LUIZ HENRIQUE GOMES

ANÁLISE E MODELAGEM DO COMPORTAMENTO DOS *SPAMMERS* E DOS USUÁRIOS LEGÍTIMOS EM REDES DE EMAIL

Belo Horizonte
05 de maio de 2006

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE E MODELAGEM DO COMPORTAMENTO DOS *SPAMMERS* E DOS USUÁRIOS LEGÍTIMOS EM REDES DE EMAIL

Tese apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

LUIZ HENRIQUE GOMES

Belo Horizonte
05 de maio de 2006

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Análise e Modelagem do Comportamento dos *Spammers*
e dos Usuários Legítimos em Redes de Email

LUIZ HENRIQUE GOMES

Tese defendida e aprovada pela banca examinadora constituída por:

Ph. D. VIRGÍLIO ALMEIDA – Orientador
Vanderbilt University, USA

Ph. D. JUSSARA M. ALMEIDA – Co-orientador
University of Wisconsin, Madison, USA

Ph. D. LUIS M. A. BETTENCOURT
Imperial College, University of London London, UK

Ph. D. VALMIR C. BARBOSA
University of California, Los Angeles, USA

Doutor EDUARDO SANY LABER
Pontifícia Universidade Católica do Rio de Janeiro, Brasil

Belo Horizonte, 05 de maio de 2006

À minha mãe, Áurea Dutra Bahia, que me ensinou na prática o significado da ética e do trabalho duro para se conseguir resultados. Aos meu filhos, Gabriel e Rafael, a quem espero transmitir parte do que aprendi e à minha esposa, Débora, que esteve o tempo todo ao meu lado nesta jornada.

Agradecimentos

A meus irmãos Paulo, Lúcia, Stael e Marcelo, pela compreensão, pelo convívio e apoio irrestrito. Palavras não são suficientes para expressar meu carinho, respeito e gratidão a vocês.

Tenho gratidão especial pelo Banco Central do Brasil, que permitiu que eu me dedicasse exclusivamente ao doutorado. Particularmente, ao meu orientador técnico José Felix Furtado por toda a ajuda no cumprimento das normas e regras estabelecidas pelo programa de pós Graduação do Banco Central do Brasil. Ao Newton Dias Passos do DEPES que me incentivou e mostrou os caminhos legais para que eu conseguisse ser aceito pelo programa do Banco Central e, especialmente, a José Roberto de Oliveira e Wallace P. de Araujo que, mesmo ficando com o setor desfalcado naquele momento, compreendeu a importância deste curso ao meu crescimento profissional e me permitiu licenciar para a realização do curso.

A meus colegas da ADBHO, DEINF e DEPES que em todos os momentos me incentivaram e ajudaram a cumprir esta jornada.

A meu orientador, Virgílio A. F. Almeida, minha gratidão por seu incentivo constante e pelos seus valiosos conselhos. Sou principalmente grato, por ter acreditado, nesse trabalho e na minha capacidade de conduzí-lo. À professora Jussara Almeida, pela paciência ao me ensinar e ao me conduzir no desenvolvimento deste trabalho. Ao professor Luis M. A. Bettencourt pelos conselhos, pela oportunidade de trabalharmos juntos e por ter me proporcionado a oportunidade visitar o "Los Alamos National Laboratory". Estar em contato com estes pesquisadores foi um privilégio para mim.

Aos colegas Cristiano Cazita, Rodrigo Barra, Fernando Castro e Fabiano Fonseca pela ajuda no desenvolvimento e construção deste trabalho, acima de tudo, pela amizade que permeou nosso relacionamento.

Aos colegas Fabrício, Tassni Cançado, Bruno Diniz, Bruno (brutti), Pedro Alcântara, José Pinheiro, Eduardo e Fabiola Nakamura, Aldri dos Santos e aos que não estão explicitamente listado mas que fizeram parte desta conquista, pela amizade. Sem vocês, com certeza, essa caminhada teria sido bem mais dura.

A todos os colegas do Laboratório *e*-Speed, do Departamento de Ciência da Computação da UFMG, por tornarem agradável e divertido o convívio diário.

Aos funcionários do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais pela ajuda e dedicação em todos os momentos.

Ao Departamento de Ciência da Computação da Universidade Federal de Minas Gerais pelo acesso as instalações e pelos recursos computacionais disponibilizados.

Aos vários profissionais que fizeram os registros das cargas de dados utilizadas nesse trabalho. Particularmente, ao Fernando Frota, Lazlo e Murilo. Sua contribuição foi inestimável.

As pessoas citadas representam um enorme apoio, mas são apenas uma parcela de todas as pessoas que me influenciaram. Este trabalho tem muito de todas elas.

Resumo

Email é um meio de comunicação cada vez mais importante e largamente utilizado para iteração entre indivíduos e/ou organizações, facilitando o contato entre indivíduos e possibilitando melhoria da produtividade nas organizações. Entretanto, o uso de ferramentas automáticas para envio de *emails* não autorizados, conhecidos como *spam*, vem, dia-a-dia, enfraquecendo a atratividade deste meio de comunicação. Até hoje, a maioria da atenção dedicada à deteção de *spam* focalizou no corpo do *email* ou nos endereços ou domínios associados aos remetentes de *spam*.

Neste trabalho, nós propusemos uma forma nova de tratar o problema causado por *spam*. Nosso objetivo é desenvolver uma compreensão profunda das características fundamentais do tráfego *spam*, do comportamento dos *spammers* e dos relacionamentos entre *spammers* e usuários legítimos em redes de *email*. Esperamos que tal conhecimento possa ser usado, no futuro, como base para projetos de técnicas mais eficazes para detectar e combater *spam*.

Primeiro, nós apresentamos uma caracterização extensiva de uma carga *emails* contendo *spam* e *emails* legítimos, que visa identificar e quantificar as características fundamentais que distinguem o tráfego *spam* do de *emails* legítimos, avaliando o impacto do tráfego *spam* no agregado e fornecendo dados para criar geradores de tráfegos sintéticos. Em seguida, nós apresentamos uma análise teórica de um modelo de redes de *email* baseado em teoria dos grafos, mostrando que existem diferenças fundamentais entre as relações desenvolvidas por *spammers* e seus pares e as relações desenvolvidas por remetentes e destinatários de *emails* legítimos. Em terceiro lugar, nós usamos as propriedades reveladas acima, do comportamento dos *spammers* e dos usuários legítimos, a fim de propôr dois novos algoritmos para deteção de *spam*. Os algoritmos propostos se utilizam das propriedades estruturais dos relacionamentos entre remetentes e destinatários de *emails* como base para a deteção de *spam*. Nossos algoritmos se propõem a corrigir classificações errôneas de um algoritmo auxiliar usado para detectção de *spam*. A precisão dessas classificações foi avaliada utilizando duas carga de dados, uma real e outra sintética. Finalmente, como a maioria do tráfego de *emails*, representada pelo tráfego do *spam*, exibe relações oportunísticas ao invés de relações sociais comuns, nós usamos este tráfego para quantificar as diferenças entre relações sociais e antisociais (representadas aqui pelo comportamento dos *spammers*) em redes de *email*.

Embora nenhuma métrica de tráfego ou comportamental estudada possa diferenciar inequivocamente *emails* legítimos de *spam*, a combinação de diversas delas mostra um retrato claro do processo por meio do qual os *email* legítimos e *spam* são criados. Por esta razão, supomos, o conhecimento gerado poderá ser usado para aumentar a eficácia, como nos algoritmos propostos, dos mecanismos de detecção de *email* ilegítimos, assim como para melhor compreender o comportamento malicioso em redes de comunicações.

Abstract

Email is an increasingly important and ubiquitous mean of communication, both facilitating contact between individuals and enabling rises in the productivity of organizations. However, the relentless rising of automatic unauthorized emails, also known as spam, is eroding away much of the attractiveness of email communication. Most of the attention dedicated to spam detection has focused on the content of the emails or on the addresses or domains associated with spam senders.

This thesis takes an innovative approach towards addressing the problems caused by spam. Our goal is to develop a deep understanding of the fundamental characteristics of spam traffic, spammers' behavior and the way spammers and non-spam (i.e., legitimate) users develop their relations in email networks, in hoping that such knowledge can be used, in the future, to drive the design of more effective techniques for detecting and combating spams.

First we present an extensive characterization of a spam-infected email workload, which aims at identifying and quantifying the characteristics that significantly distinguish spam from legitimate traffic, assessing the impact of spam on the aggregate traffic and providing data for creating synthetic workload models. Next, we present a comprehensive graph theoretical analysis of email traffic that captures the fundamental characteristics of relations among spammers and their peers, which is very different from the normal mutual relations between senders and recipients of legitimate email. Third, we use the above properties of spammers and legitimate users behavior, in order to propose two new spam detection algorithms that use structural relationships between senders and recipients of email as the basis for spam detection. Our algorithms are used to correct misclassification from an auxiliary algorithm and its classification precision is evaluated using an actual and a synthetic workloads. Finally, as the majority of email traffic, represented by spam traffic, exhibits opportunistic, rather than symbiotic social relations, we use this traffic to quantify the differences between social and antisocial (here represented by spammers behavior) behaviors in networks of communication.

Although no single behavioral or traffic metric studied can unequivocally differentiate legitimate emails from spam, the combination of several of them paint a clear picture of the processes, whereby legitimate and spam email

are created. For this reason, we suppose, they can be used to augment the effectiveness, as our proposed algorithms do, of mechanisms to detect illegitimate emails as well as to better understand malicious behavior in network of communications.

Resumo Estendido

Email é um meio de comunicação cada vez mais importante e largamente utilizado para iteração entre indivíduos e/ou organizações, facilitando o contato entre indivíduos e possibilitando melhoria da produtividade nas organizações. Entretanto, o uso de ferramentas automáticas para envio de *emails* não autorizados, conhecidos como *spam*, vem, dia-a-dia, enfraquecendo a atratividade deste meio de comunicação. Até hoje, a maioria da atenção dedicada à deteção de *spam* focalizou no corpo do *email* ou nos endereços ou domínios associados aos remetentes de *spam*.

Neste trabalho, nós propusemos uma forma nova de tratar o problema causado por *spam*. Nosso objetivo é desenvolver uma compreensão profunda das características fundamentais do tráfego *spam*, do comportamento dos *spammers* e dos relacionamentos entre *spammers* e usuários legítimos em redes de *email*. Esperamos que tal conhecimento possa ser usado, no futuro, como base para projetos de técnicas mais eficazes para detectar e combater *spam*.

Primeiro, nós apresentamos uma caracterização extensiva de uma carga *emails* contendo *spam* e *emails* legítimos, que visa identificar e quantificar as características fundamentais que distinguem o tráfego *spam* do de *emails* legítimos, avaliando o impacto do tráfego *spam* no agregado e fornecendo dados para criar geradores de tráfegos sintéticos. Em seguida, nós apresentamos uma análise teórica de um modelo de redes de *email* baseado em teoria dos grafos, mostrando que existem diferenças fundamentais entre as relações desenvolvidas por *spammers* e seus pares e as relações desenvolvidas por remetentes e destinatários de *emails* legítimos. Em terceiro lugar, nós usamos as propriedades reveladas acima, do comportamento dos *spammers* e dos usuários legítimos, a fim de propôr dois novos algoritmos para deteção de *spam*. Os algoritmos propostos se utilizam das propriedades estruturais dos relacionamentos entre remetentes e destinatários de *emails* como base para a deteção de *spam*. Nossos algoritmos se propõem a corrigir classificações errôneas de um algoritmo auxiliar usado para detectção de *spam*. A precisão dessas classificações foi avaliada utilizando duas carga de dados, uma real e outra sintética. Finalmente, como a maioria do tráfego de *emails*, representada pelo tráfego do *spam*, exibe relações oportunísticas ao invés de relações sociais comuns, nós usamos este tráfego para quantificar as diferenças entre relações sociais e antisociais (representadas aqui pelo comportamento dos *spammers*) em redes de *email*.

Embora nenhuma métrica de tráfego ou comportamental estudada possa diferenciar inequivocamente *emails* legítimos de *spam*, a combinação de diversas delas mostra um retrato claro do processo por meio do qual os *email* legítimos e *spam* são criados. Por esta razão, supomos, o conhecimento gerado poderá ser usado para aumentar a eficácia, como nos algoritmos propostos, dos mecanismos de detecção de *email* ilegítimos, assim como para melhor compreender o comportamento malicioso em redes de comunicações.

Resumo do capítulo "Workload Models of Legitimate and Spam Emails"

Apesar do grande número de artigos sobre custo e do imenso número de métodos para deteção e filtragem de *spams* existentes, os esforços para analizar as características deste tipo de tráfego na Internet foram um tanto limitados. Além de algumas caracterizações de carga de *emails* [1, 2], nós estamos cientes somente de algumas análises limitadas de um número reduzido de características do tráfego do *spam* na literatura [3, 4, 5].

Neste capítulo nós apresentamos uma caracterização extensiva do tráfego *spam*. Nosso objetivo é desenvolver uma compreensão profunda das características fundamentais do tráfego *spam* e do comportamento dos *spammers*, com o intuito de que tal conhecimento possa ser aplicado no projeto de técnicas mais eficazes para detectar e combater *spam*.

Nossa caracterização baseou-se em um *log* de oito dias, perfazendo um total de 360 mil *emails* que chegaram ao servidor SMTP de borda da Universidade Federal de Minas Gerails. As mensagens foram classificadas como *spam* ou não utilizando técnicas padrão de detecção de *spam*. Para cada um das duas cargas resultantes e, também, para a carga agregada, nós analisamos um conjunto de métricas estatísticas, utilizando informações disponíveis nos cabeçalhos dos *emails*. Nosso objetivo era identificar as características quantitativas e qualitativas que, significativamente, distinguam o tráfego *spam* do de *emails* legítimos, avaliando o impacto do tráfego *spam* sobre o agregado. Além disso, avaliamos como o último se diferencia do tráfego de *emails* legítmos, fornecendo dados para gerar carga sintética de *emails* contendo *spam*.

Os resultados mais importantes que obtivemos foram:

- Diferentemente do tráfego de *emails* legítimos que exibe um claro padrão diário e semanal com picos durante as horas de trabalho do dia e durante os dias de semana, os números de *spams*, de *bytes* de *spam*, de *spammers* distintos e de destinatários de *spam* distintos são, de forma geral, insensíveis aos períodos do dia ou aos dias da semana, permanecendo quase estáveis durante todo o tempo nos *logs* analisados.

