e-mail, padrão de comportamento, estilo de escrita.

- Anonimidade procedimental - esta anonimidade depende do protocolo de comunicação utilizado. Por exemplo, o endereço de um nodo da rede pode revelar a identificação de um usuário, se este nodo estiver associado a um único usuário.

Segundo *Schreck* ([51], [40]), para proteger a privacidade de um usuário através de anonimidade, os três tipos de anonimidade devem estar presentes simultaneamente em um sistema adaptado ao usuário.

Pseudo-anonimidade

O uso de pseudônimos pode representar uma solução parcial para os problemas relacionados à anonimidade. Uma desvantagem do uso de pseudônimos está no fato de que permanece uma ligação entre uma ação/obra e seu autor, o que representa um ponto de vulnerabilidade.

É comum haver um terceiro, intermediando a troca de informações. Há dois problemas com essa abordagem: o primeiro é fazer com que a comunidade entre em acordo com relação a qual entidade é confiável. A segunda diz respeito ao fato de que esse terceiro, como centralizador de informações, pode se tornar um ponto vulnerável para ataque e um ponto do qual todos dependam para que o sistema funcione.

Lucent Personalized Web Assistant (LPWA) [44] é a mais conhecida ferramenta baseada no uso de pseudônimos. Inicialmente conhecida por *Janus* e, atualmente, por *Proxymate*, o LPWA foi projetado para utilizar um mesmo pesudônimo todas as vezes que um determinado usuário retorne a um mesmo *site*, mas usa um pseudônimo diferente para cada *site*. Esse pseudônimo também é utilizado em formulários que solicitem o nome do usuário. A vantagem dessa tecnologia é que permite que *sites* da Web definam um perfil de cada usuário, a fim de personalizar o conteúdo das páginas, sem permitir que este perfil esteja associado a um nome de usuário ou que este seja combinado com informações reveladas por outros *sites*. Mas essa tecnologia possui o mesmo problema que o uso de pseudônimos, no mundo real: se alguém descobre a identidade real que está por trás de

um pseudônimo, todas as ações passadas do indivíduo, que foram realizadas sob o mesmo pseudônimo, estarão automaticamente expostas.

3.2.5 Máscaras

Na Web, da mesma forma que no mundo real, uma pessoa pode se esconder atrás de máscaras ou *personae* (vide Seção 3.1.3). Uma *persona* digital é um modelo da personalidade pública de um indivíduo, uma representação simplificada de alguns aspectos da realidade. Uma pessoa pode ter múltiplas máscaras, até mesmo para interagir com uma mesma organização. Cada máscara pode refletir um dos vários papéis ou interesses que uma pessoa representa, ou possui, ao se relacionar com uma organização [10].

Apresentamos, a seguir, algumas propostas de trabalhos que se baseiam na definição de máscaras, também por conhecidas por *personae* ou perfis:

- *Persona* [17] - uma *persona* é uma coleção de dados pessoais que o cliente irá disponibilizar para um dado site. Esses dados não incluem os interesses, nem tampouco o comportamento do usuário.

- *Information Crystals* [3] - Este mecanismo se propõe a preservar a anonimidade de uma pessoa, ao mesmo tempo que gera perfis que podem ser utilizados por companhias, para mineração de dados. Essa anonimidade se baseia na idéia de camuflagem, ou seja, pacotes de informação de um indivíduo ao se misturarem com outros com características similares, acabam se escondendo.

  Os perfis e preferências de um indivíduo são definidos em códigos móveis e cifrados, denominados *infoatoms*. Cada pessoa gera seus próprios infoatoms em seu computador pessoal. Dessa forma, cada um decide quais informações deseja disponibilizar e quais deseja manter privadas.

  Os *infoatoms* têm a capacidade de interagir e se ligar a outros *infoatoms*, formando cristais de informação. O cristal irá coletar e emitir um perfil dos dados e as estatísticas dos dados agregados serão transmitidas ao minerador de dados.

  A motivação para o desenvolvimento desse método se baseou no fato de que os consumidores aceitam disponibilizar suas informações, se forem recompensados de alguma forma e se sua privacidade estiver protegida. No caso desse método, a

compensação oferecida aos usuários é o recebimento de pagamentos automáticos, quando outros usam porções de seus dados.

Uma vantagem dessa proposta está em resolver o grande conflito que existe entre privacidade de usuários e a demanda de mineradores de dados. Mas o problema da personalização continuou sem solução.

Um problema dessa tecnologia é que usuários continuam sem saber como ou com qual objetivo seus dados serão utilizados. Assim, por exemplo, os resultados agregados dos dados podem ser utilizados para objetivos invasivos de privacidade, como saber se uma determinada população tem propensão a ter um certo tipo de doença.

3.2.6 Protocolos para especificação de uso e coleta de dados de usuários

Em geral, a privacidade é discutida como um problema social, isto é, uma negociação em uma comunidade sobre o processamento de informação pessoal. Um dos métodos para negociação é através das políticas de privacidade. Uma política de privacidade é um conjunto de especificações sobre práticas de coleta e uso da informação de uma organização. A política deve poder ser modificada para poder satisfazer aos requisitos de privacidade do usuário. Até há pouco tempo atrás, o usuário só tinha duas opções: aceitar um sistema ou não. Hoje, ele já tem como ajustar suas preferências.

Um problema das políticas de privacidade divulgadas pelas empresas reside na sua falta de clareza e legibilidade. Um trabalho realizado por um equipe de pesquisadores da *North Carolina State University* [5] identificou que para compreensão de 40 políticas examinadas, 12 requeriam um nível de escolaridade superior e 7 requeriam o equivalente ao nível de pós-graduação. Isso quer dizer que a compreensão total de uma política só será possível para aproximadamente 1/6 da população adulta da Internet.

O *Platform for Privacy Preference Project* (P3P) do *World Wide Web Consortium*[12] (W3C) é uma tentativa de padronização da linguagem de especificação de política de privacidade. P3P permite que *sites* Web negociem com o usuário, quais informações serão coletadas, como e para o quê serão utilizadas, da seguinte forma: P3P define um

[12]http://www.w3.org/P3P

protocolo que permite que administradores de *sites* publiquem a política de privacidade do *site*, em um formato padrão que pode ser recuperado automaticamente. Quando um usuário visita um *site*, o navegador lê a política de privacidade do *site* e a compara com as definições de segurança configuradas pelo usuário. Se as políticas forem satisfatórias, o navegador continua a requisição de páginas do *site*. Caso contrário, dúvidas podem ser resolvidas através de interação com o usuário.

Esse mecanismo possui várias desvantagens:

1. No mundo há várias leis que regulamentam a privacidade e será muito difícil unificar todas essas leis.

2. "Apesar de P3P fornecer um mecanismo para assegurar que usuários, antes de disponibilizar informações, estejam informados sobre políticas de privacidade, ele não fornece um mecanismo para assegurar que os *sites* ajam de acordo com suas políticas"[13]. Portanto, exige-se que os usuários confiem nos *sites*.

3. O objetivo da coleta de dados deverá estar claro para o usuário e os dados só poderão ser usados da forma proposta. Entretanto, conforme exposto na seção 2.5.1, isto não é possível.

4. A escolha que o usuário faz não é exatamente a de como proteger a sua privacidade, mas sim, de quanto de privacidade se estará dispensando.

5. Segundo Ackerman, "para ser realmente útil a uma grande quantidade de pessoas, P3P e outros protocolos similares irão requerer interfaces que sejam suficientemente fáceis de usar e adaptar às características do usuário" [1].

6. A União Européia rejeitou explicitamente o P3P, por considerar que esta proposta só visa a formalizar baixos padrões de proteção de privacidade[14].

7. P3P pode deixar usuários confusos, sob vários pontos de vista, por exemplo, achar que um *site* que apresenta uma política de privacidade é um *site* seguro ou achar que todo *site* que não implementa P3P é um *site* que viola sua privacidade [31].

[13]http://www.w3.org/P3P

[14]http://www.computerworld.com/securitytopics/security/privacy/story/0,10801,75389,00.html

8. Nos Estados Unidos, alguns críticos colocaram o P3P como uma tentativa das empresas, de se evitar uma legislação de proteção de privacidade na Internet. Na verdade, P3P pode realmente atuar como uma ferramenta eficiente a favor da argumentação que defende as políticas de privacidade, o que contraria a necessidade de uma legislação [33].

3.2.7 Agências de controle de confiabilidade

O que colocamos como agências de controle de confiabilidade na verdade são serviços que fornecem, ao consumidor, uma certa segurança de que a política do *site* realmente reflete suas práticas. Em geral, esses serviços exigem que os *sites* paguem uma taxa, aceitem certos acordos contratuais e, possivelmente, passem por um processo de auditoria, em troca da autorização para divulgar algum tipo de selo de aprovação. Esse tipo de solução também irá requerer a confiança dos usuários em uma determinada "agência".

Como exemplos de "agências" que oferecem esses serviços, citamos: TRUSTe[15], BBBOnLine[16] e Verisign[17].

A título de exemplo, apresentamos, a seguir, os requisitos que um *site* deve respeitar para obter uma licença do TRUSTe [6]:

- O *site* divulga, para o usuário, suas práticas de coleta e divulgação de informação, em uma linguagem fácil de ler e compreender.
- O *site* permite que o usuário opte se aceita ou não que suas informações sejam repassadas para terceiros.
- O *site* deve proteger os dados dos usuários, no sentido de que eles não poderão ser perdidos, mal utilizados ou alterados sem autorização.
- O *site* fornece algum mecanismo para que os usuários possam atualizar suas informações.
- O *site* estará passando periodicamente por revisões e verificações.

[15] http://www.truste.org
[16] http://www.bbbonline.com
[17] http://www.verisign.com

3.3 Camadas de proteção de privacidade

3.3.1 Exposição X Privacidade

Inicialmente, ao pensarmos em uma proposta de taxonomia para proteção de privacidade, tivemos a idéia de trabalhar com níveis de privacidade. Contudo, como esclarece *Millar*, privacidade está diretamente relacionada com a noção de consentimento, que é uma decisão completamente pessoal [47]. Dessa forma, por não ser possível definir uma taxonomia para níveis de privacidade, optamos por trabalhar com níveis de exposição. No nosso entender, a vantagem dos níveis de exposição era que eles seriam os mesmos para quaisquer pessoas, pois, independente do fato do que cada um considera invasão de privacidade, o grau de exposição diante de um determinado comportamento seria o mesmo.

Entretanto, após analisar alguns tópicos que contribuiriam na definição dos níveis de exposição, percebemos que vários fatores estariam interferindo e até mesmo impedindo uma formalização de níveis de exposição:

- Não pode haver uma única classificação para o uso de *cookies*. Alguns *cookies* seguem unicamente o objetivo existente quando da sua criação, ou seja, facilitar a construção de aplicações Web que devem "lembrar" o estado no qual o usuário estava, durante a última interação com o *site*, por exemplo, carrinho de compras ou preferências para personalização da página. Outros *sites* utilizam *cookies* para invadir privacidade, por exemplo, estudar o comportamento do usuário, sem o consentimento deste. Como nos traz *Kristol* [41], não há como fazer com que a tecnologia, sozinha, distinga o bom uso de *cookies*, do mau uso destes.

- Para identificar o nível de exposição do usuário, seria necessário conhecer não só o comportamento atual, mas também o passado. Entretanto, em geral, o conjunto de dados retidos por um *site* é maior do que o conjunto de informações disponíveis para uma ferramenta ou arquitetura de identificação do nível de exposição do usuário, pois as fontes de informações desta seriam restritas ao arquivo de histórico do usuário e às ações do usuário a partir do momento de sua implantação. Portanto, não haveria como saber, com precisão, se o nível de exposição associado a um usuário estaria correto e coerente com a quantidade de dados de posse dos *sites*.

- Classificar um usuário em um determinado nível de exposição poderia lhe dar a idéia incorreta de que, estando em um determinado nível, então estaria garantida a sua privacidade naquele nível. Entretanto, essa garantia não pode ser dada, já que, a todo momento, surge uma nova idéia ou uma nova pesquisa que apresenta alguma nova forma de invasão de privacidade.

- Sob um outro ponto de vista, o fato de um usuário se expor muito para um determinado *site* não quer dizer que a sua privacidade esteja sendo ou será invadida. Primeiro, porque a exposição de um usuário, com o seu consentimento, já descaracteriza a invasão de privacidade. Segundo, porque o fato de estar se expondo muito facilita a invasão de privacidade, mas não necessariamente implica que ela irá ocorrer, pois, por exemplo, nem todo *site* armazena informações de usuários para serem analisadas posteriormente.

- Não há como implementar uma tecnologia com a capacidade de identificar quando, como e quais informações dos usuários estão sendo armazenadas e muito menos se elas serão analisadas ou transmitidas para terceiros.

Portanto, passamos a falar não mais de níveis de exposição, mas sim, de níveis de proteção do usuário. Esses níveis de proteção de privacidade de usuários envolveriam desde o governo até o próprio usuário, baseando-se portanto, na conscientização deste com relação ao problema de se proteger e na mudança de atitude da sociedade, como um todo, quanto ao respeito à privacidade alheia. Entretanto, mais uma vez, essa terminologia estaria incorreta, porque quando se fala em níveis, fica implícito que todos os níveis inferiores estão presente, o que não ocorre no caso da privacidade. Por isso, optamos por utilizar a denominação de camadas de proteção de privacidade [35].

Usuários podem ter sua privacidade protegida através de diferentes camadas de proteção. Cada camada é independente das demais e, da mesma forma que camadas geológicas, a existência de uma camada não implica que as anteriores devam existir. Contudo, quando mais de uma camada está presente, a organização delas seguirá sempre a mesma ordem, conforme mostrado na Figura 3.1.