- Os tempos entre chegadas de *spam* e *emails* legítimos são, aproximadamente, distribuídos segundo uma distribuição exponencial. Entretanto, enquanto as taxas de chegada de *spam* permanescem aproximadamente

estáveis em todos os períodos analisados, as taxas de chegada de *emails* legítimos variam por um fator de cinco.

- O tamanho dos *email* nos tráfegos *spam*, não *spam* e agregado seguem distribuições Log-normal. Entretanto, o tamanho médio de um *email* não *spam* é de seis a oito vezes maior do que o tamanho médio de um *spam* nos nossos *logs*. Além disso, o coeficiente de variação (CV) dos tamanhos de *emails* legítmos são, em torno, de três vezes maior que o CV dos tamanhos dos *spams*. Assim, o impacto do tráfego *spam* no agregado implica em uma diminuição no tamanho médio dos *emails*, além de um aumento na variabilidade nos tamanhos destes.

- A distribuição do número de destinatários por *email* tem cauda muito mais pesada na carga *spam* do que na carga de *emails* legítmos. Visto que somente 5% dos *email* não *spam* são endereçados a mais de um usuário, enquanto 15% dos *spams* têm mais de um destinatário. Na carga agregada, a distribuição é fortemente influenciada pelo tráfego *spam*, desviando-se significativamente do observado na carga de não *spams*.

- A principal distinção entre os tráfegos *spam* e não *spam*, quando consideramos a popularidade dos remetentes e destinatários, acontece quando medimos esta em número de *emails* por destinatários. Enquanto no tráfego não *spam* e na carga agregada a popularidade em número de *emails* recebidos por dia por destinatário é modelada aproximadamente por uma distribuição Zipf mais uma probabilidade constante para os destinatários de uma única mensagem por dia, no tráfego *spam*, o modelo mais adequado compõe-se da concatenação de duas distribuições Zipfs, além da probabilidade constante atribuída aos destinatários de um único *spam* por dia.

- Ao analisar a localidade temporal entre os destinatários no tráfego não *spam*, encontramos dois conjuntos distintos de usuários, um com uma forte localidade temporal e outro que recebe *emails* somente esporadicamente. Por outro lado, estes dois conjuntos não estão claramente definidos no tráfego *spam*. De fato, a localidade temporal é, em média, muito mais fraca entre os destinatários de *spam* e ainda mais fraca nos destinatários quando consideramos o tráfego agregado. Resultados similares foram encontrados para a localidade temporal em torno dos remetentes.

- A distribuição dos tamanhos das listas de contatos dos remetentes e destinatários no tráfego não *spam* é muito mais enviesada para listas com tamanhos menores do que no tráfego *spam*. De fato, um *spammer* típico envia *email*, em média, para duas vezes mais destinatários distintos que um remetente de *email* legitímo. Além disso, um destinatário de *spam* típico recebe *emails* de um número de *spammers* distintos quase três vezes maior que o número dos remetentes distintos de um destinatário de *emails* legitímos. Por último, o

tráfego *spam* impacta significativamente a distribuição dos tamanhos da lista de contatos para remetentes e destinatários no tráfego agregado.

Nossa caracterização revela diferenças significativas entre as cargas de *emails spam* e não *spam*. Estas diferenças são, possivelmente, devido à inerente natureza distinta de remetentes de *emails* e seus relacionamentos com os destinatários dos *email* em cada grupo. Visto que uma transmissão de *email* legítimo é o resultado de um relacionamento bilateral, iniciado tipicamente por um ser humano, dirigido por algum relacionamento social. Enquanto, por outro lado, uma transmissão de *spam* é basicamente uma ação unilateral, tipicamente executado por ferramentas automáticas e dirigido pela vontade dos *spammers* de alcançar tantos alvos quanto possível, indiscriminadamente, sem ser detectado.

Resumo do capítulo "Characterization of Graphs of Legitimate and Spam Email"

Neste capítulo nós apresentamos uma caracterização do modo como os *spammers*, os remetentes e *emails* legítimos e seus pares desenvolvem seus relacionamentos através da troca de *emails*. Para tal, modelamos os tráfegos de *emails* legítimos e *spam* utilizando teoria dos grafos e usamos estas estruturas para identificar métricas estruturais que podem ser utilizadas para diferenciá-los. Outro objetivo deste capítulo é encontrar um conjunto de métricas que podem ser utilizadas, no futuro, em um modelo preditivo da disseminação de *spams*.

Este estudo vai além de diversos outros estudos recentes [6, 7] da natureza do tráfego *spam*. Nós tratamos de uma base de dados diferente, envolvendo um número muito maior dos usuários e mensagens e analisamos um conjunto muito maior de métricas estáticas e dinâmicas. Nós mostraremos que não há nenhuma métrica do grafo que distingue inequivocamente um *email* legítimo de um *spam*. Entretanto, identificamos diversas métricas podem ser combinadas em um mecanismo probabilístico para deteção de *spam*, como os que propomos no Capítulo 6.

Os resultados mais importantes que obtivemos foram:

- O grau de saída dos remetentes externos nos gráfos tanto de usuários quanto de domínios são bem modelados por uma lei de potências. Entretanto, no grafo de usuários, os graus de saída abaixo de 20 são muito mais prováveis para *spammers* do que para os remetentes de *emails* legítimos. Por outro lado, remetentes com grau de saída maior que 400 quase não existem no tráfego *spam*. No grafo de domínio, a distribuição dos graus de saída mostra uma probabilidade muito mais elevada para nós com graus baixos no tráfego *spam* do que no tráfego de *emails* legítimos.

- A análise da reciprocidade da comunicação sugere que uma forte assinatura dos *spammers* é seu desequilíbrio estrutural na comunicação entre o conjunto dos remetentes e seus destinatários associados. O mesmo

desequilíbrio é encontrado na análise dos grafos de domínio.

- O conjunto de assimetria ("Asymmetry set") mostra o número de usuários *spam* no conjunto diferença entre os "in-set"e "out-set"de um nodo em um grafo. Encontramos que, tanto no grafo de usuário como no de domínio, há uma forte correlação estatística entre o tamanho deste conjunto e o número dos *spammers* nele. Esses dois resultados mostram, juntos, que as mensagens de *spam* quase não são respondidas.

- Nossos resultados para o coeficiente de aglomeração mostram que os usuários do tráfego *spam* têm uma coesão muito mais baixa na sua comunicação quando comparado com usuários legítimos. Mostrando que *spammers*, geralmente, enviam *emails* a destinatários não correlacionados.

- Em conseqüência da assimetria na comunicação entre usuários do tráfego *spam*, encontramos que, em ambos os gráficos, há muito menos probabilidade de se visitar um nodo participante do tráfego *spam* do que um nodo do legítimo, durante uma caminhada aleatória.

- O subgrafo construído a partir dos usuários do tráfego *spam* é uma estrutura que cresce muito mais rapidamente quando comparado com o grafo construído com os usuários do tráfego legítimo. Alé do mais, devido ao tamanho de nossos *logs*, uma semana de dados, não encontramos um ponto de saturação em nenhum dos subgrafos analisados.

- A localidade temporal entre pares de usuários (remetente-destinatário) é mais forte no tráfego legítimo do que no *spam*, mostrando a concentração da comunicação entre usuários legítimos no tempo.

- Ao analisar a entropia do fluxo das mensagens/*bytes* das comunicações entre remetentes de *spam* e de *emails* legítimos e seus pares, encontramos que remetentes de *emails* legítimos comunicam-se de maneira muito variável com seus pares do que *spammers*.

Resumo do capítulo "Using Structural Similarity for Improving Spam Detection"

Nenhum mecanismo de detecção de *spam* existente é infalível [3, 8]. Entre os principais problemas encontrados estão os falsos positivos e os erros nas classificações. Além disso, os filtros devem ser atualizados continuamente para capturar a diversidade de mecanismos introduzidos por *spammers* para evitar detecção.

Neste capítulo, nós propomos e avaliamos dois algoritmos para a deteção de *spam* que usam a estrutura dos relacionamentos entre remetentes e destinatários como base para a deteção de mensagens de *spam*. O algoritmos trabalham em conjunto com um outro detector de *spam* (chamado daqui por diante de algoritmo auxiliar), necessário para produzir um histórico de classificações de *emails*, de modo que as novas classificações possam ser inferidas

baseando-se na estutura dos relacionamentos dos novos remetentes e destinatários com os usuários existentes e no histórico de envio para recebimento destes. A idéia chave do algoritmo é que as listas dos destinatários distintos para os quais os *spammers* e os usuários legítimos enviam mensagens, assim como as listas dos remetentes distintos dos quais usuários recebem mensagens [1], podem ser usadas como identificadores de remetentes e destinatários no tráfego de *emails* [3, 9, 10]. Nós mostramos que a aplicação dos nossos algoritmos sobre os resultados da classificação de um classificador auxiliar conduz à correção de um número significativo de erros nas classificações.

Os algoritmos que apresentamos neste capítulo visam melhorar a eficácia de mecanismos de detecção de *spam*, reduzindo falsos positivos e fornecendo informações para ajustar suas coleções de regras. Diferentemente da maioria das alternativas para detecção de *spam*, nós focalizamos nossa análise nas características que conjecturamos são as mais difíceis para *spammers* mudarem, isto é, a estrutura dos seus relacionamentos com seus destinatários. Outros estudos recentes focalizaram em técnicas do combate a *spam* baseadas em características de modelos de grafos construídos com usuários do tráfego de *emails* [6, 7, 11].

Resumo do capítulo "Social versus Anti-Social Behavior in Email Traffic"

A caracterização do tráfego de *emails* como uma rede social complexa [2] foi tópico principal de diversos estudos recentes. Pelo melhor de nosso conhecimento, todas estas caracterizações têm lidado através do uso de estratégias simples com o fato de que o tráfego de *emails* apresenta não somente interações sociais, mas também outros tipos de relacionamentos entre usuários de *email*. Exemplos de tais estratégias são a limitação do tráfego aos *emails* aos internos a uma organização [13, 14, 15, 16, 17], limitação de usuários de *email* de tal maneira que somente usuários que participam em comunicações em dois sentidos são considerados [15, 17, 16], eliminação de usuários de *email* com um volume elevado de menssagens [15, 16, 17] ou, ainda, eliminação das ligações que conectam usuários de *email* que apresentam um número baixo de mensagens trocadas [14].

Por outro lado, tem crescido o interesse em descobrir a evidências em antisociais do comportamento em redes. Entre os tópicos analisados em trabalhos recentes estão: observações proibidas, comportamento hostil e grupos de polarização [18, 19]. *Email* como um possível meio de distribuição maciça é associado, particularmente, com a disseminação de vírus eletrônico, assim como com tráfego *spam* [20] que tem inundado a Internet com mensagens não desejadas e que, geralmente, contêm propostas comerciais ou, mais recentemente, uma variedade de outras fraudes. Este comportamento, que nós chamamos genericamente de *antisocial*, apresenta características diferentes das relações sociais, as quais já foram extensivamente analisadas no contexto de redes.

[1] As listas do contato como definidas no Capítulo 4

[2] De acordo com Newman, uma em rede social é um conjunto de pessoas ou grupo de pessoas conectadas segundo os padrões da interação social, que podem ser representados como nós e ligações, respectivamente, em um grafo [12].

O que é conceitualmente interessante sobre este comportamento antisocial é que gera um grafo com propriedades quantitativas e dinâmicas não triviais que refletem algum tipo de interação que nós caracterizamos quantitativamente e contrastamos com as propriedades gerais de redes sociais.

Neste capítulo é apresentado uma caracterização dos relacionamentos no tráfego do *email* que aborda o problema sob um novo ponto de vista. Primeiramente, a maioria das limitações adotadas em estudos precedentes são retiradas. Em segundo, nós consideramos um único tráfego agregado de *emails* que forma dois componentes diferentes: um componente legítimo gerado por usuários legítimos durante sua interação social através da troca de *emails* e um componente não legítimo gerado pelos usuários que usam seu *email* para enviar *spam*. Note que, o componente legítimo de nosso estudo é dominado pelo comportamento social, enquanto o componente não legítimo apresenta um comportamento que nós chamaremos de "comportamento antisocial".

Os resultados mais importantes que obtivemos foram:

- Nós encontramos uma distribuição de lei de potências como melhor modelo para distribuição dos graus dos nodos nas quatro redes analisadas nos tráfegos sociais e antisociais. Entretanto, o modelo se adequa melhor à distribuição do grau nas redes construídas com o tráfego social do que nas redes construídas com o tráfego antisocial.

- Nós encontramos que usuários no tráfego antisocial têm coeficiente de agregação significativamente mais baixo que no tráfego social, principalmente nas redes que incluem usuários externos.

- Nós encontramos graus de nodos vizinhos não correlacionados, para as redes sociais e antisociais que incluem os usuários externos. Por outro lado, devido ao grau não usual de emissão de *spam* a usuários internos forjados, em uma de nossas cargas, encontramos correlação positiva na rede antisocial interna. Finalmente, todas as redes sociais internas apresentaram correlação positiva entre os graus dos nodos vizinhos.

- Nós encontramos uma distribuição de lei de potências como melhor modelo para a distribuição dos tamanhos das comunidades em todas as redes definidas, nas duas cargas. Não obstante, encontramos que as comunidades nas redes antisociais que inclui usuários externos têm em média 30% mais nós que as comunidades sociais. Adicionalmente, encontramos que a maior comunidade em cada rede social é quase duas vezes maior que a correspondende maior comunidade na rede antisocial.

- Encontramos que os nodos no tráfego social têm um coeficiente de troca preferencial significativamente maior que seus pares no tráfego antisocial. Além disso, nodos na rede antisocial que inclui os usuários externos têm um baixo, mas às vezes não zero, coeficiente da troca preferencial.

- Finalmente, os resultados mostram que o tráfego social de *emails* tem uma entropia mais baixa (mais informação estrutural) do que o tráfego antisocial para os dois períodos analisados, o de trabalho e o complementar. Além disso, quanto maior o intervalo sob a análise, maior a diferença.

Nossa caracterização revela diferenças significativas entre os tráfegos social e antisocial. Estas diferenças acontecem, possivelmente, devido à natureza distinta inerente dos relacionamentos dos remetentes de *email* e suas conexões com os destinatários em cada grupo. Dado que, o envio de um *email* legítimo é o resultado de um relacionamento bilateral, tipicamente iniciado por ser humano, dirigido por algum relacionamento social, e a transmissão de um *spam* é bàsicamente uma ação unilateral, executada tipicamente por ferramentas automáticas e dirigido pela vontatde dos *spammers* de alcançar tantos alvos quanto possível.

Resumo do capítulo "Conclusions and Future Work"

Esta tese fornece uma análise extensiva do comportamento de usuários legítimos e *spammers* no tráfego de *emails*, clarificando as características que os distinguem mais significativamente.

Em resumo, nós mostramos que o comportamento dos *spammers* e dos usuários legítimos exibem as fortes diferenças que refletem em diversas distinções estruturais e dinâmicas. Em segundo lugar, nós usamos estes conhecimentos como base para proposição e análise de dois algoritmos que se utilizam da similaridade estrutural das listas de contatos dos usuários de cada grupo para deteção de *spam*. Finalmente, nós estudamos as características estruturais e dinâmicas das redes de *spam* e de usuários legítimos no tráfego de *emails* a fim descobrir o comportamento antisocial dos *spammers*.

Possíveis direções trabalho futuro incluem:

- Verificação de nossos resultados em relação ao tempo e a cargas distintas, a fim certificar de que os resultados não são específicos das cargas analisadas;

- Projeto e avaliação de novas técnicas de detecção de *spam* que exploram as distinções entre os tráfegos *spam* e legítimo e entre os comportamentos dos usuários apresentados nesta tese;

- Projeto e execução de um gerador sintético de carga de *emails* contendo *spam* a ser usado na avaliação experimental de novas estratégias de deteção de *spam*;

- Usar grafos com pesos nas arestas (por exemplo, com número das mensagens como peso) na análise de características antisociais e sociais dos relacionamentos nas redes. Como mostramos a intensidade do tráfego é uma das principais distinções entre comportamentos antisocial e sociais em redes do *email*;

- Caracterizar outras cargas antisociais como ataques de vírus ou *worms*, a fim tentar descobrir assinaturas gerais do comportamento antisocial;
- Estudar como grandes volumes de *spam* em redes do email podem influenciar a disseminação de vírus ou outro conteúdo malicioso em redes do *email*, a fim avaliar se as técnicas de imunização que se aplicam hoje continuam a ser eficazes em redes de *email* infectadas de *spam*.

Sumário

Lista de Figuras

Lista de Tabelas

Chapter 1

Introduction

Electronic mail (email) has become a *de facto* means to disseminate information to millions of users in the Internet. Among the strengths of electronic communication media, such as email, are the relatively low cost, high reliability and, generally, fast delivery. Electronic messaging is not only cheap and fast, but it is also easy to automate [21]. These properties tend to make social email communication among users more indiscriminate, as email senders may easily send copies of a message to multiple parties that play no active role in the relationship between sender and recipients. Furthermore, the low cost [1] involved in delivering a message to a large group of recipients makes email obviously also very attractive for commercial advertising purposes and distribution of a variety of other kind of scams. Email as a means of potential mass distribution is particularly associated with the dissemination of computer viruses, worms as well as spam messages [20].

Our goal in this thesis is to develop an in-depth study of spam email traffic and of the relationships developed by spammers [2] and their peers using email. As well as, uncover their main characteristics and their distinctions from legitimate email traffic and from the relationship developed by legitimate email senders [3] and their peers. Throughout the analysis, we show how these properties can be used to improve spam detection techniques. Finally, we use these properties in order to propose, implement and test two innovative algorithms based on the relationships among email users for improving spam detection.

1.1 Spam definition

Spamming is the act of sending unsolicited (commercial) electronic messages in bulk, and the word spam has become the synonym for such messages. This word is originally derived from spiced ham, which is a registered

[1] In energy, time and reputation of the sender

[2] It is a common used term to designate the one who sends spams.

[3] It is a commonly used term to designate the one who sends non spam email.

trademark of Hormel Foods Corporation [22]. Based on the origin of the word spam, all other emails are so called ham. Official terminology for spam is defined in [21, 20, 23]: Unsolicited Bulk Email (UBE) or Unsolicited Commercial Email (UCE) - "*It is Internet mail ("email") that is sent to a group of recipients who have not requested it. A mail recipient may have at one time asked a sender for bulk email, but then later asked that sender not to send any more email or otherwise not have indicated a desire for such additional mail; hence any bulk email sent after that request was received is also UBE*".

Classical direct marketing has been used for a long time. For these marketing methods costs increase proportionally with the number of potential customers reached and revenue is only created by selling real products or services. In this classical approach, frauds are almost excluded because lots of investments is necessary in order to lead profitability. Spamming, as a new form of direct marketing, in contrast, is subject to lots of frauds, as the costs of sending millions of spam emails is very low and the cost increase with the number of potential customers reached is near zero. Therefore, in order to generate profit, it suffices that only a very small rate of them result in true business transactions. Spams are sent out by companies and by individuals, primarily for profit. Spam advertises, generally, offers a diversity of goods like; such as pornography, computer software, medical products, investments and credit card accounts.

1.2 Motivation

Spam is an increasing threat to the viability of Internet email and a danger to Internet commerce. Spammers take away resources from users and service suppliers without compensation or authorization. Defending against spam has the characteristics of an arms race, the effectiveness of spam fighting techniques is permanently challenged and constant upgrades and new developments being necessary.

1.2.1 Economical

The volume of spam is increasing at a very fast rate. In January 2003, 24% of all Internet emails were spam. By March 2005, this fraction had increased to 83% [24]. A report from October 2003 shows that, in North America, a business user received on average 10 spams a day, and that this number is expected to grow by a factor of four by 2008 [25]. Furthermore, AOL and MSN, two large ISPs, report blocking, daily, a total of 2.4 billion spams from reaching their customers' in-boxes. This traffic corresponds to about 80% of daily incoming emails at AOL [26].

This rapid increase in spam traffic is taking its toll in end users, business corporations and system administrators. Results from a survey in March 2004 with over two thousand American email users report that over 60% of them are less trusting of email systems, and over 77% of them believe being online has become unpleasant or annoying

due to spam [27]. The impact of spam traffic on the productivity of workers of large corporations is also alarming. Research enterprises estimate the yearly cost per worker at anywhere from US$ 50 to US$ 1400, and the total annual cost associated with spam to American businesses is in the range of US$ 10 billion to US$ 87 billion [26]. Finally, estimates of the cost of spam must also take into account the costs of computing and network infra-structure upgrades as well as quantitative measures of its impact on the quality of service available to traditional non-spam email traffic and other “legitimate" Internet applications.

1.2.2 Technical

A number of approaches have been proposed to alleviate the impact of spam. These approaches can be categorized into pre-acceptance and post-acceptance methods, based on whether they detect and block spam before or after accepting the email [3]. Examples of pre-acceptance methods are black listing [28] and gray listing or temp-failing [29]. Pre-acceptance approaches based on server authentication [4, 30], economic disincentives [31] and accountability [32] have also been proposed. Examples of post-acceptance methods include Bayesian filters [33, 34], collaborative filtering [35], email prioritization [3], and recent techniques that exploit specific properties of the relationship established between spammers and spam recipients [6, 7].

Although existing spam detection and filtering techniques have, reportedly, very high success rates (up to 97% of spams are detected [35]), they suffer from one key limitation. The rate of false positives, i.e., legitimate emails classified as spams, can be as high as 15% [36], incurring costs that are hard to measure. Second, the lifetime of existing techniques is compromised by spammers frequently changing their mode of operation (e.g., forging their email addresses and/or misspelling the body of spam messages). In other words, spam filters have their effectiveness frequently challenged.

Finally, despite the large number of reports on spam cost and the plethora of previously proposed spam detection and filtering methods, the efforts towards analyzing the fundamental characteristics of this type of Internet traffic have been somewhat limited.

1.3 Contributions of this thesis

This thesis takes an innovative approach towards addressing the nature of spamming and presents:

- A thorough understanding of the qualitative and quantitative fundamental characteristics of spam traffic and its impact on the aggregate email traffic, uncovering the way network resources are being stressed by spam traffic;

- A quantitative exhaustive understanding of the main distinctions between spammers' and legitimate senders' and recipients' relationships and how it can be used in spam detection;
- Two innovative algorithms that uses quantitative knowledge of user relationships for spam detection improving;
- The structural and dynamical characteristics of the antisocial spam networks and their distinction relative to social email networks.

We hope that such knowledge can be used, in the future, to drive the design of more effective techniques for detecting and combating spam emails or for improving our proposed detection algorithms, as well as to better understand the malicious behavior in complex networks and its distinction relative to social behavior.

1.4 Outline of the thesis

In this thesis we present a deep analysis and modeling of spammers' characteristics as well as their distinctions relative to legitimate senders' and recipients' behaviors. Chapter 2 presents an overview of the studies related to the content of this thesis. Next, in Chapter 3 is presented the definitions, formulae and applications in computer science area of the theory used in the other Chapters of this thesis. Thirdly, in Chapter 4, we present a statistical characterization of spam and legitimate email traffics, uncovering their qualitative and quantitative distinctions. In Chapter 5 we model the relationship established among email users and between users and domains using graph theory and present their structural and dynamical characteristics. In Chapter 6 we propose, implement and test a new spam detection algorithm that uses the structural characteristics of the relationship between senders and recipients to classify emails as spam or legitimate. In Chapter 7, we present the main structural and dynamical distinctions between antisocial email traffic (here taken as spam) and social traffic, which corresponds to legitimate email traffic. Finally, Chapter 8 presents our conclusions an outlook for future work.

Chapter 2

Related Work

In this Chapter we present an overview of the studies related with the content of this thesis. In order to do so, we divided the studies into three groups. In Section 2.1 we present studies related with statistical characterization of Internet workloads and, specifically, email workloads. Section 2.2 describes an overview of the studies related with anti-spam techniques. Finally, Section 2.3 describes studies characterizing and modeling social and antisocial behaviors in networks.

2.1 Workload characterization

Developing a clear understanding of a workload is a key step towards the design of efficient and effective distributed systems and applications. A number of traffic characteristics, most of them not available in current management tools, can be clearly understood through the analysis of real traffic. An in-depth understanding of a workload, among other uses, is valuable for the formulation of policies of use, development of techniques to detect network anomalies, and for extracting guidelines for design of new services.

Calzarossa and Serazzi [37] presented in a survey of the workload characterization by pointing out through a few applications, the various steps required for the construction of a workload model and the sets of parameters and techniques to be considered in various types of studies. The steps proposed in the survey were summarized by the authors as: Choice of the set of parameters able to describe the behavior of the workload; choice of the performance/monitoring tools; experimental measurement collection; analysis of workload data and finally; construction of the workload models. In Chapter 4 we present an extensive characterization of a spam email workload where we use a diversity of steps and techniques proposed in [37].

A number of characterization studies of different workload types that led to valuable insights into system design are available in the literature, including the characterization of Web workloads [38], streaming media work-

loads [39, 40, 41] and, recently, peer-to-peer [42] and chat room workloads [43]. To the best of our knowledge, previous efforts towards characterizing spam traffic have been very limited.

Next, we discuss previous characterizations and analysis of email and spam workloads.

2.1.1 Email traffic characterizations

In [1, 2], the authors provide an extensive characterization of several email server workloads, analyzing email inter-arrival times, email sizes, and number of recipients per email. They also analyze user accesses to email servers (through the POP3 protocol), characterizing inter-access times, number of emails per user mailboxes, mailbox sizes and size of deleted emails, and propose user behavior models. In Chapter 4 of this thesis, we characterize not only a general email workload, but also a spam workload, in particular, aiming at identifying a signature of spam traffic, which can be used in the future for developing more effective spam-controlling techniques. Furthermore, in Chapter 4 Sections 4.3 and 4.4, we contrast our characterization results for legitimate emails with those reported in [1, 2].

2.1.2 Spam email traffic characterizations

In [5] the authors analyze two sets of SMTP packet traces, the second collected four years after the first, and identify a significant increase in the use of DNS black lists over this four-year period. Furthermore, they also show that the distribution of the number of email connections per spam sender, although not Zipf-like in the first collected set, does exhibit a heavy-tailed Zipf-like behavior in the second set, collected four years later. In contrast, the characterization presented in Chapter 4 covers a much broader range of workload aspects. Nevertheless, we compare our results with those presented in [5] whenever it is appropriate.

Twining *et al* [3] present a simpler server workload characterization, as the starting point for investigating the effectiveness of novel techniques for detecting and controlling junk emails (i.e., virus and spams). They analyze the logs of two email servers that include a virus checker and a spam filter, and characterize the arrival process of each type of email (i.e., spam, virus, and “good") per-sender, the percentage of servers that send only junk emails, only good emails, and a mixture of both. A major conclusion of the paper is that popular spam detection mechanisms such as blacklisting, temp-failing, and rate-limiting are rather limited in handling the problem. Chapter 4 presents a much more thorough characterization of spam traffic, and contrasts, whenever appropriate, our findings to those reported in [3].

In [4, 44], the authors analyze the temporal distribution of spam arrivals as well as spam content in two distinct workloads. The authors also discuss the factors that make users and domains more likely to receive spams and

the reasons that lead to the use of spam as a communication and marketing strategy. In [4] is also included a brief discussion of the pros and cons of several anti-spam strategies. While in [44] it is discussed the privacy of users in the Internet and the correlation between the spammer business and the spam content. In Chapter 4 we do not analyze content of emails, but we present an extensive characterization of spam and legitimate traffics using a workload not only bigger and more recent but with a much greater percentage of spams.

2.2 Describing anti-spam techniques

A diversity of previous studies have focused on techniques for reducing the impact of spam email traffic. These approaches can be categorized into pre-acceptance and post-acceptance methods, based on whether they detect spam before or after accepting messages. Examples of pre-acceptance methods are server authentication [4, 30], black listing [28], white listing [45], Gray lists or Temp-failing [29] and accountability [32]. Generally, pre-acceptance tend to catch spam messages coming from well know senders or from relays not properly configured. While post-acceptance methods are mostly based on information available in the body or header of the messages and include Bayesian filters [33], collaborative filtering [35] and, more recently, user social and antisocial behaviors [6, 7]. Post-acceptance filters are, generally, more complex than pre-acceptance filters. Therefore, they tend to be more expensive in terms of computational power, since they need to process the messages before being able to classify them properly.