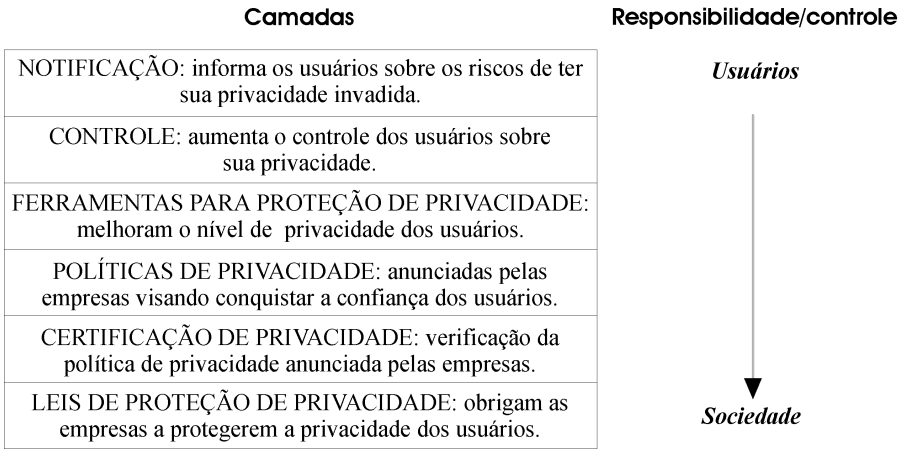


Figura 3.1: Camadas de proteção de privacidade

3.3.2 Camada 1: Notificação

Em geral, usuários não estão conscientes dos riscos que estão correndo de ter sua privacidade invadida. Em outras palavras, eles não sabem que tipo de informação pode ser derivada a partir de sua interação com um *site* [2]. Por exemplo, muitos usuários ainda não sabem o que é um *cookie* ou que, ao bloquear *cookies*, poderão perder a oportunidade de receber serviços personalizados. É também comum que usuários não saibam que cada clique em um *objeto* do *site* (*links*, figuras, botões, etc) possa ser utilizado para construir um perfil detalhado a seu respeito, ou até mesmo para descobrir quais *sites* visitou previamente. Assim, há uma necessidade clara de manter usuários informados sobre os riscos que estão correndo. Uma estratégia que pode ser adotada é "cenarizar" para os usuários, o que pode ser feito com as informações que libera.

Os navegadores atuais têm buscado informar os usuários e dar a eles a opção de definir o que querem proteger. Entretanto, o mecanismo utilizado se baseia no modelo *opt-out*, ou seja, sempre que os usuários quiserem proteger seus dados, eles devem comunicar isso explicitamente e, portanto, sempre que nada for informado, pode-se deduzir que os *sites* têm a liberdade de utilizar os dados dos usuários da forma como quiserem. Portanto, o mecanismo de *opt-out* coloca toda a responsabilidade da proteção de privacidade nas mãos dos usuários. Dessa forma, o ideal seria o modelo *opt-in*, no qual os usuários só necessitam informar explicitamente o que autorizam divulgar.

Essa primeira camada deve oferecer ao usuário acesso a informações diversas que

incluem:

- O que é um *cookie*, para que serve e, inclusive, as vantagens e desvantagens de sua utilização. Em outras palavras, não basta simplesmente bloqueá-los, deve-se também saber o que se estará perdendo.
- Cada clique em um *objeto* da página é uma informação que permitirá definir um padrão de comportamento e interesses do usuário.
- O uso de ferramentas de anonimidade protege a privacidade, mas também implica em menor grau de personalização.
- Um *site* pode descobrir se um usuário já acessou ou não um outro determinado *site*, através de, por exemplo, *Web bugs* e ataques a cache (Seção 2.4), mesmo usando ferramentas de anonimização [22]. Portanto, se alguém quiser ter privacidade quanto ao acesso a um determinado *site* ou pesquisa sobre um determinado assunto, deve evitar usar a Internet.
- A transmissão de dados pessoais deve ser sempre feita com segurança, mas mesmo que a transmissão seja segura, não há garantias de que haverá proteção da privacidade do usuário.

Privacy Critics [1] é um exemplo de ferramenta que se encaixa nessa primeira camada de proteção de privacidade. É importante ressaltar que informar sobre riscos é completamente diferente de agir, automaticamente, em defesa da privacidade de alguém. Não podemos nos esquecer que a privacidade está relacionada com a noção de consentimento e, por isso, usuários devem ter o direito de escolher o que querem e em quem confiam.

3.3.3 Camada 2: Controle

Apesar da privacidade ser um conceito pessoal, há algumas tecnologias que, sem dúvida alguma, criam condições para que a invasão de privacidade possa ocorrer. Como exemplos dessas tecnologias podemos citar os *cookies* de terceiros e os *Web bugs*, porque o propósito de ambos é oferecer condições para análise do comportamento do usuário sem o seu conhecimento e consentimento explícito. A camada 2 inclui mecanismos que permitem que

os usuários tenham controle sobre suas informações através de mecanismos ou ferramentas que ataquem essas tentativas explícitas de violação de sua privacidade.

Nessa camada, a tecnologia chave é o navegador da Web e suas extensões, como os *plugins*, que devem permitir que os usuários facilmente rejeitem ou filtrem métodos indesejáveis de coleta de dados.

Em geral, navegadores oferecem aos usuários a opção de rejeitar *cookies* ou *cookies* de terceiros. Entretanto, a seleção dessa opção não é uma tarefa muito fácil para muitos usuários da Web. Além disso, somente usuários que já têm alguma noção sobre seus riscos terão interesse em bloquear esse tipo de serviço. Por isso é tão importante que o usuário tenha acesso a alguma ferramenta da primeira camada de proteção de privacidade.

O mesmo ocorre com os arquivos de histórico, que registram todas as páginas já visitadas pelo usuário. Códigos maliciosos podem facilmente recuperar esse tipo de arquivo e divulgá-lo para terceiros. Por isso, os navegadores devem facilitar a tarefa do usuário de periodicamente editar ou apagar esses arquivos.

Uma outra opção que os usuário têm é a de instalar filtros em suas máquinas. Conforme trazido na seção 3.2.3, estes não eliminam o risco de invasão de privacidade. Por isso, usuários necessitam de camadas adicionais para proteção de sua privacidade.

3.3.4 Camada 3: Ferramentas para proteção de privacidade

Essa camada engloba a maior parte das ferramentas conhecidas para proteção de privacidade. A principal diferença entre esta camada e a anterior está no local onde reside o mecanismo de proteção de privacidade. Na camada 2, o mecanismo se encontra na própria máquina do usuário; na camada 3, o mecanismo opera de algum lugar da Web.

Os mecanismos mais explorados são a anonimidade (Seção 3.2.4) e a pseudo-anonimidade (Seção 3.2.4). Nesta camada, se localizam, por exemplo: *Anonymizer*[18], *Lucent Personalized Web Assistant* (LPWA)[19], *Onion Routing*[20] e *Crowds* [49].

[18] http://www.anonymizer.com
[19] http://www.bell-labs.com/projects/lpwa
[20] http://www.onion-router.net

3.3.5 Camada 4: Políticas de privacidade

A idéia desta camada é fornecer aos usuários informações sobre a política de privacidade do *site*, e deixá-los negociar a forma de coleta e uso da informação. Para isso, pode ser utilizado o P3P[21], o *Privacy Bird*[22], ou mecanismos similares. O grande problema que essas políticas oferecem é que não há como garantir que os *sites* estão agindo de acordo com a política divulgada, o que implica que os usuários deverão confiar inteiramente nos *sites*, salvo se estiverem protegidos pelas duas camadas apresentadas a seguir.

3.3.6 Camada 5: Certificação de privacidade

Esta camada está associada à preocupação de se garantir que os *sites* estejam obedecendo às políticas de privacidade divulgadas. Para isso, a política de privacidade anunciada por um *site* deve ser periodicamente verificada por organizações de auditoria e grupos de privacidade. Estas organizações podem simplesmente fornecer aos *sites* um selo de garantia de qualidade ou uma nota. Estas notas podem ser baseadas na taxonomia proposta por *Wang et al.* para as preocupações do usuário (Seção 2.4) [60]: acesso impróprio, coleta imprópria, monitoramento impróprio, análise imprópria, transferência imprópria, transmissão não desejada e armazenamento impróprio.

Como esse processo pode estar muito além das possibilidades financeiras de muitos provedores de serviços da Web, *Cranor* [13] propõe o uso da tecnologia para automatizar o processo de auditoria, por exemplo, monitorando a propagação de dados.

Uma pesquisa recente enfatizou a importância das políticas de privacidade ao identificar que a grande maioria dos usuários da Web esperam ver e compreender as políticas de privacidade, quando visitam um *site* [19]. Entretanto, devemos ter um certo cuidado com as certificações de privacidade. Já foi divulgado na mídia, casos de venda de companhias, como a Toysmart.com[23], que incluíram a venda de dados dos clientes. O grande problema é que as empresas que adquirem esses conjuntos de dados não se sentem na obrigação de respeitar a política de privacidade da antiga companhia. Estes tipos de situações ressaltam a necessidade da sexta camada de proteção de privacidade.

[21] http://www.w3.org/P3P
[22] http://www.privacybird.com
[23] http://abcnews.go.com/sections/tech/DailyNews/toysmartftc000711.html

3.3.7 Camada 6: Leis que regulamentem a proteção de privacidade

As leis que regulamentam a proteção de privacidade variam de país para país, sendo que, em alguns países, elas nem existem. Até que essas leis existam, as empresas não terão muito incentivo em proteger e respeitar a privacidade dos usuários, principalmente porque os usuários nem sabem que sua privacidade pode estar sendo invadida.

Um problema central reside no fato de que não é possível controlar o comportamento na Web. Ao invés disso, os governos devem tentar regular os códigos ou o funcionamento das aplicações da Web (navegadores, sistemas de e-mails e outros) [42].

Uma outra dificuldade está em se conseguir um consenso internacional, porque o conceito de privacidade é extremamente dependente de questões políticas e culturais. Apesar dessas dificuldades, existe um conjunto de atividades que estas leis devem regular:

- usuários devem ser notificados sobre quais dados serão coletados e o objetivo do processamento destes;
- a coleta de dados só pode ser feita para um uso específico e todos os dados coletados devem ser necessários para o objetivo para quais serão usados;
- todo armazenamento de dados deve ter um tempo limite de armazenamento;
- dados não poderão ser repassados para terceiros;
- no caso de venda de empresas, obrigar os novos proprietários de uma determinada coleção de dados a estarem respeitando a política de privacidade da antiga empresa;
- usuários devem ter acesso aos dados coletados e, igualmente, devem poder atualizá-los ou removê-los;
- apesar dos *sites* serem registrados em um único país, sempre que forem acessados por usuários de outro país, deverão obedecer às restrições de coleta de dados do país do usuário.

3.3.8 Comentários adicionais

A literatura técnica não reporta ferramentas que implementem as seis camadas de proteção de privacidade, principalmente porque é difícil implementar as duas últimas camadas, por constituírem um compromisso da sociedade. Mas é possível projetar e implementar ferramentas que cubram várias dessas camadas. Para cobrir um maior número de camadas é importante que os pesquisadores da área estejam devidamente atentos à noção de consentimento relacionado à privacidade. Conforme nos traz *Roger Clarke* [11], "a proteção de privacidade é um processo de encontrar o balanceamento apropriado entre privacidade e múltiplos interesses competitivos". Portanto, uma arquitetura para proteção de privacidade na Web deve não somente oferecer recursos para proteção do usuário, mas é fundamental que essas ferramentas não impeçam os usuários de se beneficiarem de alguns serviços da Web, como a personalização.

3.4 Segurança X Proteção de privacidade

Segundo *Garfinkel* [25], há uma grande sobreposição entre os conceitos de segurança de sistemas e a privacidade de dados de usuários. Por exemplo, sempre que informações confidenciais forem transmitidas, deve-se utilizar um canal seguro. Para se ter uma certa garantia da privacidade de dados armazenados, deve-se utilizar mecanismos de segurança, como a criptografia e controle de acesso. Mas não está muito claro saber o que está sobreposto e o que não está. Na verdade, muitas pessoas confundem os dois conceitos. Entretanto, deve-se ter em mente que, dentre outras justificativas, se privacidade fosse igual a segurança, a OECD não teria lançado um documento específico para cada um dos dois tópicos: segurança[24] e privacidade[25]. Nesta seção, apresentaremos uma discussão que tem por objetivo esclarecer a distinção entre estes dois conceitos.

A segurança possui as seguintes características ([50], [37]):

1. Confidencialidade - somente usuários autorizados podem ler ou ter acesso a informações.

[24] *OECD Guidelines for the Security of Information Systems and Networks* - http://www.oecd.org/document/42/0,2340,en_2649_201185_15582250_1_1_1_1,00.html

[25] *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data* - http://www.oecd.org/document/20/0,2340,en_2649_201185_15589524_1_1_1_1,00.html

2. Integridade - somente usuários autorizados podem alterar/escrever informações.

3. Disponibilidade - um serviço está disponível sempre que for necessário ou, em outras palavras, um computador/rede é seguro se nada, nem ninguém, pode evitar que usuários autorizados acessem os serviços desejados.

4. Responsabilidade - é possível identificar a entidade responsável por cada ação.

Por outro lado, as questões que envolvem privacidade são muito mais amplas do que segurança e incluem aspectos políticos, culturais e sociais, além dos tecnológicos.

Agre [4] faz a distinção entre privacidade e segurança, da seguinte forma: "privacidade de informação significa que eu consigo controlar minhas informações pessoais. Segurança de dados significa que alguém, em alguma organização, consegue controlar minhas informações pessoais (...). O problema começa quando a própria organização deseja invadir minha privacidade, por exemplo, usando informações sobre as minhas transações para objetivos secundários. Esse uso secundário dos dados pode ser tão seguro quanto possível, mas mesmo assim constitui invasão de privacidade".

Para se ter uma melhor noção da distinção entre privacidade e segurança, podemos analisar algumas classificações existentes, relacionadas à privacidade e verificar de que forma se encaixam com a segurança. Iniciando pelos 8 (oito) princípios da OECD (Seção 2.3), podemos observar que somente o princípio da segurança está diretamente relacionado à segurança. Dos 5 (cinco) requisitos do TRUSTe (Seção 3.2.7), somente o transcrito a seguinte envolve segurança: "O *site* deve proteger os dados dos usuários, nos sentido de que eles não poderão ser perdidos, mal utilizados ou alterados sem autorização".