Among pre-acceptance methods is black listing [46, 28], where the connections are refused by the server after finding the sender in a list of “non trustable senders". These black lists can be private or public. While the first is constructed based on the site administrators experiences, the public type is provided by specialized sites, generally, as a paid information service. Note that, black listing is subject to several problems, on the other hand trustable senders can be wrongly classified as non trustable, avoiding to receive legitimate messages. While, on the other hand, spammers change their sender names very frequently, causing a strong reduction of the effectiveness of black listing.

White lists are lists of trustable ISP (*Internet Service Providers*) [45] from which the server always accepts messages without applying any post-acceptance spam detection method. This kind of listing scheme is to be used in conjunction with any post-acceptance spam detection method in order to avoid delaying and to reduce resource usage by the post-acceptance method adopted. This scheme is also subject to errors, due to the use of forged identifiers by spam senders.

In the Gray lists or Temp-failing methods the server returns a “temporary fault message"every time it receives a try to open a connection where the pair (sender, recipient) appears for the first time. In doing so, badly config-

ured servers, which is supposed to be the case of SMTP servers [47] of spam senders, do not try do re-send the message [29].

In the Delaying techniques the SMTP servers try to maximize the delay of the spammers connection, in order to reduce their effectiveness [48, 49]. Also in this case, wrong sender classification can cause lots of problems in the flow of legitimate messages.

In the authentication approaches every SMTP server must authenticate origin servers, by using digital certification sites, before acceptance of a connection [30, 32]. As spammers use forged domains to send their emails they should not be authenticated. Moreover, authentication of forged domains is not possible.

Assigning costs for sending/receiving emails is a method in which the SMTP server must pay an amount [32, 31], using electronic money, for every email sent. Thereafter, as spammers must send millions of emails per day, their activities turn to be not cost effective. The main problem with these two previously described techniques is the need of changing in the actual structure of Internet SMTP servers which is very hard.

The spam detection algorithms we propose in Chapter 6 are not in the class of pre-acceptance filters. But, as will be shown, they can work in conjunction with any of the previous pre-acceptance techniques. Moreover, they need to work in conjunction with a post-acceptance filter.

On the other hand, there is a great number of spam detection methods that run after the email acceptance. Among them, the body-based email filters are the most common [50, 35, 51, 52, 53, 54, 55, 56]. Moreover, the Bayesian [50, 52, 55, 56] filters, which use Bayesian statistical techniques in the classification of message body, are in the group of the more frequently used. One interesting characteristic of Bayesian filters is their learning capability, which in principle increase their accuracy after each new classification.

Parallel to Bayesian filters we find the simple-text-classification-based filters, which use text classification techniques and a pre-established table of probabilities to classify messages using the words and sentences that appear in their body or headers [51, 54].

In [35] a collaborative filter is proposed, where the classification of trustworthy users are added to a database and used in the subsequent classifications. The authors related that the filter reached 98% of accuracy.

In [53] the authors experimentally compare a diversity of body-based filters. The results demonstrate that there is a plateau of accuracy at 99.9%, reached by Bayesian filters and that is surpassed only by computationally very expensive filters. The success behind the Bayesian filters is closely related to their learning capability. In doing so, this type of filter is capable of capturing the concept of spam each user has. Running on the mailbox of each user, it is trained by the users with new classifications and unclassifications and, after some instances it reaches a very precise knowledge of the concept of spam for each particular user.

Recent studies, as can be seen in [6, 7], focus on the structural characteristics of the relationships established

between senders and recipients of spam and legitimate email messages, in order to use them as the basis for the design of new post-acceptance spam detection techniques.

In [6] a graph is created to represent the email traffic captured in the mailbox of individual users. The clustering coefficient of each of these graphs is used to classify messages as spam or legitimate. The results show that 53% of the messages were precisely classified using the proposed approach. Our algorithms, proposed in Chapter 6, are different from the one proposed in [6], since we base our analysis on several users other than focusing on a single one. Moreover, we need an auxiliary algorithm whereas the algorithm proposed in [6] is a stand alone approach.

In [7] the authors propose to detect machines that behave as spam senders by analyzing a border flow graph of sender and recipient machines. In contrast to our work, presented in Chapter 6. The authors propose trying to detect relays that are commonly used to send spam, instead of identifying spam messages one-by-one. They use a metric that evolves through time to determine the periods in which certain relays are being used by spammers and block the traffic coming from those relays.

2.3 Analyzing and modeling social, virus, and malicious behavior

The ease with which messages can be distributed to many recipients is at the root of most opportunistic behavior involving email. These behaviors displays different characteristics from social relations and, beside social networks properties, is beginning to be subject to many studies. Next, we present a diversity of studies that analyze and model social and, malicious behaviors in email networks.

2.3.1 Analyzing and modeling social networks

Social networks were focus of several prior studies, which have provided evidence that there are several structural graph properties that are characteristic of graphs that represent social iteration between users, but not of other complex networks [12, 57]. Among these properties are positive assortative mixing, high clustering coefficient and small word property.

In [12] the authors describe several new models of social networks, based on random graphs with given degree distributions, that allows to explore directly the effects of various degree distributions. Moreover, they compare the predictions of the models to data for a number of real-world social networks. They found that the predictions for typical vertex-vertex distances, clustering coefficients, and expected vertex degree agree well with empirical data. One of the main conclusions of the paper is that if a comparison between a network and the equivalent random model reveals substantial disagreement, it strongly suggests that there are significant social forces at work in the

network. In Chapter 7 we show that some of the models proposed in this paper are in agreement with the empirical results we found for social networks of emails and, on the other hand, are not applied for our antisocial networks.

In [57], Newman studied assortative mixing in networks and, as a special case, he studied degree assortative mixing in a variety of networks. He proposed a number of measures of assortative mixing appropriate to various mixing types, and he applied them to a variety of real-world networks, showing that assortative mixing is a pervasive phenomenon found in many networks. In the particular case of mixing by degree, he found strong variation with assortativity in the connectivity of the network. The results found in Chapter 7 for the mixing by degree of our social and antisocial networks are in agreement with the results presented in [57].

In [14, 15, 17, 58] the authors suggested that social networks can be understood in terms of their organization in communities, i. e., groups of nodes that have a greater density of links among them proportionally with other nodes in the network. They also show, as can be seen in [15, 58], that such communities can be quantified and used to infer information about the existence of informal groups inside the organizations as well as its hierarchical structure.

In [14] the authors describe a method for the automatic identification of communities from email logs within an organization, using a betweenness centrality algorithm. The method was verified using nearly 1 million email messages, and it was effective at identifying true communities, both formal and informal. Moreover, the proposed method was able to identify leadership roles within the communities. In Chapter 7 we use the minimum spanning tree (MST) algorithm in order to extract communities from our social and antisocial networks and characterize their size distribution.

In [15, 17, 58] the authors propose a procedure for analyzing and characterizing complex networks and apply it to social networks. Their results reveal self-organization of social network, and a power law distribution of community sizes. In Chapter 7 we extracted the communities of our social and antisocial networks, but using a distinct algorithm. Nevertheless, we also found a power law distribution as the best statistical model that describes community sizes.

In summary, in Chapter 7 we separate the social from the antisocial, represent by spam email traffic, component of the email networks and present a characterization of their skeleton properties and, of the dynamic of their communication.

2.3.2 Communication patterns in social networks

The communication patterns in social networks have been extensively studied in order to explain characteristics of the dissemination of information inside organizations and other social groups [59, 60]. A clear understanding of these patterns can influence positively organizations in the management of their information flow.

In [59] the authors present a study and a model of information flow within an organization. Due to the fact that an item relevant to one person is more likely to be of interest to other individuals in the same social circle than to those outside of it. An epidemic model on a scale-free network with this property has a finite threshold, implying that the spread of information is limited. The study was carried out using a workload of emails exchanged inside an organization. In Chapter 7 we study the dynamical of the incoming communication of social and antisocial email traffics, the last represented by the spam email traffic.

Eckmann, Moses and Sergi [60] have shown, via the consideration of mutual information between sets of two and three time series of email received and sent by a single user, that there are coherent dynamical structures among the behavior of users, with patterns of sending and receiving that exhibit temporal interdependence, and to a large extent approximate synchronization. Conversely, in Chapter 7 we study the incoming email traffic, aggregated over all users, and show that spam in-flow of emails has much less temporal structure when compared with legitimate in-flow of emails.

Finally, in a related work, Barabasi [61], has shown that the statistics of time intervals between email messages is typically well described by power law distributions, with characteristic bursting of activity alternating with long silences. He also proposed a simple decision-based queuing model to explain the process. In Chapter 7 we study the dynamics of communication in social and antisocial email groups by two simple metrics, the coefficient of reciprocity and entropy of the incoming traffic.

2.3.3 Virus propagation analysis and models in email networks

Analyzing and modeling virus propagation is the subject of several recent studies [13, 62, 63, 64]. A clear understanding of the infection mechanism and how it propagates through a network of email users is a important and necessary step toward developing an efficient immunization scheme.

In [13, 62] the authors show that there are several structural properties characteristic of social networks that can be used to explain the speed and amplitude of the spread of email viruses through organizations. Consequently, knowledge of the properties helps the design of techniques for virus immunization. Particularly, the authors study the virus propagation over a diversity of different networks and show that on networks with scale free degree distribution (ex. social networks) the immunization of only 10% of the nodes with higher degrees is very effective in stopping a virus dissemination. Moreover, it is shown that if the scale free degree distribution is not present in the network, this immunization scheme can be very ineffective.

In [63] authors present a virus model that accounts for the behavior of email users, such as email checking frequency and the probability of opening an email attachment. They found that the topology of email network plays an important role in determining way an email virus spreads. Moreover, in the study, email virus propagation on

three topologies are compared: power law, small world and random graph topologies. One of the key conclusions is that while in power law networks email viruses spread more quickly than on a small world or a random graph topology, the immunization defense against viruses is more effective on a power law topology.

Finally, in [64] it is shown that error tolerance is not shared by all redundant systems. They show that only scale free networks have this property. However, they also show that these networks are vulnerable to attacks on some vital nodes.

Chapter 3

Background

3.1 Introduction

The aim of this Chapter is to present the definitions, formulae and applications in computer science area of the theory used in the other Chapters of this thesis.

3.2 Statistical distributions

A probability distribution is a function that assigns probabilities to the measurable sets of a measurable space [65].

A random variable is defined as a measurable function from a probability space to some measurable space. This measurable space is the space of possible values of the variable, and it is usually taken to be the real numbers. Every random variable gives rise to a probability distribution, and this distribution contains most of the important information about the variable. As an example, if X is a random variable, the corresponding probability distribution assigns to the interval $[a, b]$ the probability $f(X) = P(a \leq X \leq b)$, i.e. the probability that the variable X will take a value in the interval $[a, b]$, which is called probability distribution function (PDF). The probability distribution of the variable X can be uniquely described by its cumulative distribution function (CDF) $F(x)$., which is defined by: $F(x) = P(X \leq x)$.

Several probability distributions are so important in theory or applications that they have been given specific names. In this Section we focus in some of them that are commonly found in the literature as models for computer science phenomenas and that we use in the next Chapters of this thesis.

The **exponential** distributions [65], are continuous probability distributions, often used to model the time between independent events that happen at a constant average rate, which is defined by:

$$f(x) = \lambda e^{-\lambda x}, \tag{3.1}$$

where $\lambda > 0$ is the distribution rate parameter.

As examples of use of **exponential** distributions models are: time between submission of queries to search machines in the Internet and time executing a process in a computer. In Chapter 4 we found exponential distributions as the best model for inter arrival rates of emails in legitimate (non-spam emails), spam (emails flagged as spam) and aggregate email traffics.

The **log-normal** distribution [65] is the probability distribution of any random variable whose logarithm is normally distributed. If X is a random variable with a normal distribution, then $\exp(X)$ has a **log-normal** distribution; likewise, if Y is log-normally distributed, then $\log(Y)$ is normally distributed. The **log-normal** distribution has probability density function (PDF) defined as:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}, \tag{3.2}$$

for $x > 0$, where μ and σ are the mean and standard deviation of the variable's logarithm.

Log-normal distributions are useful to model positively skewed data, whereas the natural log of the data are normally distributed, such as movement data and electrical measurements . In Chapter 4 we found **log-normal** distributions as the best model for the distribution of the size of the emails in non-spam, spam and aggregated email traffics.

The **gamma** distribution [65] is a continuous probability distribution. The probability density function of the **gamma** distribution can be expressed in terms of the gamma function:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{x}{\beta}}, \tag{3.3}$$

for $x > 0$, where $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter of the **gamma** distribution.

The **gamma** distribution is commonly used in modeling skewed data such as movement data and electrical measurements.

The **Weibull** distribution [65] is a continuous probability distribution with the probability density function defined as:

$$f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}, \tag{3.4}$$

for $x > 0$, where $\alpha > 0$ is the rate parameter and $\beta > 0$ is the shape parameter.

One example of application of **Weibull** distribuition is for time to failure modeling.

In Chapter 4 we found **gamma** and **Weibull** distributions as the models for the distribution of sender and recipient stack distances in non-spam and spam email traffics.

Computer science is plenty of high variability phenomenas, the above distributions are not capable of correctly modeling this kind of phenomena. As examples of high variability phenomena we have: file sizes in a computation environment, link bandwidth in Internet environment and in/out degree of graph models of router connections in the Internet [66].

In order to treat with this models we use **power law** distribution models. Generally, **power law** distributions have the following formulas:

$$P(X > x) \approx Cx^{-\alpha}, \tag{3.5}$$

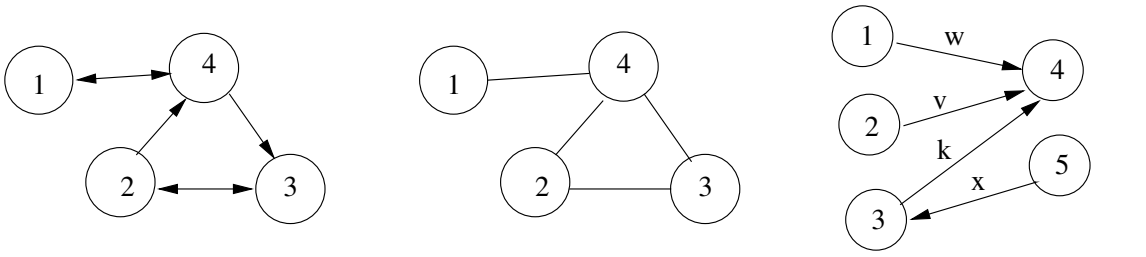
where $C \neq 0$ is a constant, and α is the distribution parameter.

In Chapters 4, 5 and 7 we use a **power law** distribution to model the popularity of a diversity of distinct objects. In some these case we referred as a **power law** distribution called Zipf's Law [67].

Originally, Zipf's law stated that, in a corpus of natural language written text, the frequency of any word is roughly inversely proportional to its rank in the frequency table. So, the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. The term has come to be used to refer to any of a family of related **power law** probability distributions. From this on, it was observed that Zipf's law could be applied to the popularity of a diversity of types of objects in computer science.

3.3 Network definitions

The popularity of network models for representing virtually any natural structure reflects the flexibility and generality of the network models, including those undergoing dynamical changes of topology. Several investigations in networks, as ours, involve the representation of the structure of interest as a graph, followed by an analysis of the topological features of the obtained representation performed in terms of a set of informative measurements [66]. Such activities can be understood as being aimed at the topological characterization of the studied structures, as we do in Chapter 5 for spam and legitimate networks. Another related application is to use the obtained measurements in order to discriminate between different classes of structures, as we do in Chapter 7 in order to find distinctions between social and antisocial networks.



a) Directed unweighted b) Undirected unweighted c) Bipartite directed weighted

Figure 3.1: Graph examples

Characterization and classification of structures using networks imply important question of how to choose the most appropriate measurements. To begin with, there is an unlimited set of topological measurements. Next, we have the fact that measurements are often correlated, implying redundancy. Ultimately, one has to rely on her/his knowledge of the problem and available measurements in order to select a suitable set of features to be made. For such reasons, it is of paramount importance to have a good knowledge not only of the most representative measurements, but also of their respective properties and interpretation.

The purpose of this Section consists precisely in providing an integrated and comprehensive guide to the network topological measurements commonly found in the literature and, used in this thesis, as well as their main interpretation applied to the characterization of email networks. We begin presenting the basic network concepts and notation, following we present several network topological measurements and its basic interpretation.

3.3.1 Graph basic definitions

Definition 1 *A **weighted directed graph**, G, is defined as a triple (V, E, ω), where $V(G)$ is a set with N vertices (or nodes), $E(G)$ is a set of M directed edges (or directed links), and a mapping ω defined as: $\omega(E) \rightarrow \Re$.*

Each vertex can be identified by an integer value $i = 1, 2, \ldots, N$. Moreover, each edge is identified by a pair $(i, j) \in E(G)$ that represents a connection going from vertex i to vertex j to which a weight $\omega((i, j)) \in \Re$ is associated.

Based on Definition 1, we define an **unweighted direct graph**, G, as a pair (V, E) defined exactly as previously, note that in this case we do not have an ω mapping defined [68, 69]. Moreover, if we does not consider the directions of the edges, i.e., if $(i, j) = (j, i) \; \forall \; (i, j)$ and $(j, i) \in E(G)$, we have a **unweighted undirect graph**.

When the set $V(G)$ is such that $E(G) = V_1(G) \times V_2(G)$, where $V_1(G)$ and $V_2(G) \subset V(G)$ and $V_1(G) \cap V_2(G) = \emptyset$, we have a **bipartite graph**, which can be directed, undirected, weighted or unweighted.

Figures 3.1 a-c show examples of graph representations, where nodes are represented by circles and edges are the lines linking the circles. In Figure 3.1-a is presented an example of an **unweighted directed graph** and in Figure 3.1-c we present an **weighted directed bipartite graph** [68, 69].

In Chapters 5 and 7 we use graph models in context of email networks. There, we define users or domains of users who send or receive emails as nodes and, the edges representing messages being exchanged between two nodes. In accordance with the analysis being done, one of the previously defined type of graphs is chosen.

3.3.2 Network properties

Distinct network topologies manifest according with the type of relation being modeled. As said, there exist a diversity network topology metrics. As a matter of the fact, many of the metrics present characteristic values, for each relation being modeled, that can be found in the literature. This Section defines the network topological and dynamical metrics we use in this thesis.

The **degree** of a node i, represented by k_i, is the number of links connected to that node, often called connectivity of i. In the case of directed networks, there are two kinds of degrees: the **out-degree**, k_i^{out} , which is the number of outgoing links of i, and the **in-degree**, k_i^{in}, corresponding to the number of incoming links of i. Note that, generally, $k_i \leq k_i^{out} + k_i^{in}$. As an example, in Figure 3.1-a, $k_2^{out} = 2$ and $k_2^{in} = 2$.

Two nodes i and j of a graph G are said to be **neighbors** if there exist an edge linking them. The **out-set** of a node i in a directed graph $G = (V, E)$, represented by $OS(i)$, is the set of nodes j, such that there is an edge $(i, j) \in E(G)$. Analogously, we define the $IS(j)$ as the **in-set** of nodes i such that $(i, j) \in E(G)$. As it can be seen, in Figure 3.1-a, the $IS(2) = 3$ and $OS(2) = 3, 4$.

In Chapters 5 and 7 we use node **degree** in order to characterize social and antisocial email networks. Aiming at using this to detect spam, as in Chapter 5, or to extract social/antisocial components of email networks, as in Chapter 5. Moreover, in Chapters 7 and 6 we use the **in-set** and **out-set** definitions in order to extract groups of nodes called **clusters** and use them in characterizing social/antisocial networks, as in Chapter 7 or to improve spam detection, as in Chapter 6.

The **clustering coefficient** (CC_i) [66] of a node i measures the probability of neighbors of a node i being neighbors themselves. Let k_i be the number of neighbors of a node i, then if all of its neighbors are neighbors of themselves, we have $k_i(k_i - 1)$ edges linking them. In an actual network we can measure the number of nodes in the set of neighbors of a node (i) that are neighbors themselves, lets represent this number as l_i, so the CC_i of a node i is defined as:

$$CC_i = \frac{l_i}{k_i(k_i - 1)} \tag{3.6}$$

The CC measures the transitivity of a network and is defined as the average CC_i for all node i in the network and is defined as:

$$CC = \frac{1}{N}\sum_{i=0}^{N} CC_i, \tag{3.7}$$

where N is the total number of nodes of the network.

The CC is used in the study of propagation throughout a network and is important in the understanding of the information propagation, diseases, computer virus and other in actual networks.

In Chapters 5 and 7 we use node **clustering coefficient** in order to characterize non-spam, spam, social and antisocial email networks. Again, aiming at using this to detect spam, as in Chapter 5, or to extract social/antisocial components of email networks, as in Chapter 5.

Link reciprocity [70] in a directed network is defined as the tendency of a pairs of vertices to form mutual connections between each other. In Chapter 5 we use **link reciprocity** in order to characterize legitimate and spam networks, looking for distinctions that enable us to improve spam detection. In another way **link reciprocity** tell us how much information is lost when a directed network is regarded as undirected.

We use following definition of **link reciprocity**, in Chapter 5, that we called, **communication reciprocity** (CR). The **communication reciprocity** of a node i is defined as:

$$CR(i) = \frac{|OS(i) \cap IS(i)|}{|OS(i)|}, \tag{3.8}$$

where $OS(i)$ is the **out-set** and $IS(i)$ is the **in-set** of the node i, respectively. With our choice of normalization this metric measures the probability of a node i be targeted by one of its targets.

The **path** between two non-adjacent vertices i and j of a network is defined as a sequence of m edges (i, k_1), (k_1, k_2), . . . , (k_{m-1}, j); such set of edges is called a path between i and j, and m is the length of the path. We say that two vertices are connected if there is at least one path connecting them.

A **connected component** (CC) is a subset of the nodes of a graph, so that one node can be reached from any other node in the set following edges between them [66]. Besides that, a **giant connected component** is defined as the greatest CC in a graph. Typically a network has one GCC which contains significant part of the network components. The characterization of GCC in networks is used in order to understand the behavior of the flow between nodes in networks.

In Chapters 5 and 7 we study the GCC of our email networks of non-spam, spam, social and antisocial traffics.

One **shortest path** between two vertices i and j in a graph [68, 69] is defined as one path between them with minimum size among all the paths linking them.

In networks, the greater the number of paths in which a vertex or edge participates, the higher the importance of this vertex or edge for the network. Thus, assuming that the interactions follow the shortest paths between two vertices, it is possible to quantify the importance of a vertex or an edge (u) in terms of its **betweenness** (B_u). Defined as:

$$B_u = \sum_{(i,j) \in E} \frac{\sigma(i,u,j)}{\sigma(i,j)}, \tag{3.9}$$

where $\sigma(i,u,j)$ is the number of shortest paths between vertices i and j that pass through vertex or edge u, $\sigma(i,j)$ is the total number of shortest paths between i and j, and the sum is over all pairs $(i,j) \in E$ of distinct vertices.

Another interesting structural characteristic of networks is the probability of visiting a node during a **random walk** through the graphs [71]. At each step of the **random walk** we need to select the next node to be visited. This can be done in two ways. The next node can be randomly selected from the **out-set** of the current node or we can perform a jump. For a jump, one of the nodes of the graph is selected randomly as the next node. Note that, this measure is closely related to node **betweenness** since higher node **betweenness** tends to generate a higher probability of visitation. The probability $P(i)$ of finding a node i in a random walk is computed iteratively as follows:

$$P(i) = \frac{d}{N} + (1-d) * \sum_{j \ \in \ IS(i)} \frac{P(j)}{|OS(j)|}, \tag{3.10}$$

where d is the probability of performing a jump during a random walk, N is the number of nodes in the graph.

In Chapter 5 we study the probability of a node participate of a **random walk** in our non-spam and spam email networks.

For networks with different types of vertices one interesting structural characteristic to be studied is the way the types of vertices connect, which is called the **mixing pattern** or **assortative mixing** of the network [66].

In Chapter 7 we analyze the nature of degree correlations between neighbors in undirected versions of our email networks, by looking at the correlation between node degrees of the neighbors of nodes [72]).

3.4 Vector basic definitions

The purpose of this Section consists precisely in providing an integrated and comprehensive guide to the mathematical vector basic definitions commonly found in the literature and, used in this thesis, as well as their main interpretation applied to the characterization of email networks.

A **vector** of dimension n is an ordered collection of n elements, which are called components or dimensions. We often represent a vector by some letter, just as we use a letter to denote a scalar (real number) in algebra. An n-dimensional vector $\vec{A}$ has n elements denoted as $A_1, A_2, \ldots, A_n$. Symbolically: $\vec{A} = \langle A_1, A_2, \ldots, A_n \rangle$.

Two vectors are equal if their corresponding components are equal. The **norm** of a vector $\vec{A}$ of dimension n, denoted $|\vec{A}|$, is defined as:

$$|\vec{A}| = \sqrt{A_1^2 + A_2^2 + ... + A_n^2} \tag{3.11}$$

The sum of two vectors $\vec{A} = (A_1, A_2, \ldots, An)$ and $\vec{B} = (B_1, B_2, \ldots, B_n)$ is a third vector defined as: $\vec{C} = \vec{A} + \vec{B} = (A_1 + B_1, A_2 + B_2, ..., A_n + B_n)$.

The **inner product** of two vectors $\vec{A}$ and $\vec{B}$ is a number defined as:

$$\vec{A} \bullet \vec{B} = \cos(\vec{A}, \vec{B})|\vec{A}||\vec{B}|, \tag{3.12}$$

In Chapter 6 we define a vectorial representation for email senders and recipients. We then, represent group of users by the summation of their vectorial representation. Finally, we define similarity of users and group of users by the normalized inner product of their vectorial representation.

3.5 Clustering algorithms

Complex networks, generally, present an inhomogeneous connecting structure characterized by the presence of groups whose vertices are more densely interconnected to one another, called **clusters**, than with the rest of the network. These grouping structure was previously found and analyzed in many kinds of networks such as social [19, 73, 14], biological [73] and Internet web pages networks. Clustering identification in large networks is particularly useful because vertices belonging to the same cluster are more likely to share properties and dynamics. In addition, the number and characteristics of the existing clusters provide subsidies for identifying the category of a network as well as understanding its dynamical evolution and organization. In the case of the World Wide Web, for instance,

pages related to the same subject are typically organized into communities, so that the identification of these communities can help the task of seeking for information.

In Chapters 6 and 7 we use clustering algorithms in order to find and use group distinctions of non-spam, spam, social and antisocial email networks to improve spam detection as well as extract antisocial components. In Chapter 6 we use historical information about clusters in a new spam detection algorithm, and show that it is very effective in improving spam detection. However, in Chapter 7 we use studied clustering structure of social and antisocial networks in order to characterize their distinct structure.

A diversity of clustering algorithms have been proposed and used in the literature [73, 14, 17]. In Chapters 7 we use use the following algorithm called **Minimum Spanning Tree Clustering Algorithm** (MST) to extract communities of our email networks [74]:

1. Start by assigning each node its own cluster, so that if you have N nodes, you now have N clusters, each containing just one node.
2. Let the distances between the clusters equal the distances between the nodes they contain.
3. Find the closest (or most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
4. Compute distances between the new cluster and each of the old clusters (as the number common neighbors the users inside the new cluster and the users inside each old cluster have).
5. Repeat steps 1 to 4 until the happening of any of the previously chosen stop condition.

However, in Chapter 6 we use an adapted version of MST as part of the new spam detection algorithm proposed.

3.6 Conclusions

In this Chapter we present a diversity of definitions that we use and refer in next Chapters of this thesis. Moreover, we presented here a few collection of applications of each measurement or property defined.

Chapter 4

Workload Models of Legitimate and Spam Emails

4.1 Introduction

Despite the large number of reports on spam cost and the plethora of previously proposed spam detection and filtering methods, the efforts towards systematically analyzing the characteristics of this type of Internet traffic have been somewhat limited. In addition to some previous email workload characterizations [1, 2], we are aware of only a few limited analysis of some spam traffic characteristics in the literature [3, 4, 5].

In this Chapter we take an innovative approach towards addressing the problems caused by spam and present an extensive statistical characterization of a spam traffic. A preliminary version of this work was published in [10]. Our goal is to develop a deep understanding of the fundamental characteristics of spam traffic and spammer's behavior, hoping that such knowledge can be used, to drive the design of more effective techniques for detecting and fighting spam.