Das sete preocupações que um usuário deve ter com relação à sua privacidade (Seção 2.4), somente duas estão diretamente relacionadas à segurança, conforme mostrado a seguir:

- Acesso impróprio: essa preocupação pode ser diretamente tratada por metodologias de segurança, como o uso de *firewalls*, antivirus, filtragem de código executável, como *JavaScript*, programas *Flash* e *ActiveX*.

- Coleta imprópria: a segurança pode evitar, de forma eficiente, a coleta imprópria por terceiros, mas não do servidor do *site* que um usuário esteja pesquisando, pois os servidores Web são entidades autorizadas para acessar os dados dos usuários.

- Monitoramento impróprio: a segurança pode evitar o monitoramento impróprio por terceiros, mas não pelo servidor do *site* que o usuário está acessando.
- Análise imprópria: a segurança não tem como evitar essa prática, se os dados já estiverem nas mãos dos mineradores de dados.
- Transferência imprópria e transmissão não desejada: não há métodos seguros que evitem essa prática, que é uma questão cultural e dependente de leis que as controlem.
- Armazenamento impróprio: da forma como foi definida, essa preocupação é completamente dependente de segurança.

Para finalizar, apresentamos, sucintamente, uma lista das diferenças entre segurança e privacidade:

- Privacidade é um conceito pessoal, mas segurança, não.
- Nem todo problema de privacidade pode ser resolvido através de meios computacionais.
- A mineração de dados, em geral, determina uma invasão de privacidade mas não, necessariamente, falta de segurança.
- Dentre os oito princípios da OECD, somente o princípio da segurança tem relação com a segurança.
- Existem meios, como *cookies* e *Web bugs*, de se violar a privacidade de usuários, mesmo em computadores/redes seguras.

Portanto, está claro que privacidade não é sinônimo de segurança, nem tampouco um subconjunto desta. Na Web, não há como propor uma solução completamente independente de segurança, porque a proteção de dados que são transmitidos entre usuários e *sites* é uma das tarefas relacionadas à segurança de redes. Portanto, pode-se afirmar que a segurança é um meio auxiliar para obter privacidade.

Capítulo 4

MASKS: *Managing Anonymity while Sharing Knowledge to Servers*

Neste capítulo nós apresentamos as principais características da arquitetura do MASKS, cujo acrônimo significa **M**anaging **A**nonymity while **S**haring **K**nowledge to **S**ervers (figura 4.1).

4.1 Características de projeto

MASKS é uma arquitetura baseada no método de revelação seletiva [34]. A idéia básica desse método é colocar uma barreira entre os dados privados e o analista de dados e controlar as informações que podem atravessar esta barreira. Esse método minimiza a divulgação de dados pessoais sem impedir uma análise contínua desses dados.

O conceito chave do MASKS é o conceito de máscara. Uma máscara é uma identificação temporária que um usuário pode adotar enquanto estiver interagindo com um *site*. Essa identificação é associada a um usuário, de acordo com seu interesse em um tópico e *site* específico. Sempre que um usuário visita um *site*, ele pode utilizar uma máscara para interagir com o *site*, sem ser identificado. Da mesma forma que máscaras psicológicas (Seção 3.1.3), usuários podem ter, e em geral têm, diversas máscaras.

Esse esquema levanta várias questões. A primeira é garantir que o usuário tenha controle sobre suas informações pessoais. A segunda é como associar máscaras ao comportamento do usuário. A terceira diz respeito à compatibilidade com protocolos padrões

da Web. E, por fim, como oferecer o serviço de máscaras, sem aumentar o tempo de resposta percebido pelo usuário.

MASKS satisfaz os seguintes requisitos:

- Proteção de privacidade – MASKS aplica a anonimidade, através do uso de máscaras, como um mecanismo de proteção de privacidade.

- Compatibilidade parcial com o processo de personalização – ao contrário de outras ferramentas de privacidade, MASKS permite personalização porque disponibiliza dados que podem ser usados por *sites* Web para oferecer serviços personalizados, sem que seja possível criar um perfil individualizado de cada usuário.

- Segurança – quanto maior a quantidade de informação armazenada, maior a probabilidade de se tornar alvo de um ataque. Por isso, MASKS reduz esse risco, baseando o seu processamento somente na última requisição de cada usuário.

- Eficiência – os serviços oferecidos pelo MASKS devem ser eficientes, no sentido de que a latência percebida pelos usuários não deve ser maior que a existente, sem o uso do MASKS. Na verdade, os algoritmos implementados são eficientes, com relação ao tempo de resposta, porque o mecanismo aplicado é muito simples. Dessa forma, espera-se que os usuários não percebam qualquer atraso, quando estiverem usando MASKS.

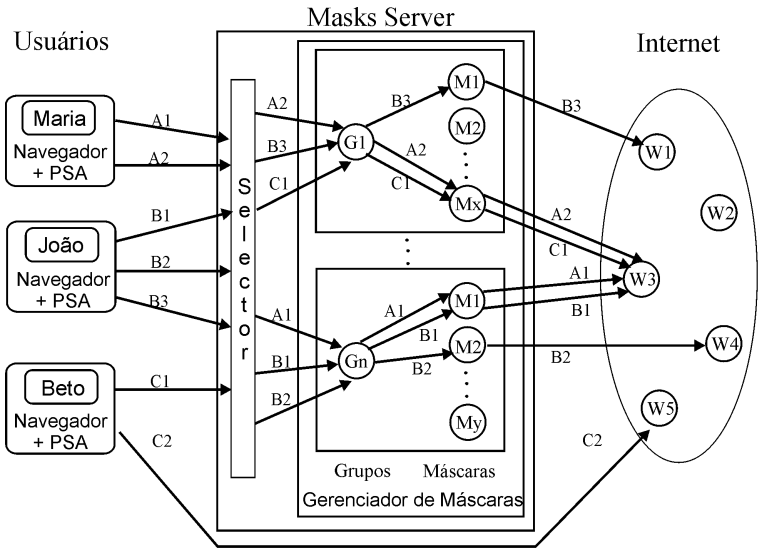
- Flexibilidade – os serviços do MASKS se adaptam dinamicamente a mudanças de comportamento do usuário. Perfil é o conjunto de interesses do usuário. Um fator importante na perfilização é ser capaz de adaptar a mudanças nos interesses do usuário no decorrer do tempo.

- Interoperabilidade e facilidade de implantação – MASKS deve empregar os protocolos padrões HTTP e TCP e trabalhar com os mecanismos usuais de identificação, como os *cookies*.

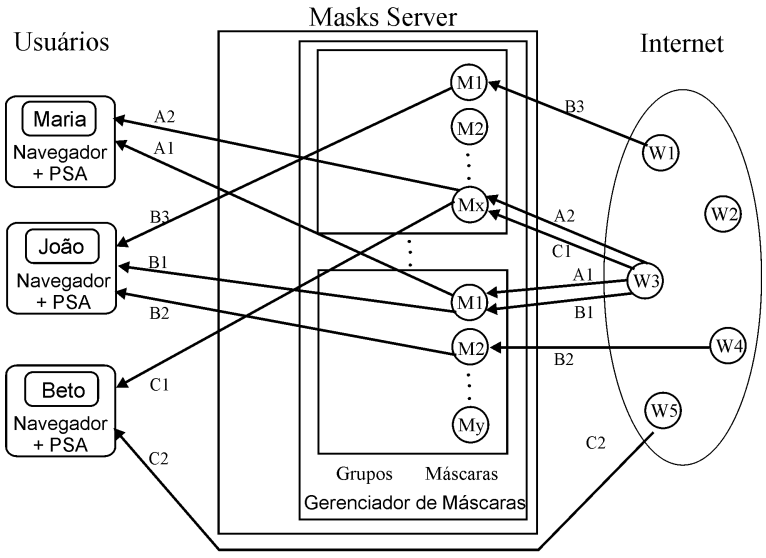
- Facilidade de uso – usuários não necessitam fornecer informações prévias ao MASKS.

- Proteção mais ampla da privacidade do usuários – considerando as seis camadas de proteção de privacidade – Seção 3.3 – (notificação, controle, ferramentas para

proteção de privacidade, políticas de privacidade, certificação de privacidade e leis que regulamentem a proteção de privacidade), MASKS é a única arquitetura que conhecemos que cobre as três primeiras camadas de proteção de privacidade (notificação, controle, ferramentas para proteção de privacidade).



a) Tratamento de requisições dos usuários



b) Tratamento de respostas dos sites

Figura 4.1: Arquitetura simplificada do MASKS

4.2 A arquitetura do MASKS

A arquitetura do MASKS possui dois componentes principais: o agente de privacidade e segurança (PSA - *Privacy and Security Agent*) e o servidor de máscaras (*Masks Server*).

O PSA é um programa que cada usuário executa em conjunto com o navegador. O PSA é um intermediário entre o *Masks Server* e os usuários, responsável por: cifrar as requisições dos usuários, manter o usuário informado sobre os seus riscos de ter sua privacidade invadida e sobre as máscaras que lhe estão sendo atribuídas; permitir que os usuários desliguem o processo de mascaramento, no caso deles preferirem interagir diretamente com os *sites*, sem anonimidade; bloquear e filtrar métodos conhecidos de invasão de privacidade, como os *cookies* de terceiros e os *Web bugs* (Seção 2.4). Devido ao seu conjunto de funções, o PSA oferece aos usuários as duas primeiras camadas de proteção de privacidade: notificação e controle (Seção 3.3).

O segundo componente é o *Masks Server*, que é o intermediário entre os usuários e os *sites* da Web, trabalhando como um *proxy*. O *Masks Server* é responsável pelo gerenciamento de máscaras e atribuição destas aos usuários. A atribuição de máscaras é baseada no conceito de grupo. Um grupo representa um tópico de interesse. Cada requisição de um usuário é associada a um grupo, de acordo com a semântica do objeto requisitado. Dessa forma, por trás das requisições teremos grupos, e não mais indivíduos. Essa característica do MASKS permite a divulgação de dados sobre os interesses dos usuários, sem que seja necessário identificá-los. Esses dados podem ser utilizados para oferecer serviços personalizados preservando, ao mesmo tempo, a privacidade da identidade do usuário.

Podemos distinguir dois componentes no *Masks Server*: o Selector e o gerenciador de máscaras. Ao *Selector* cabe a seleção do grupo de interesse de cada requisição do usuário. Ao gerenciador de máscaras cabe a tarefa de, dado um grupo, determinar a máscara correta para o usuário.

A figura 4.1 mostra todos os componentes do MASKS e exemplifica a interação entre os clientes e os *sites* da Web.

4.2.1 O processo de atribuição de máscaras aos usuários

Algumas interações anônimas estão ilustradas na figura 4.1-a). O processo inicia quando um usuário envia uma requisição cifrada pelo PSA ao *Masks Server*. Então, o *Selector* escolhe o melhor grupo para a requisição recebida para que, em seguida, o *Masks Server* possa enviar a requisição mascarada para o *site* da Web. A associação de máscaras é, portanto, realizada a cada requisição.

Há duas justificativas que reforçam essa opção de se trabalhar no nível de requisições. A primeira considera que, como um usuário pode demonstrar vários interesses durante uma única sessão, é mais simples caracterizar seus interesses de acordo com a semântica dos objetos requisitados. A segunda diz respeito à proteção de privacidade do usuário: como a informação principal será a requisição, o usuário não terá que disponibilizar informações adicionais. Além disso, não haverá a necessidade de armazenar dados dos usuários para identificação de grupo.

Conforme mostrado na figura 4.1-a), cada grupo poderá ter diversas máscaras associadas a ele, uma para cada *site* que ofereça o tipo de informação associada ao grupo. Por exemplo, há vários *sites* que oferecem informações sobre turismo. Portanto, o grupo associado ao tema turismo terá uma máscara para cada *site* de turismo conhecido. Na figura 4.1-a), esta situação é representada pelas requisições B1 e B2 de João.

É interessante observar que os *sites* da Web continuarão a ver as requisições de cada grupo como se fossem requisições comuns, provenientes de um único indivíduo. Por exemplo, na figura 4.1-a), o *site* W3 "achará" que as requisições A2 da Maria e C1 de Beto vieram de uma única pessoa e poderá oferecer serviços personalizados ao interesse do grupo. Entretanto, os *sites* não terão informações suficientes para criar um perfil de cada usuário, individualmente. Uma outra proteção da individualidade dos usuários é o fato de que usuários poderão possuir várias máscaras, mesmo enquanto estiverem navegando por um único *site*.

Suponhamos que o *site* W3, na figura 4.1-a), seja um portal que ofereça diferentes classes de informação, tais como turismo e investimentos. Suponhamos também que Maria requisite informações sobre serviços de turismo (A1) e, depois, sobre investimentos (A2). W3 verá as duas requisições de Maria (A1 e A2) como provenientes de dois usuários distintos, como conseqüência do fato de que virão de dois grupos diferentes.

Por fim, o MASKS também permite que usuários interajam diretamente com um *site*, conforme mostra a requisição C2 da figura 4.1-a).

Capítulo 5

PSA: *Privacy and Security Agent*

O *Privacy and Security Agent* (PSA), ou agente de privacidade e segurança, é um programa executado em conjunto com o navegador (*plugin*), que atua como intermediário entre os usuários, o *Masks Server* e os *sites* da Web. Conforme mostrado na figura 5.1, é o PSA que recebe as requisições dos usuários e as repassa ou ao *Masks Server* ou aos *sites* da Web, conforme o desejo dos usuários em estarem mascarados ou não. Também é o PSA que recebe as respostas e as transmite ao usuário, juntamente com uma avaliação dos riscos de invasão de privacidade do usuário. Devido ao seu conjunto de funções, o PSA oferece aos usuários as duas primeiras camadas de proteção de privacidade: notificação e controle. Nas seções que se seguem estaremos detalhando: as suas funções, de que forma atende às duas primeiras camadas de proteção de privacidade, as suas características relacionadas à segurança das informações dos usuários do MASKS.