Our characterization is based on an eight-day log of over 360 thousand incoming emails to a large university in Brazil. Standard spam detection techniques are used to classify the emails into two categories, namely, spam and non-spam (i.e., legitimate email). For each of the two resulting workloads, as well as for the aggregate workload, we analyze a set of statistics, based on the information available in email headers. We aim at identifying the quantitative and qualitative characteristics that significantly distinguish spam from non-spam traffic, assessing the impact of spam on the aggregate traffic by evaluating how the latter deviates from the non-spam traffic, and providing data for generating realistic synthetic spam-infected email workloads

Our key findings are:

- Unlike traditional non-spam email traffic, which exhibits clear weekly and daily patterns, with load peaks during the day and on weekdays, the numbers of spam emails, spam bytes, distinct active spammers and

distinct spam email recipients are roughly insensitive to the period of measurement, remaining mostly stable during the whole day, for all days analyzed.

- Spam and non-spam inter-arrival times are approximately exponentially distributed. However, whereas the spam arrival rates remain roughly stable across all periods analyzed, the arrival rates of non-spam emails vary by as much as a factor of five in the periods analyzed.

- Email sizes in the spam, non-spam and aggregate workloads follow Log-normal distributions. However, the average size of a non-spam email is from six to eight times larger than the average size of a spam, in our workload. Moreover, the coefficient of variation (CV) of the sizes of non-spam emails is around three times higher than the CV of spam sizes. Thus, the impact of spam on the aggregate traffic is a decrease on the average email size but an increase in the size variability.

- The distribution of the number of recipients per email is more heavy-tailed in the spam workload. Whereas only 5% of non-spam emails are addressed to more than one user, 15% of spams have more than one recipient. In the aggregate workload, the distribution is heavily influenced by the spam traffic, deviating significantly from the one observed in the non-spam workload.

- Regarding daily popularity of email senders and recipients, the main distinction between spam and non-spam email traffics comes up in the distribution of the number of emails per recipient. Whereas in the non-spam and aggregate workloads, the distribution is well modeled by a single Zipf-like distribution plus a constant probability of a user receiving only one email per day, the distribution of the number of spams a user receives per day is more accurately approximated by the concatenation of two Zipf-like distributions, in addition to the constant *single-message* probability.

- There are two distinct and non-negligible sets of non-spam recipients: those with very strong temporal locality and those who receive emails only sporadically. These two sets are not clearly defined in the spam workload. In fact, temporal locality is, on average, much weaker among spam recipients and even weaker among recipients in the aggregate workload. Similar trends are observed for the temporal locality among email senders.

- The distributions of contact list sizes for senders and recipients are much more skewed towards smaller sizes in the non-spam workload. In fact, a typical spammer sends emails on average to twice as many distinct recipients as a typical legitimate email sender. Moreover, a typical spam recipient receives spams from a number of distinct spammers that is almost three times the number of non-spam senders a typical recipient

has contact with. Furthermore, the spam traffic significantly impacts the distribution of contact list sizes for senders and recipients in the aggregate traffic.

Therefore, our characterization reveals significant differences between the spam and non-spam workloads. These differences are possibly due to the inherent distinct nature of email senders and their relationships with email recipients in each group. Whereas a non-spam email transmission is the result of a bilateral relationship, typically initiated by a human being, driven by some social relationship, a spam transmission is basically a unilateral action, typically performed by automatic tools and driven by the spammers' will to reach as many targets as possible, indiscriminately, without being detected.

The remaining of this chapter is organized as follows. Our email workloads and the characterization methodology are described in Section 4.2. Section 4.3 analyzes temporal variation patterns in the workloads. Email traffic characteristics are discussed in Section 4.4. Email recipients and senders properties are analyzed in Section 4.5. Finally, Section 4.6 presents our conclusions.

4.2 Email workload

This section introduces the email workload analyzed in this Chapter. Section 4.2.1 describes the data source and collection architecture. The methodology used in the characterization process is presented in Section 4.2.2. Section 4.2.3 provides an overview of our email workload.

4.2.1 Data source

Our email workload consists of anonymized SMTP logs of incoming emails to a large university, with around 22 thousand students, in Brazil. The logs were collected at the central Internet border email server of the university. This server handles all emails coming from the outside addressed to the vast majority of students, faculty and staff with email addresses under the major university domain name. Only emails addressed to two out of over 100 university sub-domains (i.e., departments, research labs, research groups) do not pass through and, thus, are not logged by, the central border server.

The central email server runs the Exim email software [75], the AMaViS virus scanner [76] and the Trend-micro Vscan anti-virus tool [77]. It also runs a set of pre-acceptance spam filters, including local black lists and local heuristics for detecting suspicious senders. These filters block on average 50% of all daily SMTP connection arrivals. The server also runs SpamAssassin [51], a popular spam filtering software, over all emails that are accepted. SpamAssassin detects and filters spams based on a set of user-defined rules. These rules assign scores to each received email based on the presence in the subject or in the email body of one or more pre-categorized

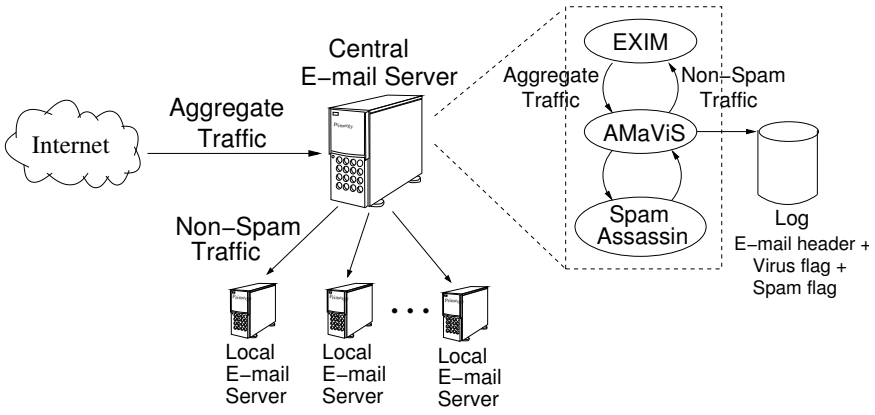


Figure 4.1: Data Collection at the Central Email Server

keywords taken from a constantly changing list. Highly ranked emails are flagged as spams. Spam Assassin also uses size-based rules, which categorize emails larger than a pre-defined size as non-spam [1]. Emails that are neither flagged as spam nor as virus-infected are forwarded to the appropriate local servers, indicated by the sub-domain names of the recipient users.

We analyze an eight-day log collected by the AMaViS software at the central email server, during academic year at the university. Our logs store the header of each email that passes the pre-acceptance filters, along with the results of the tests performed by SpamAssassin and the virus scanners. In other words, for each email that is accepted by the server, the log contains the arrival time, the size, the sender email address, a list of recipient email addresses and flags indicating whether the email was classified as spam and whether it was detected to be infected with a virus. Figure 4.1 shows the overall data collection architecture at the central email server.

Emails that are flagged with virus or addressed to recipients in a domain name outside the university, for which the central email server is a published relay, are *not* included in our analysis. These emails correspond to only 0.8% of all logged data.

Note that the central server does not perform any test on the existence of the accepted email recipient addresses. Such tests are performed by the local servers. Thus, some of the recipient email addresses in our logs may not actually exist. These recipient addresses could be result of honest mistakes or the consequence of dictionary attacks [78], a technique used by some spammers to automatically generate a target distribution list with a large number of *potential* email addresses.

[1] We note that this rule was applied to only 4% of all emails in our log

4.2.2 Characterization methodology

As the basis for our characterization, we first group the emails logged by AMaViS into two categories, namely, *spam* and *non-spam* (also referred to as “ham" in the literature [52]), based on whether the email was flagged by SpamAssassin. Three distinct workloads are then defined:

- *Spam* - only emails flagged as spam by SpamAssassin.
- *Non-Spam* - only emails not flagged as spam by SpamAssassin.
- *Aggregate* - all emails logged by AMaViS.

We characterize each workload separately. The purpose is fourfold. First, we can compare and validate our findings for the non-spam workload with those reported in previous analysis of traditional (non-spam) email traffic [3, 1, 2, 79]. Second, we are able to identify the characteristics that significantly distinguish spam from non-spam traffic. Third, we are also able to assess the quantitative and qualitative impact of spam on the overall email traffic, by evaluating how the aggregate workload deviates from the non-spam workload. Finally, we provide useful data for generating synthetic spam and non-spam workloads, which in turn, can be used in an experimental evaluation of alternative spam detecting techniques.

Our characterization focuses on the information available in the email headers, logged by AMaViS. In doing so, we characterize the email arrival process, distribution of email sizes, distribution of the number of recipients per email, distribution of contact list sizes, popularity and temporal locality among email recipients and senders.

Each workload aspect is analyzed separately for each day in our eight-day log (except when explicitly stated), recognizing that their statistical characteristics may vary over time. The email arrival process is analyzed during periods of approximately stable arrival rate, as daily load variations may also impact the aggregate distribution.

To find the statistical distribution [2] that best models each workload aspect, on each period analyzed, we compare the least square differences of the best fitted curves for a set of alternative distributions commonly found in other characterization studies [1, 39, 40, 41, 43, 80, 13, 81]. We also visually compared the curve fittings at the body and at the tail of the measured data, favoring a better fit at either region whenever appropriate to capture the aspects of the workload that are most relevant to system design. For instance, shorter inter-arrival times and larger email sizes have a stronger impact on server capacity planning. Thus, we favor a better fit at the body (tail) of the data for determining the arrival process (distribution of email sizes). In Sections 4.4-4.5, we show only the results for the best fitted distributions.

[2] In Chapter 3 we presented a formal definition of a probability distribution, as well as, the definition of all statistical distributions we propose as models for the traffic metrics presented in this Chapter

Measure	**N**on-Spam	**S**pam	**A**ggregate
Period	2004/01/19-26	2004/01/19-26	2004/01/19-26
Number of days	8	8	8
Total # of emails	191,417	173,584	365,001
Total size of emails	11.3 GB	1.2 GB	12.5 GB
Total # of distinct senders	12,338	19,567	27,734
Total # of distinct recipients	22,762	27,926	38,875
Avg # distinct recipients/msg (CV)	1.1 (0.74)	1.7 (1.38)	1.4 (1.27)
Avg # msgs/day (CV)	23,927 (0.26)	21,698 (0.08)	45,625 (0.17)
Avg # bytes/day (CV)	1.5 GB (0.39)	164 MB (0.19)	1.7 GB (0.37)
Avg # distinct senders/day (CV)	3,190 (0.22)	5,884 (0.10)	8,411 (0.11)
Avg # distinct recipients/day (CV)	8,981 (0.15)	14,936 (0.24)	19,935 (0.20)

Table 4.1: Summary of Workloads (CV=Coefficient of Variation)

A visual inspection of the list of sender *user names* in the spam workload indicated that a large number of them seemed to be a random sequence of characters, suggesting forging. Note that sender IP addresses may also be forged, although we expect it to happen less frequently. Our logs contain only sender domain names. However, sender IP addresses are separately collected by the Exim software. By analyzing the Exim logs collected at the same period of our AMaViS logs, we found that, on average, a single sender domain name is associated with 15 different IP addresses, whereas the average number of different domains per sender IP address is only 6. In other words, there is no indication of which information is more reliable. Because the results of Spam Assassin are available only in the AMaViS logs and a merge of both logs is hard to build, our per-sender analysis focuses only on sender domain names.

Thus, throughout this chapter, we simply use:

- *Email sender* - to refer to the email sender domain.
- *Email recipient* - to refer to an email recipient user name.

4.2.3 Overview of the workloads

An overview of our three workloads is provided in Table 4.1. Note that, although spams correspond to almost 50% of all emails, spam traffic corresponds to only 10% of all bytes received during the analyzed period. Furthermore, the total number of distinct spammers is almost 60% larger than the number of distinct senders in the non-spam workload. Thus, the average number of emails originated from the same domain is smaller in the spam workload, possibly due to spammers periodically changing their email domain names to escape from black lists. Note, also, that the total number of spam recipients as well as the number of recipients per spam are also, significantly, larger than the corresponding metrics in the non-spam workload. This may be explained by spammers effort to target

Group	Senders		Recipients	
	%	% Msg	%	% Msg
Only Non-Spam	29	31	25	10
Only Spam	56	23	38	20
Mixture	15	46	37	70

Table 4.2: Distribution of Senders and Recipients

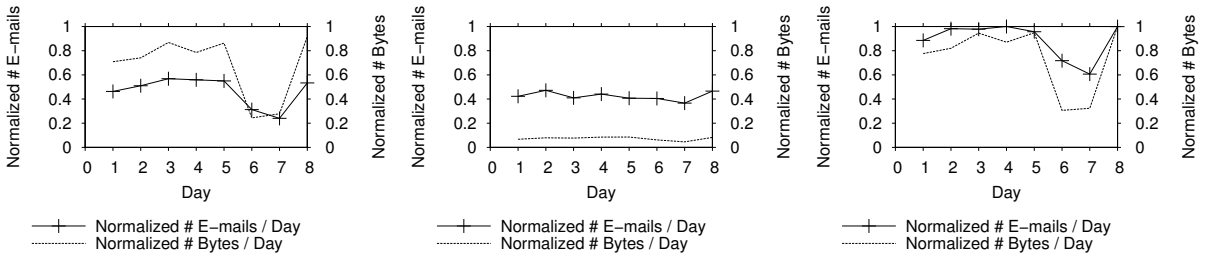
as many addresses as possible (e.g., dictionary attacks). Another interesting point is the much lower variability in spam traffic, which is further discussed in Section 4.3. Similar conclusions hold on a daily basis, as shown in the last five rows in Table 4.1.

Table 4.2 shows the percentages of senders and recipients that send and receive only non-spam emails, only spams and a mixture of both. It also shows the percentage of emails each category of sender/recipient is responsible for. More than half of all domains send only spams, whereas 15% of them send both types of emails. We also point out that, on average, six out of the ten most active spam senders on each day send only spams. Nevertheless, the spam-only servers are responsible for only 23% of all emails, whereas 46% of the emails originate from domains that send a mixture of spams and non-spams. These results may be explained by spammers frequently “forging" new domains. In [3], the authors also found a large fraction of senders who send only junk (i.e., virus or spam) emails. However, they found those senders represent a larger fraction of the emails in their workloads than we find in ours.