5.1 Funções básicas

As funções do PSA são as descritas a seguir:

Cifrar URLs que trafeguem entre o PSA e o *Masks Server*. O objetivo desta função é evitar que terceiros conheçam informações privadas dos usuários, tais como: *sites* acessados, tópicos pesquisados, o endereço IP associado a um determinado conjunto de dados trasmitidos via formulários.

Manter o usuário informado sobre os seus riscos. *Ackerman & Cranor* [1] trazem

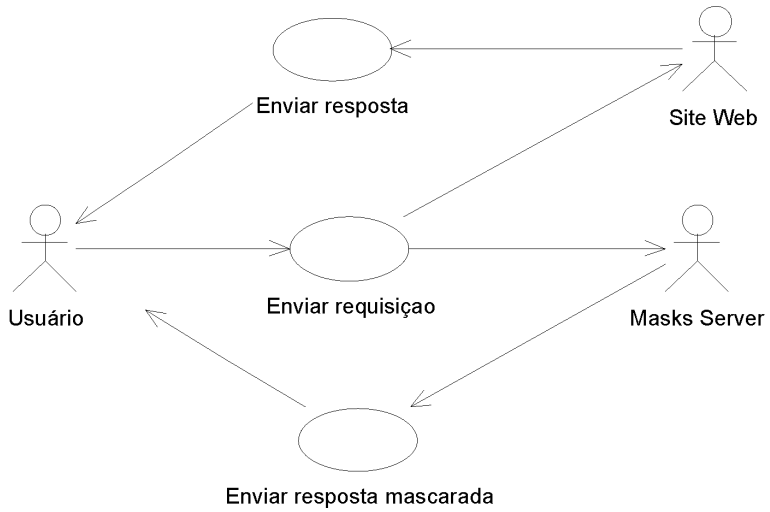


Figura 5.1: Caso de uso do PSA

que usuários se beneficiariam de sistemas que os ajudassem a identificar situações nas quais a privacidade estivesse em risco. Portanto, esta função do PSA atende a esta necessidade dos usuários da Web. Esta função também é importante para chamar a atenção das pessoas para o fato de que, ao trocarem de navegador ou de computador, as suas especificações de privacidade da máquina/navegador de origem se perdem.

Manter o usuário informado sobre as suas máscaras. A qualidade das informações divulgadas aos servidores é completamente dependente da precisão das máscaras associadas aos usuários. A fim de melhorar o grau de confiança dos usuários sobre as máscaras escolhidas, o PSA tem, como uma de suas funções, a interação com os usuários, de forma a deixá-los informados sobre os grupos associados a eles e permitindo-lhes escolher outro grupo, se preferirem, para interações futuras com um mesmo *site*. Essa estratégia endereça algumas questões práticas para personalização, como as levantadas por *Soltysiak & Crabtree*. Esses pesquisadores afirmam que "o perfilador não deve operar automaticamente sem mostrar seus resultados para os usuários e sem obter sua aprovação" [52]. Eles também colocam que os usuários devem ser capazes de revisar e corrigir seu perfil, o que indica que, na prática, o processo de mascaramento pode obter melhores resultados, com a ajuda do usuário.

Bloquear métodos conhecidos de invasão de privacidade. Algumas tecnologias, como os arquivos de históricos e os *scripts*, oferecem vantagens e desvantagens ao usuário. Outras, como os *Web bugs*, são reconhecidamente invasivas e não trazem qualquer benefício ao usuário. Portanto, esse tipo de tecnologia deve ser automaticamente bloqueado, para garantir ao usuário um nível maior de privacidade.

Permitir que os usuários desliguem o processo de mascaramento. Apesar do MASKS divulgar algumas informações para que os usuários possam ter acesso a serviços personalizados, o grau de personalização que lhes é ofertado não é o mesmo que poderiam obter através de interação direta com os *sites*. Portanto, é perfeitamente possível que os usuários prefiram interagir diretamente, sem anonimidade, com *sites* que considerem confiáveis. E o MASKS representaria um ataque à liberdade do usuário, se não lhe permitisse fazer essa opção.

Remover informações que permitam identificar o usuários. Esta função inclui a remoção de identificações dos pacotes de informações que serão submetidos pelos usuários, como, por exemplo, a página que estava sendo visitada.

5.2 Interface com o usuário

A interface com o usuário é um dos principais aspectos do PSA, pois é através dela que o usuário recebe serviços de proteção de privacidade da primeira camada de proteção (Seção 3.3.2). Pelas suas funções e características, o projeto da interface com o usuário também requer muito cuidado. *Cranor et al.* [14] nos trazem que "um dos maiores problemas de sistemas para controle de privacidade será o projeto de interfaces adequadas. Esses sistemas devem informar o usuário sempre que a sua privacidade estiver em risco. Entretanto, (...) isso deve ser feito de forma discreta". *Hochheiser* [32] complementa, dizendo que "sistemas de privacidade devem ser tão simples quanto possível, mas não simples demais. Evitar complexidade de projeto, implementação e interface com o usuário irá reduzir o risco de falhas e erros dos usuários". Portanto, a notificação dos usuários é uma função necessária, que deve ser realizada de forma discreta e simples.

Uma forma de simplificar a interface é através de um conjunto de um número pequeno de opções consistentes. É interessante dar ao usuário a opção de configurar o tipo de

interface que prefere ter.

Baseando-nos nesses requisitos, ficou definido que, inicialmente, a interface do PSA com o usuário teria as seguintes características:

- No momento em que o usuário executa um navegador, o PSA verifica a configuração deste e apresenta uma janela contendo uma avaliação dessa configuração. Esta mesma janela contém uma ligação para uma outra janela de configurações, na qual o usuário pode configurar como quer receber os avisos sobre sua privacidade dali para frente.

- O usuário pode optar por um dos seguintes formatos de recebimento de avisos:

  1. janelas *pop-up* contendo o aviso;
  2. diagnósticos de avaliação na barra de status, sob forma de ícone - neste caso, se o usuário estiver interessado em obter maiores informações, ele poderá clicar sobre o ícone e, somente após o clique, uma janela contendo informações adicionais irá se abrir na tela do usuário;
  3. nenhum aviso - os usuários devem ter o direito de optar por uma interface mais enxuta, se eles estiverem seguros com relação às suas práticas de navegação e configuração do navegador.

- O usuário pode, a qualquer momento, alterar a sua configuração de interface, ou seja: se não estava recebendo avisos, pode optar por passar a recebê-los e vice-versa; se estava interagindo via ícones, passar a janelas e vice-versa.

- Os avisos serão dados aos usuários, sempre que eles assim o desejarem e for possível identificar que um usuário está passando de um estado mais protegido para outro estado menos protegido. Por exemplo, quando um usuário está enviando um formulário preenchido ou, então, quando não aceitava *scripts* e passa a aceitá-los.

5.3 Arquitetura

Uma apresentação esquemática dos módulos que compõem o PSA é apresentada na Figura 5.2. Os componentes presentes estão descritos a seguir:

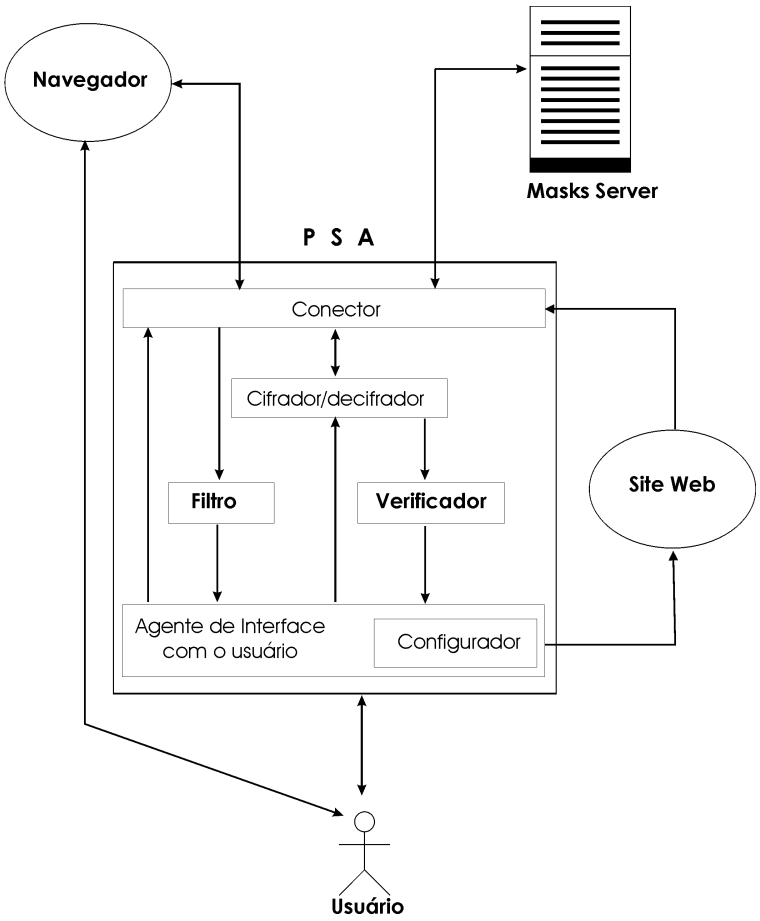


Figura 5.2: Arquitetura simplificada do PSA

Conector: Módulo responsável pela conexão entre o navegador e o *Masks Server*, atuando como intermediário. Em outras palavras, o conector realiza as comunicações entre o navegador e o PSA e entre o PSA e o *Masks Server*. Este módulo intercepta as requisições disparadas no navegador, processando-as internamente. Além disso, é responsável pelo repasse das respostas vindas do servidor de máscaras para o navegador.

Cifrador/decifrador: Módulo utilizado para cifrar e decifrar as URLs enviadas e recebidas do servidor de máscaras.

Filtro: Módulo responsável pelo filtro de informações do cabeçalho *HTTP* que possam conter algum tipo de identificação do usuário. Exemplos de informações a serem filtradas: a identificação do navegador que acompanha o cabeçalho *HTTP* e a página

na qual o usuário estava antes de efetuar a requisição atual (*referer*).

Verificador: Este módulo bloqueia mecanismos mais comuns utilizados para coletas de informações dos usuários, como, por exemplo, os *Web bugs*. Neste módulo do PSA, os documentos recebidos do MASKS também são avaliados para verificar se possuem formulários. Páginas que possuem formulários são, potencialmente, páginas que não podem ser mascaradas, uma vez que estas geralmente buscam enviar informações pessoais e identificáveis como nome, número do cartão de crédito, dentre outras, para o servidor Web. Porém, páginas de pesquisa como o *Google*, apesar de possuírem formulários, não se enquandram no perfil de páginas de coleta de dados e podem passar pelo processo de mascaramento. Sendo assim, o verificador deve prever esta situação e ser capaz de distinguir estes casos. A princípio, a estratégia utilizada se baseará na existência ou não de senha, no formulário. No caso de ser encontrado um pedido de senha, o usuário será comunicado e, com sua autorização, a requisição será transmitida diretamente ao Servidor Web.

Agente de Interface com o Usuário: Módulo responsável pela comunicação com os usuários do sistema. Através deste agente, os usuários recebem informações como o risco de invasão de privacidade que os mesmos estão correndo.

Configurador: O configurador compõe o Agente de Interface com o Usuário e, como seu próprio nome diz, possui a função de permitir ao usuário a configuração do PSA. Mais especificamente, este módulo permite que o usuário informe o tipo de interface desejada com o PSA e se quer ou não que suas sessões sejam mascaradas.

A Figura 5.3 apresenta o processamento de uma requisição do usuário, pelo PSA. Conforme pode ser visto na figura, uma requisição do usuário que chega ao PSA é filtrada. Qualquer problema identificado é comunicado ao usuário, que poderá sempre optar por continuar a sua navegação mascarado, ou não. Antes de enviar a requisição filtrada para o *Masks Server*, esta é cifrada.

A Figura 5.4 destaca o processo inverso, ou seja, o processamento de uma resposta enviada pelo *Masks Server*. Após receber uma resposta, esta é decifrada. Uma vez decifrada, a resposta passa por um processo de "limpeza", realizada pelo módulo *Verificador*. Neste processo de verificação de conteúdo, caso seja identificado algo que possa facilitar

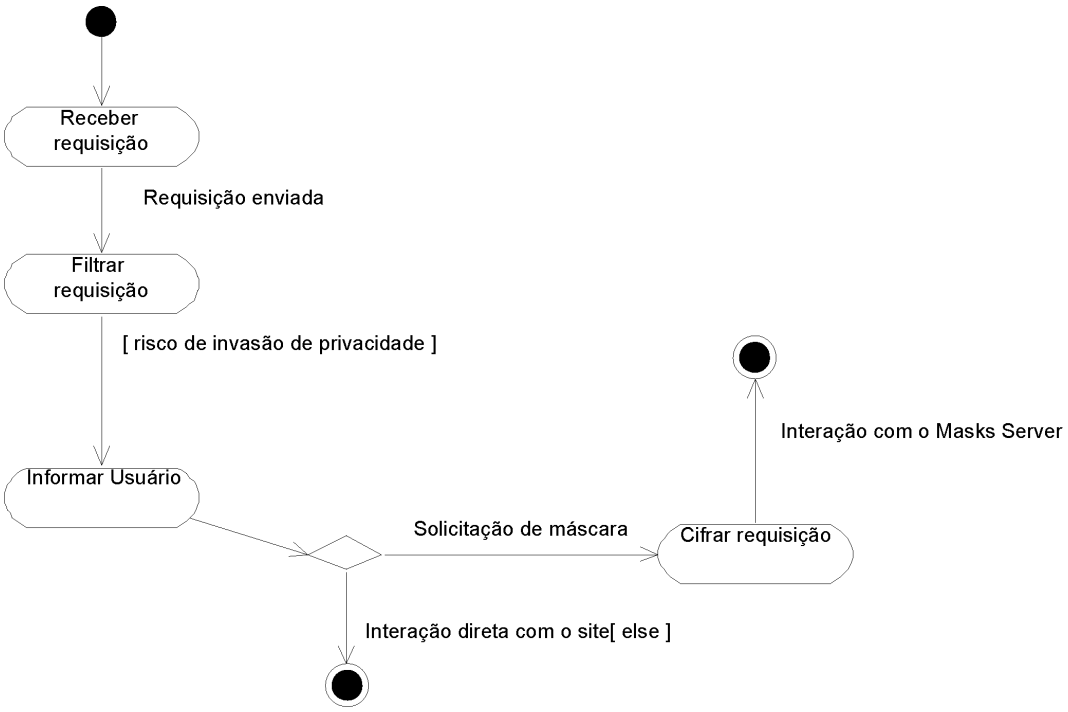


Figura 5.3: Processamento de uma requisição do usuário, pelo PSA

ou caracterizar invasão de privacidade, o usuário será notificado. Se a resposta satisfizer aos requisitos do PSA e do usuário, ela será transmitida ao navegador.

5.4 Implementação

A primeira versão do protótipo do PSA foi implementada para o navegador *Mozilla*. Isto se deve ao fato deste facilitar a criação de *plugin's* e interfaces de usuários através da linguagem conhecida como *XUL* (*XML-based User Interface Language*).

Os seguintes componentes foram implementados através de uma arquitetura conhecida como *XPCOM*: o conector, o filtro e, parcialmente, o agente de interface com o usuário. As ações sobre tais componentes são disparadas pela interface de usuário disponibilizada no navegador utilizado.

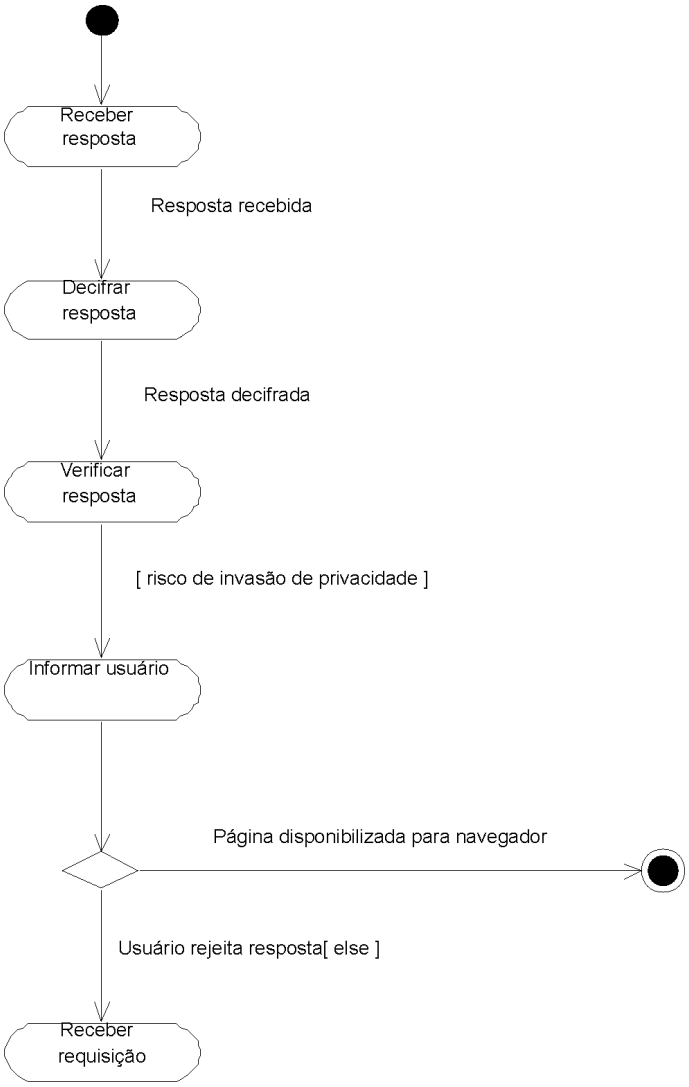


Figura 5.4: Processamento de uma resposta enviada pelo Masks Server

Capítulo 6

Masks Server

O segundo grande componente da arquitetura do MASKS é o *Masks Server*. O *Masks Server*, ou servidor de máscaras, é o intermediário entre o PSA e os *sites* da Web, trabalhando como um *proxy*. O *Masks Server* é responsável pelo gerenciamento de máscaras. Nas seções que se seguem, estaremos apresentando um detalhamento das principais características do *Masks Server*.

6.1 *Selector* e o algoritmo de seleção de grupo

O *Selector* é o componente do Masks Server, responsável por selecionar o grupo de interesse de cada requisição do usuário. Como as máscaras estão agrupadas, um ponto chave do *Selector* está na definição de grupos e de como os objetos estarão atribuídos a grupos. Na verdade, o algoritmo de seleção de grupo é, de certa forma, o ponto central de toda a arquitetura do MASKS. O objetivo do algoritmo é que o processo de seleção de grupo seja eficiente e semanticamente correto. Por semanticamente correto queremos dizer que todas as requisições associadas a um grupo estarão dirigidas a páginas que estão associadas a um mesmo tema. Dessa forma, todos os usuários que acessam um *site* da Web, utilizando uma determinada máscara, poderão receber recomendações adequadas ao seu interesse.

Para obter a simplicidade e a eficiência necessárias, os algoritmos tradicionais de mineração de dados e clusterização não são adequados, pois, necessitam de um volume grande de dados. A estratégia que adotamos se baseia no uso de uma árvore semântica, ou,

mais especificamente, a árvore de categorias definida pelo *Open Directory Projet*[1]. Essa árvore de categorias lista e organiza, semanticamente, uma parcela significativa de *sites* da Web. Essa árvore está disponível, sem custo algum, representando, portanto, um ponto de partida para definição de grupos e relacionamentos entre eles. A figura 6.1 exemplifica uma árvore de categorias.

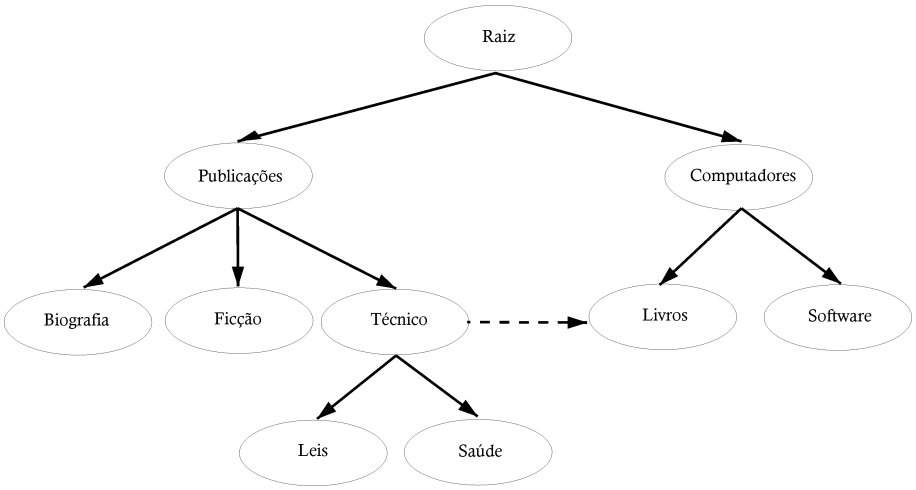


Figura 6.1: Exemplo de uma árvore de categorias

Cada nodo da árvore representa uma categoria semântica, ou um grupo, do nosso método. Um grupo é formado por um conjunto de páginas correlacionadas, um conjunto de termos que caracterizam o tópico e um conjunto de máscaras. Há uma máscara para cada *site* da Web encontrado dentre as páginas correlacionadas. Um grupo também pode ter conexões para outros nodos. Um filho de um grupo é uma especialização semântica do grupo, ou seja, cobre um tópico restrito da categoria semântica representada pelo nodo pai. Mas é possível, também, que um grupo esteja conectado a um outro grupo de uma outra subárvore. O objetivo dessas conexões, denominadas *links*, é fazer com que todos os caminhos que identifiquem um mesmo tópico acabem por apontar para um mesmo nodo. Devido a essas conexões, para atingir cada grupo, a partir da raiz, poderá haver um ou mais caminhos distintos. A figura 6.1 ilustra o relacionamento entre grupos. As arestas sólidas indicam o relacionamento entre pais e filhos. E as arestas tracejadas indicam *links*. Nessa figura, podemos observar que os nodos *Raiz* $\Rightarrow$ *Computadores* $\Rightarrow$ *Livros* e *Raiz* $\Rightarrow$

[1]http://dmoz.org

Publicações ⇒ *Técnico* ⇒ *Livros* se referem a um mesmo tópico e, por isso, o segundo possui um *link* para o primeiro.

6.1.1 Algoritmo

Esta seção apresenta o algoritmo para seleção do melhor grupo a ser utilizado para mascarar uma requisição. Como entrada de dados, esse algoritmo só necessita da requisição atual. Esse método oferece a vantagem de associar interesses dos usuários a grupos sem que seja necessário armazenar informações pessoais. E, por esse mesmo motivo, a solução proposta se adapta facilmente à natureza dinâmica da navegação dos usuários.

A idéia que está por trás do algoritmo de seleção de grupo é escolher o grupo que melhor reflita o tema da requisição do usuário. Isso é feito através de uma das seguintes ações, em ordem de prioridade:

1. determinar o grupo, de acordo com os termos da consulta, presentes na URL;
2. selecionar o grupo que indexa a URL, na árvore de categorias;
3. selecionar o grupo de acordo com algum termo existente na URL. Por exemplo, a URL *www.algum.com.br/esporte* indicaria interesse por esporte;
4. escolher o grupo raiz (*Root group*).

O algoritmo apresentado na figura 6.2 descreve como associar uma requisição a um dado grupo. Ele recebe como parâmetros a requisição *req* e a árvore de categorias *árvore*. Para compreensão do algoritmo, suporemos que um cliente envia uma requisição para a página *www.algum.com/tema*. O primeiro passo é separar os termos de consulta da URL, se existirem, e identificar o conjunto G de grupos da árvore que possuem o maior número de ocorrências dos termos de consulta. Se esse conjunto só possuir um grupo, então esse é o melhor. A título de exemplo, suponhamos que um usuário envie a requisição *www.foo.com?query=banco&credito*. Pesquisando pelos termo *banco*, encontramos os grupos *Credito* e *Bancos*. Pesquisando o termo *credito*, encontramos somente o grupo *Credito*. Obviamente, o grupo *Credito* possui o maior número de ocorrências de termos de consulta e, por ser o único grupo, é o escolhido para mascarar a requisição.

```
SelecionaGrupo (req, árvore)

G ← ϕ
Q ← termos de consulta de req
se Q ≠ ϕ então
  G ← {x ∈ árvore | x possui o maior número de ocorrências de termos de consulta}
  se G possui mais de um grupo então
    G' ← G
    G ← {i ∈ G' | graudeentrada (i) = maior grau de entrada}
    se G possui mais de um grupo então
        retorna o ancestral comum mais próximo dos nodos de G
    senão retorna G
    fim se
  senão retorna G
  fim se
senão
  site ← URL da requisição req
  se site ∈ árvore então
    G ← {x ∈ árvore | x possui site}
    se G possui mais de um grupo então
        retorna o ancestral comum mais próximo dos nodos de G
    senão retorna G
    fim se
  senão
    T ← termos da URL
    se T ≠ ϕ então
      retorna grupo associado ao tema semântico T
    senão
      retorna raiz da árvore
    fim se
  fim se
fim se
```

Figura 6.2: Algoritmo de seleção de grupo

É possível que, ainda sim, tenhamos dois ou mais nodos candidatos a representar o melhor grupo. Nesse ponto, o algoritmo tenta buscar uma generalização para os termos de consulta. Em outras palavras, o algoritmo procura por um nodo predecessor mais próximo de todos os grupos candidatos.

O segundo passo é verificar se a URL está presente na árvore de categorias. Se este for o caso e existir apenas um grupo relacionado à URL, este grupo será o selecionado. Caso exista mais de um grupo, o algoritmo retornará o ancestral comum aos grupos candidatos.

Se nem a URL, nem os termos de consulta da requisição estiverem presentes nas tabelas, então o grupo escolhido será aquele associado a algum termo da URL. No nosso

exemplo, seria o termo *tema*. Entretanto, ainda assim, em alguns casos, $G = \emptyset$. Esses casos ocorrem quando o MASKS não tem como determinar a semântica da página requisitada e, portanto, a única ação razoável será garantir a privacidade dos usuários, associando-os ao grupo raiz.

6.2 Estratégias contra ataques

Todo o tráfego entre o navegador do cliente e o anonimizador será cifrado. As requisições serão decifradas, ao atingir o *Masks Server*. Os resultados que retornarem serão novamente cifrados antes de serem repassados para os clientes.

Para evitar a correlação entre requisições que chegam e saem do *Masks Server*, o servidor de máscaras deverá enviar alguma requisição periodicamente. Também irá alterar a ordem das requisições, sempre que possível.

Convém relembrar que o PSA (Capítulo 5) faz uma limpeza do pacote que será transmitido ao *Masks Server*, ou seja, retira algumas informações, tais como a página Web ativa e sistema operacional utilizado. Esse procedimento aumenta a segurança e privacidade do usuário, pois mesmo que terceiros consigam ter acesso ao conjunto de informações que passam pelo *Masks Server*, a quantidade de informação a que terão acesso será bem menor.

6.3 Implementação

Por atuar como um *proxy*, para implementação de um protótipo do *Masks Server* foi utilizado o *Squid*[2]. *Squid* é o resultado do trabalho de inúmeras pessoas da comunidade da Internet, coordenados por *Duane Wessels* do *National Laboratory for Applied Network Research*. Esta arquitetura oferece as grandes vantagens do código aberto e de trabalhar como um *proxy* HTTP. Além dessa vantagem, ainda pode ser citado o fato deste servidor apresentar um considerável nível de escalabilidade e ser de utilização em larga escala. Um outro fato a ser ressaltado é com relação à facilidade em capturar o cabeçalho *HTTP*, o que permite processar os *cookies* enviados pelos servidores Web.