Table 4.2 also shows that whereas 25% of all recipients are not the target of spam, around 38% of them appear only in the spam workload and receive 20% of all emails in our log. Furthermore, we found that around 50% of the spam-only recipients received less than 5 emails during the whole log, and that a number of them seemed forged (e.g., randomly generated sequence of characters). These observations lead us to speculate that many spam-only recipients are the result of two frequent spammer actions: dictionary attacks and removal of recipients from their target list after finding they do not exist (i.e., after receiving a “not a user name" SMTP answer). They also illustrate a potentially harmful side-effect of spam, namely the use of network and computing resources for transmitting and processing a significant number of emails that are addressed to non-existent users and, thus, that are discarded only once they reach local email servers.

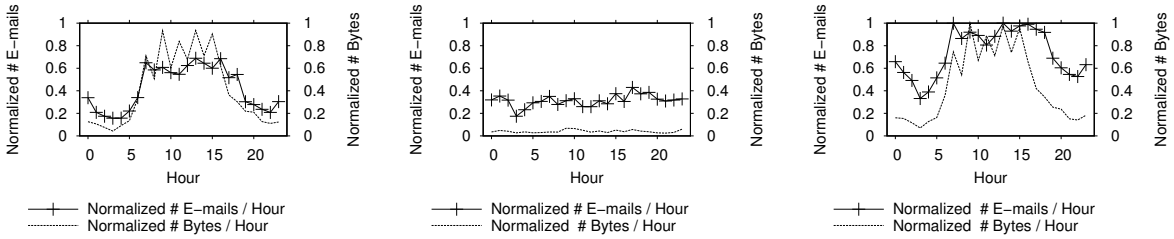
4.3 Temporal variation patterns in email traffic

This section discusses temporal variation patterns in each of our three email workloads, namely, spam, non-spam and aggregate workloads. Section 4.3.1 analyzes daily and hourly variations in load intensity. Temporal variations in the numbers of distinct email recipients and senders are discussed in Section 4.3.2.



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.2: Daily Load Variation (Normalization Parameters: Max # Emails=51,226, Max # Bytes 2.24 GB)



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.3: Hourly Load Variation (Normalization Parameters: Max # Emails=2,768, Max # Bytes 197 MB)

4.3.1 Load intensity

Figure 4.2 shows daily load variations in the number of emails and number of bytes, for non-spam, spam and aggregate workload respectively. The graphics show load measures normalized by the daily load peak observed in the aggregate traffic. The normalization parameters are shown in the caption of the figure.

Figure 4.2-a shows that daily load variations in the non-spam email traffic exhibit a bell-shape behavior, typically observed in other web workloads [39, 41, 40], with load peaks during weekdays and a noticeable decrease in load intensity over the weekend (days six and seven). On the other hand, Figure 4.2-b shows that spam traffic does not present any significant daily variation. In fact, the daily numbers of emails and bytes are roughly uniformly distributed over the whole week. This stability in the daily spam traffic was previously observed in [4] for a much lighter workload, including only 5% of all emails received. Figure 4.2-c shows that the impact of this distinct behavior on the aggregate traffic is a smoother variation in the number of emails per day. Finally, the variations in the aggregate number of bytes and in the number of non-spam bytes have very similar patterns, as shown in figure 4.2-c. This is because non-spam emails account for over 90% of all bytes received (see Table 4.1).

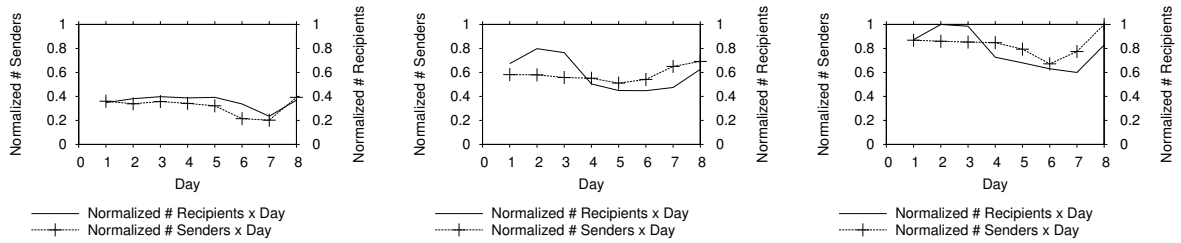
Traffic	Metric	Minimum	Maximum	Average	CV
Non-Spam	# Emails/Hr.	232 - 435	703 - 4,676	513 - 1,213	0.20 - 0.74
	# MBytes/Hr.	4 - 11	46 - 349	23 - 86	0.45 - 0.98
Spam	# Emails/Hr.	194 - 776	1,081 - 2,086	781 - 1,007	0.12 - 0.36
	# MBytes/Hr.	1.7 - 5.7	6.1 - 18.4	4.3 - 8.0	0.15 - 0.45
Aggregate	# Emails/Hr.	500 - 1,210	1,681 - 6,762	1,294 - 2,134	0.13 - 0.55
	# MBytes/Hr.	8.7 - 16.8	50 - 367	29 - 93	0.36 - 0.93

Table 4.3: Summary of Hourly Load Variation

The same overall behavior is observed for the hourly load variations, as illustrated in Figure 4.3, for a typical day. Like in [1, 2], traditional non-spam email traffic (Figure 4.3-a) presents two distinct and roughly stable regions: a high load diurnal period, typically from 7AM to 7PM, (i.e., working hours), during which the central server receives between 65% and 73% of all daily non-spam emails, and a low load period covering the evening, night and early morning. On the other hand, the intensity of spam traffic (Figure 4.3-b) is insensitive for the time of the day: the fraction of spams that arrives during a typical diurnal period is between 50% and 54%. Figure 4.3-c shows the impact of spam on the aggregate traffic is a less pronounced in hourly variation of the number of emails received than in daily variations.

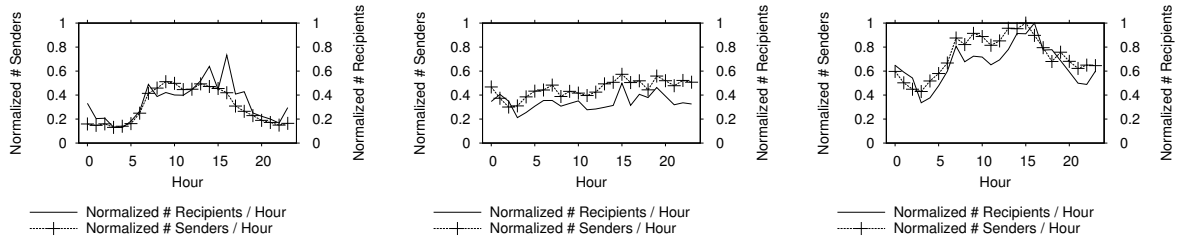
Table 4.3 summarizes the observed hourly load variation statistics. For each workload, it presents the ranges for minimum, maximum, average and coefficient of variation of the numbers of emails and bytes received per hour, on each day. In each column we show the lowest followed by the greatest values found, respectively, for the metric in a per hour base for each day of our eight-day log. Note the higher variability in the numbers of emails and bytes in the non-spam workload. Moreover, for any of the three workloads, a higher coefficient of variation is observed in the number of bytes, because of the inherent variability of email sizes. These results are consistent with that found for the daily variations in the numbers of emails and bytes, in each workload (see Table 4.1). Qualitative similar results were also found for load variations on a minute basis. The coefficients of variation of the number of emails per minute vary in the ranges of 0.45-0.78 and 0.46-0.91 for the spam and non-spam workloads, respectively. The coefficients of variation of the number of bytes per minute are in the ranges of 0.70-1.03 and 1.37-1.75, in the two workloads.

One key conclusion is that, on various time scales, whereas traditional email traffic is concentrated on diurnal periods, the arrival rate of spam emails is roughly stable over time. One question that comes up is whether this difference is also observed on a per-sender basis. We analyzed the hourly traffic generated by each of the 50 most active spam-only senders and strictly non-spam senders. We found that each of the 50 strictly spam senders sent, on average, 53% of its daily emails during the day. In contrast, the strictly non-spam senders selected send, on



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.4: Daily Variation of Number of Senders and Recipients (Normalization Parameters: Max # Senders=10,089, Max # Recipients=25,218)



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.5: Hourly Variation of Number of Senders and Recipients (Normalization Parameters: Max # Senders=956, Max # Recipients=2,802)

average, 63% of their emails during the same period. Similar results were obtained for the 100, 200 and 500 most active senders in each group. Thus, each spammer independently sends almost half of its emails over night, when computing and networking resources are mostly idle. We conjecture that, by using automatic tools, spammers try to maximize their short-term throughput, sending at the fastest rate they can get through without being noticed, throughout the day.

4.3.2 Distinct senders and recipients

This section analyzes temporal variations in the numbers of distinct senders and recipients. Daily variations and hourly variations for a typical day are shown in Figures 4.4 and 4.5, respectively. As before, we show normalized measures, expressed as fractions of the peak number of senders and recipients in the aggregate traffic, in the period. The normalization parameters are given in the captions of the figures.

As observed in the load variation, temporal variations in number of distinct email senders in the spam workload present significantly different behavior from those observed in the non-spam email workload. Whereas the number of distinct legitimate email senders does present weekly patterns, the number of distinct spammers is roughly stable

Non-Spam Traffic

Traffic	Metric	Minimum	Maximum	Average	CV
Non-Spam	# Recip./Hr.	228 - 383	589 - 2,883	411 - 978	0.21 - 0.58
	# Senders/Hr.	107 - 136	225 - 937	160 - 332	0.14 - 0.61
Spam	# Recip./Hr.	485 - 1,174	1,397 - 4,095	955 - 2,371	0.15 - 0.41
	# Senders/Hr.	147 - 406	548 - 925	433 - 577	0.10 - 0.24
Aggregate	# Recip./Hr.	828 - 1,672	2,480 - 6,580	1,505 - 3,179	0.20 - 0.41
	# Senders/Hr.	256 - 541	828 - 1,614	623 - 885	0.12 - 0.33

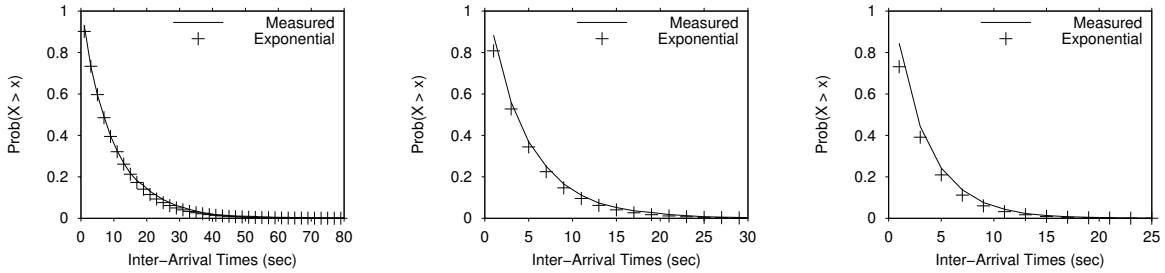
Table 4.4: Summary of Hourly Variation of Number of Distinct Recipients and Senders

over the eight days (with a slight increase by the 7^{th} day), as shown in Figures 4.4-a and 4.4-b. This difference is even more striking on a hourly basis, as shown in Figures 4.5-a and 4.5-b. Again, we speculate that the inherently different nature of the email senders in each workload (automatic tools versus human beings) are responsible for it. In the aggregate traffic, the significant variations observed in the non-spam email traffic are somewhat smoothed out by the roughly stable number of spammers, as shown in Figures 4.4-c and 4.5-c.

Regarding the daily variations in the number of distinct recipients, shown in Figure 4.4, no clear distinction between spam and non-spam traffic is observed. We found that the number of distinct spam recipients actually decreased significantly by the fourth day. We could not find any reason to explain this odd behavior. However, on a hourly basis, we found that whereas the number of distinct recipients of non-spam emails is higher during the day, the number of distinct spam recipients is roughly stable over time, as illustrated in Figure 4.5, for a typical day. These results are summarized in Table 4.4, which shows, for each of the three workloads, the observed ranges for the minimum, maximum, average and coefficient of variation of the number of distinct recipients and senders per hour.

We also measured the correlation between the number of distinct senders and the number of distinct recipients per hour, on each day, for the three workloads. We found coefficients of correlation between 0.90 and 0.99 in the non-spam workload, and between 0.58 and 0.89 in the spam workload. The lower correlation seems to indicate that there is a larger overlap in the distribution lists of typical spammers. This overlap may be the result of spammers using similar automatic tools to create their targets, trading their distribution lists to extend their reach or obtaining the same distribution list from existing web services [82]. The sharing of recipients among traditional email senders is most probably due to the fixed number of recipients, who are members of a somewhat closed university community.

In summary, our results show that, unlike traditional non-spam email traffic, which exhibits clear daily patterns,



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.6: Distribution of Inter-Arrival Times

with load peaks during the day, the numbers of spam emails, spam bytes, distinct active spammers and distinct spam recipients are roughly insensitive to the period of measurement, remaining mostly stable during the whole day, for all days analyzed.

The fundamental differences between spam and non-spam traffic discussed in this section may be explained by the inherently distinct nature of their sources. Spammers are driven by the goal of reaching as many targets as possible, without being detected. To do so, they use automatic tools to (roughly) uniformly spread the flooding of emails over time to avoid being noticed. Thus, a spam transmission is basically a unilateral action. The transmission of a non-spam email, on the other hand, is the result of a bilateral relationship [9, 60, 80, 13]. It is typically initiated by a human being, driven by some social reason (i.e., work, leisure), during his/her active hours.

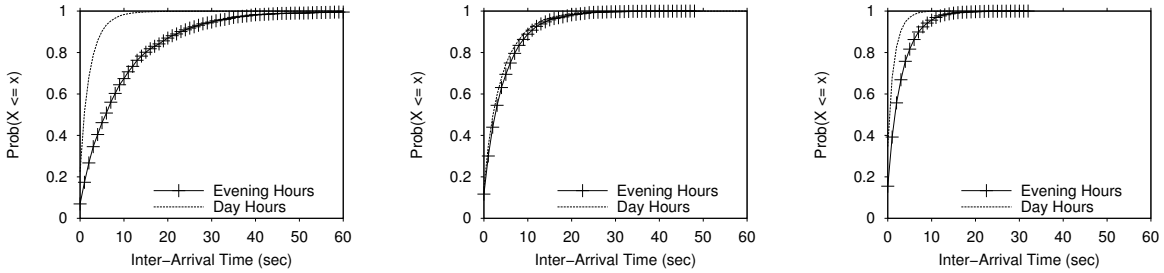
4.4 Email traffic characteristics

This section analyzes the characteristics of email traffic for the spam, non-spam and aggregate workloads. The email arrival process is characterized in Section 4.4.1. The distributions of email sizes and number of recipients per email are analyzed in Sections 4.4.2 and 4.4.3, respectively. For each workload characteristics, we discuss the differences between spam and non-spam, pointing out the impact of the former on the aggregate workload.