[2]http://www.squid-cache.org

6.3.1 Tratamento de *cookies*

A figura 6.3 exemplifica o funcionamento da arquitetura. A sequência disposta do lado esquerdo da figura representa uma sequência de requisições efetuada pelos usuários João e Maria a um servidor Web sem a existência de um *Masks Server*. Do lado direito, está representada a mesma sequência de requisições, porém com um *Masks Server* no caminho. Nas duas situações, o servidor Web retorna, junto com o conjunto resposta da requisição, um *cookie* para fins de personalização.

Na coluna da esquerda, como pode ser observado, os clientes estão recebendo diretamente um *cookie* associado à sua sessão. O servidor Web espera receber estes *cookies* nas requisições seguintes. O usuário João, ao enviar uma nova requisição ao servidor, encaminha o mesmo *cookie* **ABC** e recebe como resposta um documento contendo um novo *cookie* (**ABD**). Já com relação à requisição efetuada pela Maria, o servidor não fez nenhuma alteração no *cookie* **XYZ**. Como pode ser observado, o servidor Web pode alterar ou não os *cookies* que estavam armazenados nos clientes e que foram enviados por meio de novas requisições.

Com o *Masks Server*, esta situação é alterada. O servidor Web continua fazendo as mesmas operações sobre os *cookies* que eram feitas na situação anterior, porém, os usuários agora não perceberão mais essas alterações. O exemplo aqui exposto representa um conjunto de requisições de dois usuários que têm interesse no mesmo assunto. Isto faz com que os usuários João e Maria fiquem associados a um mesmo grupo de interesse. Dessa forma, o *Masks Server* efetua as requisições para o servidor Web como se fosse um único usuário com interesse específico. Como pode ser observado na figura 6.3, o usuário João efetua a requisição **RJ1**. O *Masks Server* irá repassar esta requisição como se fosse um usuário com interesse em, por exemplo, ficção científica. O servidor Web envia a resposta **rj1** com o *cookie* **ABC** associado. Quando Maria efetuar a sua primeira requisição (**RM1**), esta será repassada pelo MASKS ao servidor Web, porém já com o *cookie* **ABC** associado, dado que agora é como se o usuário único que tem interesse em ficção científica estivesse enviando uma segunda requisição para o servidor Web. Em seguida, o servidor Web envia a resposta **rm1** com um novo *cookie* associado **ABD**. Quando a segunda requisição (**RJ2**) efetuada por João chega ao *Masks Server*, este a repassa para o servidor Web com o *cookie* **ABD**. Esta situação pode se repetir indefinidamente.

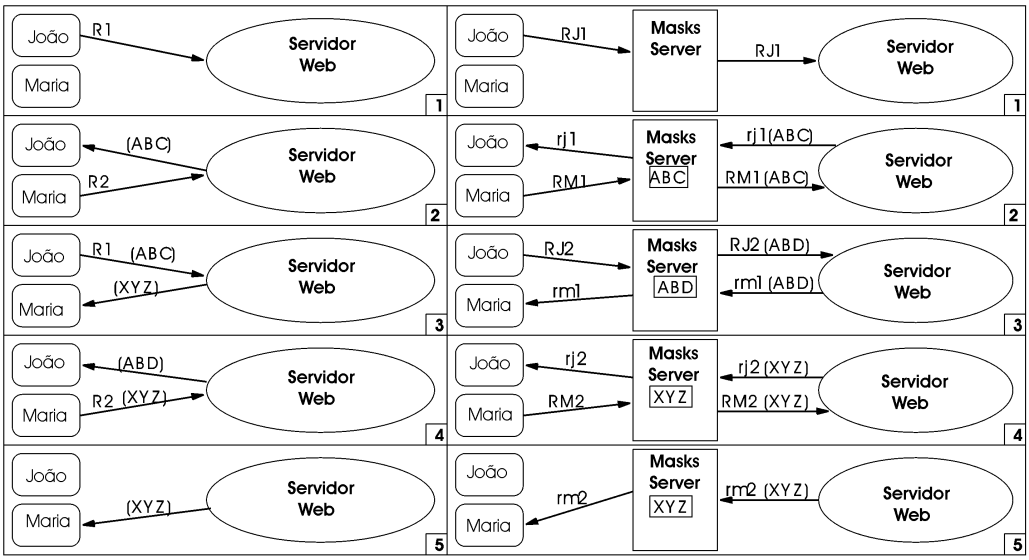


Figura 6.3: Sequência de requisições com/sem a presença de um servidor MASKS

Capítulo 7

Avaliação do MASKS

Nesta seção, é apresentada a avaliação qualitativa e quantitativa da arquitetura do MASKS. A avaliação qualitativa inclui uma análise quanto à sua aplicabilidade e suporte para proteção de privacidade e segurança, de acordo com os conceitos apresentados nos capítulos 2 e 3 deste trabalho. A avaliação quantitativa tem por objetivo verificar a qualidade de dados disponibilizados pelo *Masks Server*, a partir da aplicação da Teoria da Informação sobre um conjunto de requisições dos usuários.

7.1 Aplicabilidade

De acordo com uma pesquisa do *CyberDialogue* [16], 70% dos usuários não apreciam a idéia de estarem fornecendo informações pessoais quando estão simplemente pesquisando produtos, mas somente 24% dos usuários da Web consideram impróprio que um *site* solicite informações pessoais, se eles estiverem fechando uma transação. Usando argumento similar, deduz-se que os usuários do MASKS estarão satisfeitos porque poderão navegar e pesquisar por informações específicas, anonimamente, apesar de não poderem se esconder atrás de máscaras durante os processos de compra, pagamento e consultas mais específicas, como dados de conta corrente. O problema com esses processos é que eles necessitam de informações pessoais, como número de cartão de crédito, número da conta corrente e endereço de entrega de produto adquirido. E isso acontece não só na Web, como também no mundo real. Por um lado, isso pode parecer negativo, mas, por outro lado, é uma forma de proteger os próprios usuários, pois a anonimidade total pode facilitar e até

mesmo encorajar a prática de atividades criminais ou anti-sociais.

É importante enfatizar que o MASKS pode ser utilizado para recuperar informações da Web, anonimamente, de qualquer *site* e independente do fato do *site* fornecer serviço personalizado ou não. No pior caso, os usuários serão associados a máscaras do grupo da raiz da árvore de categorias, mas toda a requisição poderá ser mascarada.

7.2 Privacidade e segurança

O projeto do MASKS procurou respeitar as oito preocupações que o usuário deve ter com relação à sua privacidade (Seção 2.4), pois:

- o MASKS não acessa o computador do usuário, sem autorização;
- o MASKS não coleta informações dos usuários e, por conseqüência, não as armazena de forma insegura;
- toda a análise e monitoramento das atividades do usuário tem o seu consentimento, pois é o próprio usuário quem opta por utilizar o MASKS e seu processamento somente necessita da última requisição enviada;
- o MASKS não transfere informação para terceiros;
- o MASKS não transmite informações não solicitadas ao usuário.

O MASKS também reduziu a possibilidade de invasão de privacidade por parte de servidores. Ele remove informações privadas, como o *referer*, das requisições. É claro que não há como impedir que os servidores coletem informações dos usuários e, depois, as analisem e transmitam para terceiros. Mas, considerando que os servidores não terão como descobrir a identidade real dos "proprietários" das informações coletadas, então pode-se afirmar que MASKS protege a privacidade de usuários.

A fim de prover segurança, algumas estratégias tiveram que ser incorporadas ao MASKS. Para proteger os usuários, o PSA cifra todas as requisições. Para evitar análise de tráfego, as requisições devem ser reordenadas, de forma que não seja possível associar requisições enviadas pelos usuários para o *Masks Server* com as que são reenviadas do *Masks Server* para os servidores dos *sites*.

Para que a personalização possa ocorrer, *cookies* deverão ser aceitos e, por isso, MASKS os aceita. Contudo, estes não são repassados aos usuários e permanecem armazenados, no *Masks Server*, como máscaras.

Sempre que um usuário quiser disponibilizar alguma informação pessoal (nome, e-mail, número do cartão de crédito, etc), ele(a) terá que interagir diretamente com o *site* desejado, desabilitando o processo de mascaramento. Esse tipo de informação é individual e, portanto, não pode ser aplicado a todos os membros de um grupo. Nesses casos, como em qualquer outra ferramenta de anonimidade, MASKS não terá como proteger a privacidade do usuário. Entretanto, se um usuário aceitou fornecer suas informações pessoais para um *site*, então essa situação não pode ser caracterizada como invasão de privacidade, pois conta com a autorização do usuário [60]. Além disso, já foi demonstrado que os usuários gastam a maior parte do tempo navegando, pesquisando e lendo documentos, e que estas atividades fornecem a maior parte dos dados para análise do comportamento do usuário [46]. Logo, as principais atividades dos usuários da Web podem ser mascaradas e somente as informações disponibilizadas no momento do fechamento de uma transação não são suficientes para gerar um perfil dos usuários.

O processo de anonimização irá esconder várias informações, dentre elas destacamos as seguintes: URLs, histórico de navegação, o tipo de conteúdo baixado, navegador utilizado, endereço IP, sistema operacional utilizado pelo usuário.

7.3 Avaliação quantitativa

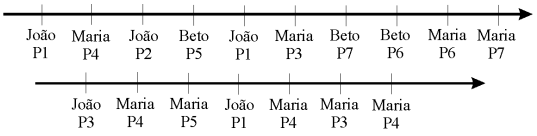
Esta seção apresenta a metodologia usada para avaliar a qualidade das informações disponibilizadas pelo MASKS para os *sites* da Web, bem como os resultados obtidos [36].

7.3.1 Metodologia

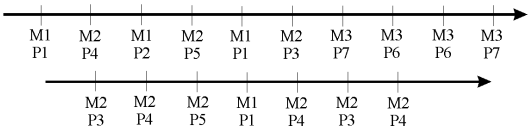
É muito difícil avaliar o quanto os dados divulgados pelo MASKS para os *sites* da Web irão afetar as estratégias de personalização adotadas por estes *sites*, pois há um número grande de técnicas de personalização que podem estar sendo utilizadas. Entretanto, é possível estimar o valor da informação disponibilizada, o qual, dado um conjunto determinado de requisições, será sempre o mesmo, independentemente da estratégia de personalização

Página	*Tema*	*Máscara*
P1	Esportes	M1
P2	Esportes	M1
P3	Finanças	M2
P4	Finanças	M2
P5	Finanças	M2
P6	Turismo	M3
P7	Turismo	M3

a) Páginas e suas respectivas máscaras



b) Ordem em que as requisições originais chegam ao servidor do site



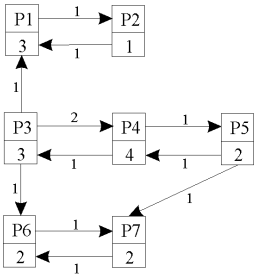
c) Ordem em que as requisições mascaradas chegam ao servidor do site

Usuário	*Sessão*	*Entropia da sessão*
João	P1, P2, P1	0,58
João	P3, P1	1,42
Maria	P4, P3, P6, P7	0,74
Maria	P4, P5, P4, P3, P4	0,95
Beto	P5, P7, P6	0,71

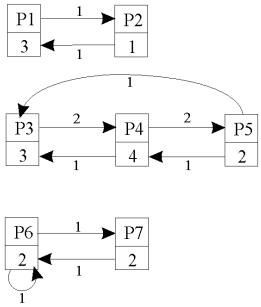
d) Sessões originais e suas entropias

Usuário	*Sessão*	*Entropia da sessão*
M1	P1, P2, P1	0,58
M1	P1	1,73
M2	P4, P5, P3	0,85
M2	P3, P4, P5, P4, P3, P4	0,66
M3	P7, P6, P6, P7	0,88

e) Sessões geradas por requisições mascaradas e suas entropias



f) Modelo de navegação para as requisições originais
(Entropia do modelo = 1,37)



g) Modelo de navegação para as requisições mascaradas
(Entropia do modelo = 1,17)

Figura 7.1: Modelos e sessões originais e mascaradas

utilizada por cada *site*. De acordo com a Teoria da Informação, a medida da quantidade de informação que uma variável contém é dada pela sua *entropia* [12]. A entropia H(X) de uma variável aleatória discreta X também pode ser entendida como a medida de incerteza de uma variável aleatória e é definida como:

$$H(X) = -\sum_{x \in X} p(x) log\, p(x)$$

Essa fórmula nos diz que, se tivermos dois experimentos X e Y, com as seguintes distribuições de probabilidade,

$$X = \begin{cases} 1, \textit{ com probabilidade } 0,5 \\ 2, \textit{ com probabilidade } 0,5 \end{cases} \quad Y = \begin{cases} 1, \textit{ com probabilidade } 0,99 \\ 2, \textit{ com probabilidade } 0,01 \end{cases}$$

então, X é muito mais incerto do que Y, pois, no caso de Y, podemos praticamente afirmar que o resultado será 1, enquanto que no caso de X, não será possível fazer nenhuma predição [38].

No nosso caso, como se deseja descobrir se uma sessão ou seqüência de requisições mascaradas irá disponibilizar informação de valor para os *sites* da Web, a entropia estará relacionada à quantidade de informação presente em cada seqüência de requisições. Quanto menos provável for a ocorrência de uma determinada seqüência de requisições, maior será a sua entropia. E, vice-versa, quanto mais provável for a ocorrência de uma seqüência de requisições, menor será sua entropia.

Para modelar um conjunto de sessões, pode-se utilizar de um grafo dirigido, no qual os nodos representam páginas e as arestas, transições de páginas. Cada aresta terá, associado a ela, o número de ocorrências da transição que representa, o que permitirá o cálculo da probabilidade de ocorrência de cada transição. Um grafo deste tipo caracteriza um modelo conhecido por Modelo de Markov [38]. De acordo com a teoria proposta por *Levene & Loizou* [43], o Modelo de Markov permite calcular: probabilidade de uma seqüência de requisições; entropia de uma sessão; entropia do modelo de Markov. A Figura 7.1 ilustra a modelagem adotada para o nosso experimento e o cálculo das entropias.

Na Figura 7.1(a), apresentamos os temas básicos de sete páginas de um *site* fictício W. Para cada um desses temas, o MASKS irá associar uma máscara diferente, conforme indicado. Cada uma dessas páginas tem uma probabilidade geral de ocorrência, calculada da seguinte forma: *p(página)= número de requisições da página / número total de requisições de todas as páginas.* No caso do exemplo da Figura 7.1, podemos verificar através da Figura 7.1(b), que o número total de requisições de páginas é 17 e que o número de requisições da página P1 é 3. Portanto, $p(P1) = 3/17 \approx 0,176$.

Na Figura 7.1(b), apresentamos uma seqüência de requisições que chega ao servidor do *site* W, em dois dias distintos, incluindo o usuário que as enviou. Na Figura 7.1(c), apresentamos essa mesma seqüência, só que com as requisições mascaradas. As Figuras 7.1(d) e

7.1(e) apresentam tabelas que detalham as requisições que compõem sessões e as entropias dessas sessões para ambos os casos: original e mascarado. Para calcular as entropias apresentadas, necessitamos dos modelos de Markov apresentados nas Figuras 7.1(f) e 7.1(g). Esses modelos são criados a partir das sessões apresentadas nas Figuras 7.1(d) e 7.1(e). Cada nodo do modelo representa uma página e contém o número de vezes que a página foi acessada. Associado a cada aresta (ou transição de página) apresentamos o número de vezes que esta transição ocorreu. Como o processo de mascaramento altera as sessões, o modelo de navegação para requisições mascaradas terá um formato diferente do modelo original, levando a um valor diferente de entropia do modelo.

O cálculo das entropias das sessões proposta em [43], foi realizado da forma detalhada a seguir. Primeiramente, se calcula a probabilidade $p(S)$ de ocorrência da sessão, que é dada por: $p(S) = p_{s_1}p_{s_1s_2}p_{s_2s_3}...p_{s_{t-1}s_t}$, onde cada s_i é um nodo do modelo de Markov, t é o comprimento da sessão e os $p_{s_is_{i+1}}$ representam a probabilidade de transição de um nodo s_i para o nodo s_{i+1}. A probabilidade de transição de um nodo s_i para o nodo s_{i+1} é dada por: $p_{s_is_{i+1}}$ = (*número de ocorrências da transição* s_is_{i+1}) / (*número total de transições que partem de* s_i). No exemplo da Figura 7.1(g), a probabilidade de transição da página P4 para a página P5 é: $p(P4P5) = 2/3 \approx 0,66$ e da página P5 para a P3 é $p(P5P3) = 1/2 = 0,5$. Como de acordo com o exemplo mostrado na Figura 7.1(a), $p(P4) = 0,235$, a probabilidade de ocorrência da seqüência [P4, P5, P3] será igual a: $p(P4P5P3) = p(P4) \times p(P4P5) \times p(P5P3) = 0,235 \times 0,66 \times 0,5 \approx 0,078$. De posse do valor da probabilidade de ocorrência da seqüência de requisições, torna-se possível calcular a entropia da seqüência, que é aproximadamente $-log(p(S))/t$. No caso do modelo do exemplo da Figura 7.1(g), a entropia de $p(P4P5P3) = -log(0,078)/3 \approx 0,85$.

O cálculo da entropia geral do modelo de Markov permite verificar o quanto o conjunto de sessões que o compõem é previsível. Um valor de entropia alto para o modelo indica que os usuários do *site* não seguem um padrão de navegação. O valor da entropia $H(M)$ do modelo M pode ser calculado de forma aproximada através da seguinte fórmula:

$$H(M) \approx -\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{m_{i,j}}{n}log\frac{m_{i,j}}{m_i},$$

onde n é o número total de páginas que compõem o modelo, $m_{i,j}$ é o número total de transições da página i para a página j e m_i é o número total de requisições da página i. Considerando o modelo da Figura 7.1(f), a sua entropia calculada por meio da fórmula

apresentada será aproximadamente $1,37$.

7.3.2 Resultados

A avaliação da qualidade de informação disponibilizada pelo MASKS foi realizada através do uso de *logs* reais de uma livraria virtual, coletados durante sete dias. Nós escolhemos os *logs* de um único *site* da Web, ao invés de um *log* mais genérico, como o armazenado por um proxy, porque queríamos avaliar o valor dos dados disponibilizados para personalização sob o ponto de vista de um *site* da Web. Dessa forma, para identificar se o processo de personalização será muito afetado pelo uso de máscara, estuda-se as diferenças entre as sessões originais e as sessões mascaradas que chegam a um determinado *site*.

Foi feita a simulação das requisições para os seguintes casos:

1. sessões originais dos usuários;

2. sessões mascaradas, considerando diferentes níveis da árvore semântica - as sessões são compostas por um conjunto de requisições cujo campo de identificação de cliente foi substituído por uma máscara de grupo. Para o processo de associação de grupo, consideramos, inicialmente, de um a cinco níveis da árvore semântica e, posteriormente, a árvore semântica completa. O objetivo era avaliar o quanto o fato de estarmos considerando um número maior ou menor de nodos, ou grupos semânticos, poderia estar influenciando nas novas sessões geradas;

3. sessões anonimizadas da forma padrão - neste caso, cada sessão conterá somente uma requisição. Esta modelagem foi feita com o objetivo de avaliar o resultado de outros mecanismos de anonimização, nos quais cada requisição é enviada com um identificador de usuário diferente e, portanto, cada sessão terá somente uma requisição.

A Tabela 7.1 apresenta a entropia de cada um dos modelos de Markov construídos. Conforme esperado, a entropia do modelo gerado por requisições mascaradas (árvore completa) é diferente da entropia do modelo original, mas comprova que as requisições geradas pelo MASKS oferecem algum valor informacional para um serviço de personalização, ao contrário do processo típico de anonimização, que gera modelos que disponibilizam informação de pouco valor para os *sites*. O seguinte fato justifica o resultado obtido: no

Modelo	Entropia do modelo
Original	8,01
MASKS (árvore completa)	4,72
MASKS (árvore 5-níveis)	5,31
MASKS (árvore 4-níveis)	5,97
MASKS (árvore 3-níveis)	7,20
MASKS (árvore 2-níveis)	7,79
MASKS (árvore 1-nível)	8,47
Anonimização	0

Tabela 7.1: Entropia de cada modelo

Modelo	Num. IDs	Num. sessões	Comp. médio da sessão
Original	227.047	242.990	1,52
MASKS (árvore completa)	12.105	166.345	2,22
MASKS (árvore 5-níveis)	9.996	157.436	2,35
MASKS (árvore 4-níveis)	8.780	147.459	2,50
MASKS (árvore 3-níveis)	7.912	138.426	2,67
MASKS (árvore 2-níveis)	7.580	131.858	2,80
MASKS (árvore 1-nível)	7.517	127.325	2,90
Anonimização	369.832	369.832	1

Tabela 7.2: Características de cada modelo

caso da anonimização, obtemos sessões de uma única requisição e, portanto, não há como ter uma visão mais geral do interesse dos usuários e nem, tampouco, transições disponíveis para cálculo da entropia do modelo.

Ainda analisando os resultados apresentados na Tabela 7.1, observamos que a diferença entre as entropias calculadas pelos dois modelos distintos, quando o número de níveis considerados para mascaramento das requisições dos usuários forem consecutivos, serão muito próximos. Isso ocorre porque a árvore semântica é muito larga e, por isso, não haverá grande variação no número de grupos disponíveis entre dois níveis próximos da árvore semântica. Além disso, a entropia do modelo da árvore completa é apenas 10% menor do

que a entropia do modelo gerado considerando apenas 5 níveis da árvore semântica. Esse resultado merece atenção, porque mostra que não é necessário considerar toda a árvore semântica para o processo de mascaramento de requisições. Dessa forma, o *Masks Server* pode economizar espaço de memória para armazenar os grupos semânticos.

Nota-se também que, quanto maior o número de níveis da árvore semântica considerados para mascaramento das requisições, menor a entropia do modelo gerado. Tal fato acontece porque quanto maior o número de grupos semânticos considerados para classificação das requisições, menor será o número de páginas associadas a cada grupo e, portanto, mais repetitivas (ou previsíveis) serão as transições entre elas.

A Tabela 7.2 apresenta algumas informações adicionais sobre cada um dos modelos de Markov construídos. A coluna **Num. IDs** representa o número de identificações de usuários distintas; **Num. sessões**, o número total de sessões geradas pelo modelo e **Comp. médio da sessão**, o comprimento médio das sessões geradas. Confirmando o resultado esperado, o número de sessões e de identificações distintas geradas pelo MASKS é menor do que os gerados pelo conjunto de requisições originais, porque, de acordo com o processo de mascaramento, as requisições originais serão semanticamente agrupadas. Esse resultado é muito bom, porque quanto maior o número de usuários que compõem um grupo, maior o grau de proteção que estará sendo atribuído a eles, pois será mais difícil definir o perfil real de um usuário específico.

Um outro ponto a ressaltar a partir dos resultados apresentados pela Tabela 7.2, é que estes reforçam os resultados da Tabela 7.1, pois demonstram que o processo típico de anonimização e o processo de mascaramento de todas as requisições como se fossem provenientes de um único usuário não constituem boas estratégias. Esses processos representam dois extremos de abordagem da utilização de anonimidade: o primeiro gera uma identificação de usuário distinta para cada requisição e o segundo, uma única identificação para um conjunto grande de requisições. Com certeza, ambos protegem a identidade real do usuário. Contudo, no primeiro caso, não será possível obter uma visão mais ampla do interesse dos usuários, pois todas as sessões terão uma única requisição. No segundo, o elevado número de requisições que chegarão ao servidor do site, como sendo provenientes de um único usuário, podem vir a confundir o servidor, ao invés de ajudá-lo no processo de personalização de serviços.

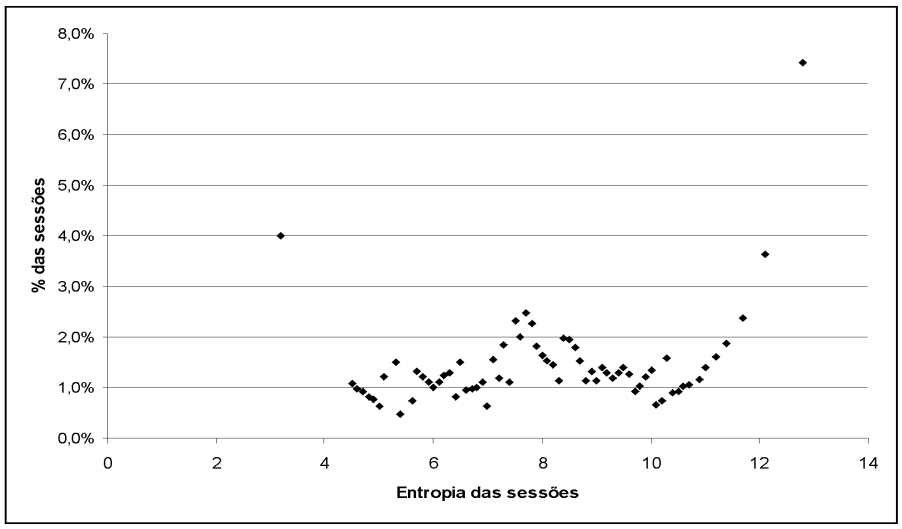


Figura 7.2: Entropia das sessões anonimizadas

A seguir, iremos analisar a entropia das sessões geradas pelos diversos modelos. Como os resultados obtidos a partir das sessões mascaradas foram muito similares, a partir desse momento, estaremos utilizando para análise, somente o gráfico gerado pelo modelo da árvore completa.

A Figura 7.2 apresenta a distribuição das entropias das sessões geradas por um conjunto de requisições tipicamente anonimizadas e a Figura 7.3, uma comparação entre a distribuição das entropias das sessões originais e as mascaradas. Conforme podemos observar, a distribuição das entropias das sessões geradas por um conjunto de requisições tipicamente anonimizadas tem uma curva muito diferente da gerada pelas sessões originais. Por outro lado, a distribuição das entropias das sessões originais possui algumas características em comum com a distribuição das entropias das sessões mascaradas. Por exemplo, com raras exceções, não há uma concentração muito grande de sessões com um determinado valor de entropia. Para quase todos os valores possíveis de entropia, o número de sessões que possuem um determinado valor de entropia não ultrapassa a 4% do total de sessões. Tanto as sessões originais quanto as mascaradas possuem um número pequeno de sessões com baixa entropia.

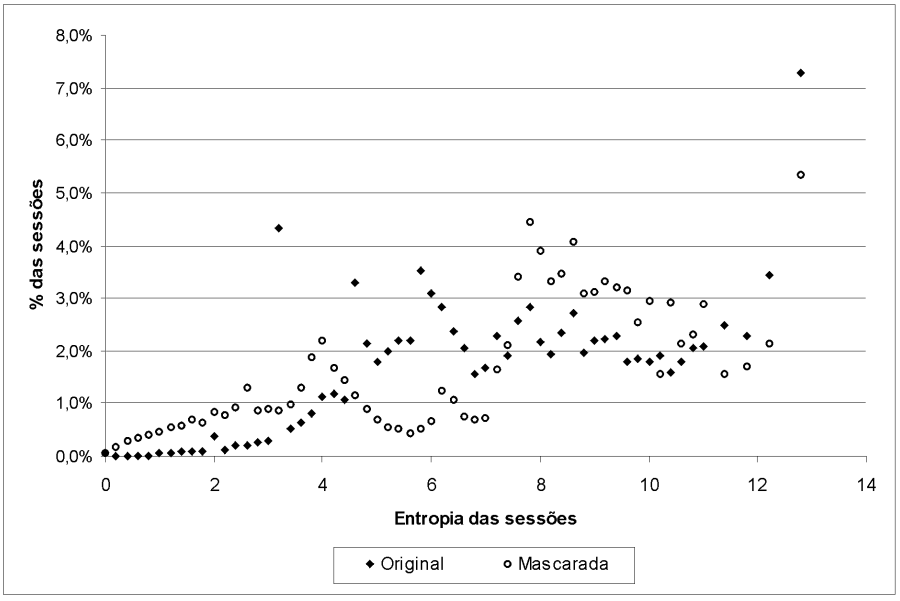


Figura 7.3: Entropias das sessões originais e mascaradas (árvore completa)

Capítulo 8

Conclusões e Trabalhos Futuros

A privacidade está se tornando uma das grandes questões levantadas pela sociedade moderna.

Este trabalho apresentou os principais aspectos relacionados à privacidade na Web, o que incluiu a proposta de uma taxonomia de camadas de proteção de privacidade de usuários da Web (Seção 3.3). Em seguida, foi descrita a arquitetura do MASKS. Dentre as suas características, destacam-se as seguintes:

- Compatibilidade para receber um serviço parcial de personalização – apesar de ser um mecanismo de anonimidade, MASKS filtra e altera as requisições dos usuários de uma forma que permite que os *sites* da Web continuem a receber informações para a oferta de serviços personalizados.

- Eficiência – os algoritmos implementados são eficientes, com relação ao tempo de resposta, porque, apesar das estruturas de dados serem complexas, o mecanismo aplicado é muito simples.

- Facilidade de uso – os usuários do MASKS não necessitam fornecer nenhuma informação prévia e o MASKS se adapta automaticamente à mudança de interesse dos usuários;

- Facilidade de implantação – MASKS não necessita de nenhum protocolo especial, pois o MASKS se baseia nos protocolos e serviços padrões da Web;

- Proteção mais ampla da privacidade do usuários – considerando as seis camadas de proteção de privacidade apresentadas na seção 3.3 (conscientização, controle, ferramentas para proteção de privacidade, políticas de privacidade, certificação de privacidade e leis que regulamentem a proteção de privacidade), MASKS é a única arquitetura que conhecemos que cobre as três primeiras camadas.

Como trabalhos futuros, deve-se incluir a busca por códigos mais eficientes na associação de máscaras a requisições de usuários, a avaliação de outras estratégias de classificação de *sites* e a necessidade de envolver especialistas da área de Interface Homem-Máquina, no projeto e desenvolvimento de um protótipo completo e mais avançado do PSA.

Na verdade, o desenvolvimento de um protótipo completo do MASKS, seguido de avaliações experimentais mais detalhadas, trariam ótimos resultados para fortalecimento da proposta ou para implantação de melhorias no projeto. Essas avaliações experimentais devem incluir a avaliação do desempenho do *Masks Server*, no caso de haver um número elevado de usuários utilizando o servidor simultaneamente, e a busca por pontos de vulnerabilidade do MASKS, tanto do ponto de vista de privacidade, quando de segurança. Também é importante procurar verificar a aplicabilidade do MASKS em contextos diversos, como no caso da computação móvel.

Dentre as metodologias que podem vir a ser aplicadas para avaliação do MASKS, encontra-se a teoria de jogos. Para isso, faz-se necessária uma melhor compreensão desta teoria para poder adaptá-la ao contexto do trabalho: resolução do conflito entre privacidade e personalização.

Um outro projeto futuro é a implementação de vários servidores de máscaras que atuem de forma cooperativa, trocando informações sobre os grupos semânticos já cadastrados, tornando mais preciso o processo de mascaramento.

Destacamos, igualmente, a necessidade de uma avaliação da privacidade no contexto cultural brasileiro, para que o MASKS, através do PSA, possa estar adaptado à nossa realidade.

Portanto, a arquitetura proposta nesta tese, pode ser ponto de partida para diversos projetos de pesquisa.

Bibliografia

[1] Mark S. Ackerman and Lorrie Faith Cranor. Privacy critics - safeguarding users' personal data. *Web Techniques*, September 1999. http://www.webtechniques.com/archives/1999/09/ackerman.

[2] Mark S. Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: Examining user scenarios and privacy preferences. *Proc. of ACM Conference on Electronic Commerce*, pages 1–8, 1999.

[3] Eytan Adar and Bernardo A. Huberman. A market for secrets. *First Monday*, 6(8), August 2001. http://www.firstmonday.org/issues/issue6-8/adar/index.html.

[4] Phil Agre. Strange ideas about privacy. *The Network Observer*, 1(10), October 1994. http://dlis.gseis.ucla.edu/people/pagre/tno/october-1994.html.

[5] Annie I. Antón, Julia B. Earp, Davide Bolchini, Qingfeng He, Carlos Jensen, and William Stufflebeam. The lack of clarity in financial privacy policies and the need for standardization. Technical Report TR-2003-14, North Carolina State University, august 2003.

[6] Paola Benassi. TRUSTe: an online privacy seal program. *Communications of the ACM*, 42(2):56–59, february 1999.

[7] Stanley Benn. Philosophical dimensions of privacy. In F. D. Schoeman, editor, *Privacy, Freedom, and Respect for Persons*, pages 223–44. Cambridge University Press, 1984.

[8] Mark Bilezikjian, John C. Tang, James Begole, and Nicole Yankelovich. Exploring web browser history comparisons. In *Conference on Human Factors in Computing Systems (CHI 2002)*, number 96-08, pages 828–829, Minneapolis, April 2002.

[9] Ann Cavoukian. Data mining: Staking a claim on your privacy. Technical report, Information and Privacy Commissioner/Ontario, January 1998. http://www.ipc.on.ca/english/pubpres/papers/datamine.htm.

[10] Roger Clarke. The digital persona and its application to data surveillance. *The Information Society*, 10(2), june 1994. http://www.anu.edu.au/people/Roger.Clarke/DV/DigPersona.html.

[11] Roger Clarke. Introduction to dataveillance and information privacy, and definitions of terms. *The Information Society*, september 1999. http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html.

[12] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[13] Lorrie Faith Cranor. The role of technology in self-regulatory privacy regimes. *National Telecommunications and Information Administration*, december 1996.

[14] Lorrie Faith Cranor, Joseph Reagle, and Mark S. Ackerman. Beyond concern: Understanding net users' attitudes about online privacy. Technical Report TR 99.4.3, AT&T Labs-Research, april 1999. http://www.research.att.com/library/trs/TRs/99/99.4.

[15] Matt Curtin, Paul Graves, and Shaun Rowland. Getting to know you (intimately): Surreptitious privacy invasion on the e-commerce web. Technical report, Interhack Corporation, july 2000. http://www.interhack.net/pubs/intimately.

[16] CyberDialogue. American internet user survey: Privacy x personalization - part I, 1999. http://www.cybersitter.com.

[17] Melissa Dunn, James Gwertzman, Andrew Layman, and Hadi Partovi. Privacy and profiling on the web. Technical note, World Wide Web Consortium, June 1997. http://www.w3/org/TR/NOTE-Web-privacy.html.

[18] Esther Dyson. Privacy protection: Time to think and act locally and globally. *Release 1.0*, April 1998. http://www.edventure.com/release1/0498.html.

[19] J. B. Earp and D. Baumer. Innovative web use to learn about consumer behavior and online privacy. *Communications of ACM*, 46(4):81–83, april 2003.

[20] Dag Elgesem. Privacy, respect for persons, and risk. In Charles Ess, editor, *Philosophical perspectives on computer-mediated communication*, chapter 3, pages 45–66. State University of New York Press, 1996.

[21] Federal Trade Commission. Privacy online: Fair information practices in the electronic marketplace, May 2000.

[22] Edward W. Felten and Michael A. Schneider. Timing attacks on web privacy. *ACM Conference on Computer and Communications Security*, pages 25–32, 2000.

[23] Susannah Fox, Lee Rainie, John Horrigan, Amanda Lenhart, Tom Spooner, and Cornelia Carter. Trust and privacy online: Why americans want to rewrite the rules. Technical report, The Pew Internet & American Life Project, august 2000.

[24] Charles Fried. Privacy. In F. D. Schoeman, editor, *Philosophical dimensions of privacy*. Cambridge University Press, 1984.

[25] Simson Garfinkel. *Web Security, Privacy & Commerce*. O'Reilly, 2nd edition, january 2002.

[26] B. Garvish and J. H. Gerdes Jr. Anonymous mechanisms in group decision support systems communication. *Decision Support Systems*, 23(4):297–328, 1998.

[27] Ian Goldberg, David Wagner, and Eric Brewer. Privacy-enhancing technologies for the internet. *Proc. of IEEE Spring COMPCON*, 1997. http://citeseer.nj.nec.com/54687.html.

[28] Bruno Gusmão, Lucila Ishitani, Virgílio Almeida, and Wagner Meira Jr. Disclosing users' information in an environment that preserves privacy. *Proc. of ACM Workshop on Privacy in Electronic Society (WPES 2002)*, November 2002.

[29] GVU's WWW Surveying Team. GVU's tenth www user survey. Technical report, Graphics, Visualization & Usability Center, College of Computing, Georgia Institute of Technology, 1998.

[30] Calvin Springer Hall and Gardner Lindzey. *Theories of Personality*. John Wiley & Sons, 3rd edition edition, 1978.

[31] James A. Harvey and Karen M. Sanzaro. P3P and IE 6: Good privacy medicine or mere placebo? *Computer and Internet Lawyer*, 19(4):1–6, april 2002.

[32] Harry Hochheiser. Principles for privacy protection software. *Proc. of 10th conf. on Computer, Freedom and Privacy: challenging the assumption*, pages 69–72, 2000.

[33] Harry Hochheiser. The platform for privacy preferences as a social protocol: An examination within the U.S. policy context. *ACM Transactions on Internet Technology*, 2(4):276–306, november 2002.

[34] ISAT. Security with privacy. ISAT 2002 Study, december 2002.

[35] Lucila Ishitani, Virgilio Almeida, and Wagner Meira Jr. Masks: Bringing anonymity and personalization together. *IEEE Security & Privacy Magazine*, 1(3):18–23, may/june 2003.

[36] Lucila Ishitani, Virgilio Almeida, Wagner Meira Jr., and Robert Pinto. Privacidade x personalização: avaliação quantitativa de uma arquitetura de compromisso. *Webmídia*, 2003.

[37] James B. D. Joshi, Walid G. Aref, Arif Ghafoor, and Eugene H. Spafford. Security models for web-based applications. *ACM*, 2001.

[38] A. I. Khinchin. *Mathematical foundations of information theory*. Dover Publications, 1957. Translated by R. A. Silverman and M. D. Friedman.

[39] Alfred Kobsa. Tailoring privacy to users' needs. *Proc. of 8th International Conference on User Modeling*, 2001. http://www.ics.uci.dcu/ kobsa/papers/2001-UM01-kobsa.pdf.

[40] Alfred Kobsa and Jörg Schreck. Privacy through pseudonymity in user-adaptive systems. *ACM Transactions on Internet Technology*, 3(2):149–183, may 2003.

[41] David M. Kristol. HTTP cookies: Standards, privacy, and politics. *ACM Transactions of Internet Technology*, 1(2):151–198, November 2001.

[42] Lawrence Lessig. *Code and other laws of cyberspace*. Basic books, 1999.

[43] Mark Levene and George Loizou. Computing the entropy of user navigation in the web. Research Note RN/99/42, Department of Computer Science, University College London, 1999. http://citeseer.nj.nec.com/levene00computing.html.

[44] LPWA. Lucent personalized web assistant. http://www.bell-labs.com/projects/lpwa.

[45] David M. Martin Jr., Richard M. Smith, Michael Brittain, Ivan Fetch, and Hailin Wu. The privacy practices of web browser extensions. *Communications of the ACM*, 44(2), February 2001.

[46] Daniel A. Menascé, Virgílio Almeida, Rodrigo C. Fonseca, and Marco Mendes. A methodology for workload characterization for e-commerce servers. In *1st ACM Conference in Electronic Commerce (EC-99)*, pages 119–128, November 1999.

[47] Melanie Millar. Protecting privacy in canada: Evaluating recent solutions proposed for and by the private sector. *Government Information in Canada*, 2(1), summer 1995. http://www.usask.ca/library/gic/v2n1/millar/millar.html.

[48] Josyula R. Rao and Pankaj Rohatgi. Can pseudonymity really guarantee privacy? *9th USENIX Security Symposium*, August 2000.

[49] Michael K. Reiter and Aviel D. Rubin. Crowds: Anonymity for web transactions. *ACM Transaction on Information and System Security*, 1(1):66–92, 1998. http://www.research.att.com/projects/crowds.

[50] Bruce Schneier. *Secrets & Lies: Digital Security in a Networked World*. John Wiley & Sons, 2000.

[51] Jörg Schreck. *Security and Privacy in User Modeling*. PhD thesis, Universität Gesamthochschule Essen, 2000.

[52] Stuart Soltysiak and Barry Crabtree. Knowing me, knowing you: Practical issues in the personalization of agent technology. *Proc. 3rd International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM98)*, March 1998.

[53] Sarah Spiekermann. Online information search with electronic agents: drivers, impediments, and privacy issues. Master's thesis, Humboldt Universität zu Berlin, november 2001.

[54] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, January 2000.

[55] Latanya Sweeney. Information explosion. In L. Zayatz, P. Doyle, J. Theeuwes, and J. Lane, editors, *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Urban Institute, 2001.

[56] Latanya Sweeney. K-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, 10(7):557–570, 2002.

[57] Herman T. Tavani and James H. Moor. Privacy protection, control of information, and privacy-enhancing technologies. *Computers and Society*, pages 6–11, March 2001.

[58] Michael Tchong. Brand conversion - personalization boosts conversion rates. *Iconocast*, October 1999. http://www.iconocast.com/issue/1999102102.html.

[59] Kurt Thearling. Data mining and privacy: A conflict in the making? *DS*, March 1998.

[60] Huaiqing Wang, Matthew K. O. Lee, and Chen Wang. Consumer privacy concerns about internet marketing. *Communications of the ACM*, 41(3), March 1998.

[61] Samuel D. Warren and Louis D. Brandeis. The right to privacy. *Harvard Law Review*, 4(5), December 1890.

[62] Alan Westin. *Privacy and Freedom*. Bodley Head, 1987.