4.4.1 Email arrival process

In this section the email arrival process in each workload is characterized during periods of roughly stable arrival rate in order to avoid spurious effects due to data aggregation. In the spam workload such periods are typically whole days, whereas in the non-spam and aggregate workloads, different stable periods are observed during the day and over night.

We found that email inter-arrival times are exponentially distributed in all three workloads, as illustrated in Figures 4.6-a, 4.6-b and 4.6-c, for typical periods of stable arrival rate in the non-spam, spam and aggregate work-



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.7: Sensitivity of Inter-Arrival Time Distribution to the Period Analyzed

Workload	Inter-Arrival Times		Exponential Parameter λ
	Mean (sec)	CV	
Non-Spam	2.1 - 9.7	1.12 - 1.90	0.10 - 0.48
Spam	3.6 - 4.9	1.08 - 1.99	0.21 - 0.26
Aggregate	1.3 - 3.2	1.07 - 1.73	0.31 - 0.75

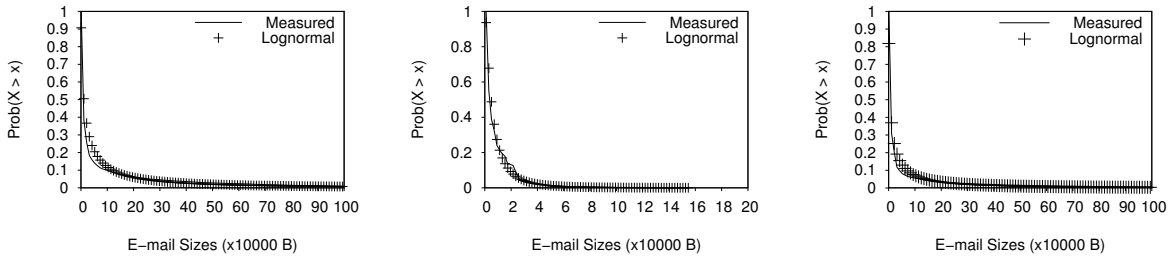
Exponential (PDF): $p_X(x) = \lambda e^{-\lambda x}$.

Table 4.5: Summary of the Distribution of Inter-Arrival Times

loads, respectively. To evaluate the sensitivity of the distribution to the period of measurement, we looked into the distribution of inter-arrival times observed in different periods. Figure 4.7 presents the cumulative distributions of inter-arrival times for two distinct periods, one during the day and the other during the evening, for each workload. Figure 4.7-a shows that non-spam email arrivals are burstier during the day, with around 86% of all inter-arrival times within 5 seconds. During the evening, only 40% of non-spam inter-arrival times are under 5 seconds. On the other hand, the distributions are the same in both periods in the spam workload, as shown in Figure 4.7-b. Figure 4.7-c shows somewhat intermediate results for the aggregate workload.

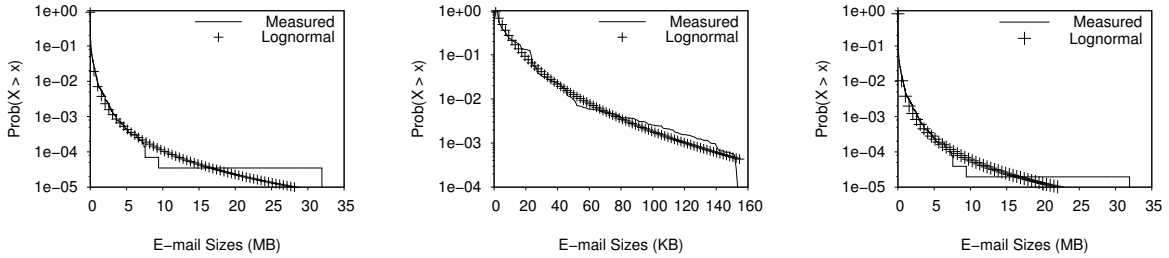
Table 4.5 summarizes our findings. It shows the ranges of the mean and coefficient of variation of the inter-arrival times measured in seconds as well as the range values of the λ parameter (email arrival rate) of the best-fitted exponential distribution, for all periods analyzed, in each workload. Note that λ remains roughly stable across all periods analyzed in the spam workload. In fact, the peak arrival rate is only 25% higher than the minimum. On the other hand, the non-spam arrival rates vary by as much as a factor of five across the periods analyzed. Aggregate traffic exhibits somewhat lower variations, as discussed in Section 4.3.

Our results are in contrast with prior work, which found that the distribution of email inter-arrival times at four email servers is a combination of a Weibull and a Pareto distributions [1, 2]. However, like in our workloads (see Table 4.5), the reported coefficient of variation of their inter-arrival times was close to 1. Moreover, our results are in close agreement with other previous work which found a non-stationary Poisson process to model with reasonable accuracy SMTP connection arrivals [79].



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.8: Distribution of Emails Sizes (Body)



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.9: Distribution of Email Sizes (Tail)

Workload	Email Sizes		Log-normal Parameters	
	Mean (KB)	CV	μ	σ
Non-Spam	34 - 75	4.27 - 5.24	8.77 - 9.72	1.72 - 1.83
Spam	5 - 9	1.37 - 1.98	7.97 - 8.51	1.03 - 1.26
Aggregate	19 - 44	5.57 - 6.39	7.97 - 8.95	1.86 - 1.94

Log-normal (PDF): $p_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$.

Table 4.6: Summary of the Distribution of Email Sizes

4.4.2 Email size

We found that the distribution of email sizes is most accurately approximated, both at the body and at the tail of the data, by a Log-normal distribution, in all three workloads, as illustrated in Figures 4.8 and 4.9, for a typical day. Figures 4.8 (a-c) show the complementary cumulative distribution functions of the data and fitted Log-normal distributions for the non-spam, spam and aggregate workloads, respectively. Semi-log plots of the same distributions are shown in Figures 4.9(a-c). Table 4.6 presents the ranges of the mean, coefficient of variation and parameter values of the best-fitted Log-normal distribution for each workload, in all days analyzed. These results are consistent with those reported in previous email workload characterizations [1, 2].

Table 4.6 shows that the sizes of non-spam emails are much more variable and have a much heavier tail in comparison with legitimate email sizes. In our workloads, approximately 83% and 61% of spam and non-spam emails, respectively, have sizes under 10 KB. However, whereas only 1% of all spams are sized above 60 KB,

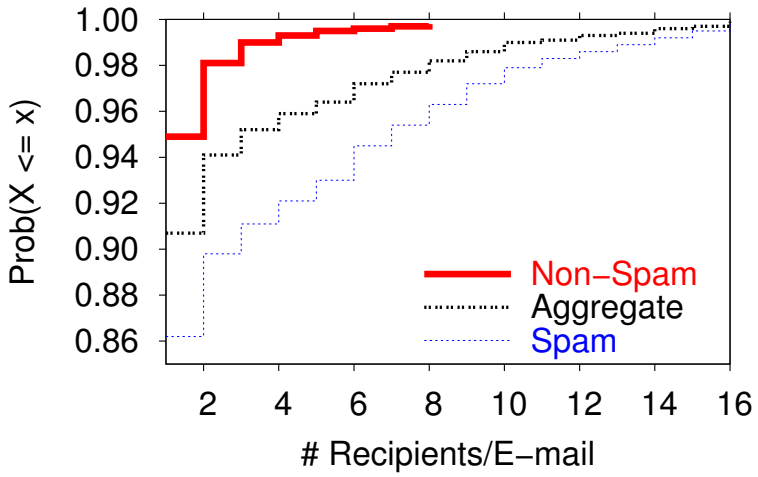


Figure 4.10: Distribution of Number of Recipients per Email

approximately 13% of non-spam emails have sizes above that mark. Note that, as explained in Section 4.2.1, the size-based rule of SpamAssassin influences only email size distribution that models less than 1% of the data in spam traffic, so this rule has almost no impact in our models. These results are consistent with those reported in [1, 2] for non-spam email traffic. The impact of spam on the aggregate traffic is, thus, a decrease in the average email size but an even more variability of email size.

We draw the following insights from these results. First, in our workload, spammers typically send (a large number of) short emails, possibly with no attachment. Second, as performed by some system administrators (including our central server administrator), email size *might* be used, together with other filtering techniques, to improve the effectiveness of spam detection.

4.4.3 Number of recipients per email

This section characterizes the distribution of the number of distinct recipients per email message. Since this distribution is discrete, we do not apply the same fitting technique as in previous sections. Instead, like in [1, 2], we subdivide each distribution into k buckets. Each bucket is characterized with an average probability, calculated over the eight days analyzed. Jointly, these probabilities represent the distribution of the number of recipients per email. For each workload, we choose a value of k in order to limit the probability of an email with more than k recipients per email to below 0.002 [1, 2]. The values of k for the non-spam, spam and aggregate workloads are $k_{non-spam} = 8$, $k_{spam} = 16$ and $k_{aggregate} = 16$, respectively.

Figure 4.10 shows the cumulative distributions for the three workloads. As mentioned in Section 4.2.3, spams

are typically addressed to a larger number of recipients than non-spam emails. Whereas, on average, 95% of all non-spam emails are addressed to one recipient, 86% of spams have a single destination. Furthermore, the distribution is heavier tailed in the spam traffic when compared to non-spam workload, possibly due to the use of automatic tools. Since almost half of all emails are spam, the distribution of the number of recipients per email in the aggregate workload is strongly influenced by the heavy tailed behavior. The authors of [1, 2] found an even heavier tail in the distribution of the number of recipients per email when compared with our findings. In that study, even though 94% of all emails are addressed to a single recipient, 20 buckets were necessary to cover 99.8% of all emails.

4.5 Analyzing email senders and recipients

This section further analyzes email senders and recipients in our three workloads. Popularity of email recipients and senders is analyzed in Section 4.5.1. Section 4.5.2 analyzes temporal locality among email recipients and senders. Finally, Section 4.5.3 analyzes contact lists of senders and recipients, i.e., the lists of contacts each sender (recipient) exchanges email with.

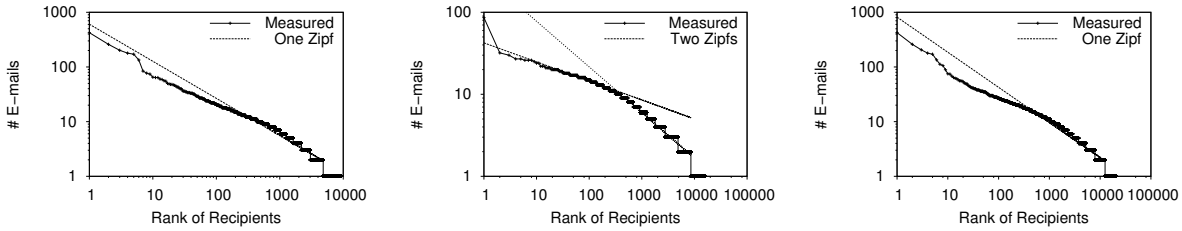
4.5.1 Popularity

Popularity of objects have been repeatedly modeled with a Zipf-like distributions (Prob(access object with rank $i = C/i^{\alpha}$, where $\alpha > 0$ and C is a normalizing constant [67]) in many contexts, including web, email and streaming media [39, 40, 41, 80]. A Zipf-like distribution appears as a straight line in the log-log plot of popularity versus object rank. However, two roughly linear regions have been observed in the log-log plots of some streaming media and peer-to-peer workloads [40, 41, 42, 83]. The concatenation of two Zipf-like distributions has been suggested as a good model in such cases [40, 41].

In the following sections, we analyze the log-log plots of email recipient and sender popularity, measured in terms of both the number of emails and the number of bytes received and sent. To assess the accuracy of our proposed models, we measure the R^2 factor of the linear regression [65] for each single Zipf-like distribution found. In our models, the values of R^2 are above 0.95 in all cases. A value $R^2 = 1$ corresponds to perfect agreement.

4.5.1.1 Recipient popularity

This section analyzes the popularity of email recipients in our three workloads. A characterization of the number of emails per recipient is presented first. The distribution of the number of bytes per recipient is discussed later in this section.



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.11: Distribution of Number of Emails per Recipient

Number of emails per recipient

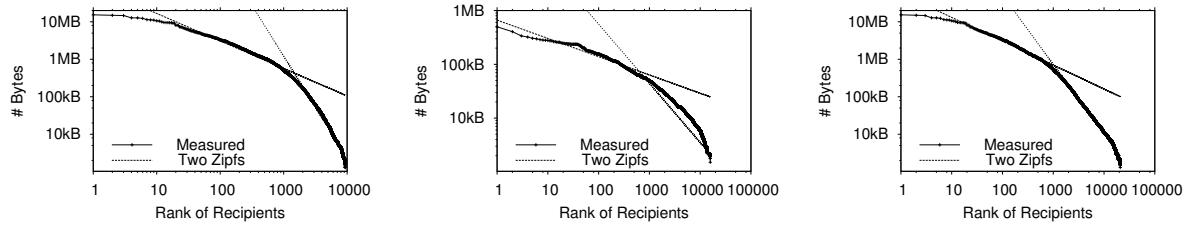
Figures 4.11-a, 4.11-b and 4.11-c show the log-log plots of the number of emails per recipient for the non-spam, spam and aggregate workloads, respectively, on a typical day. Given the large fraction of users who receive only one email per day in all three workloads, we choose to characterize the number of emails per recipient using a combination of a fixed constant probability, for recipients who receive only one email, and a probability distribution for the remaining users.

The curves in Figure 4.11 present significantly different patterns for recipients of two or more emails. Whereas Figures 4.11-a and 4.11-c show straight lines, Figure 4.11-b shows two distinct linear regions in the spam workload. These results are representative of all days analyzed. Thus, for recipients of two or more emails per day, we modeled the number of emails per recipient with a single Zipf-like distribution, for the non-spam and aggregate workloads, and with the concatenation of two Zipf-like distributions, for the spam workload. Figures 4.11(a-c) show the curves for the best fitted Zipf-like distributions in each case. The roughly flat curve over the most popular spam recipients implies they receive around the same number of spams on the analyzed day. This is true for all days analyzed, and may be explained by the larger average number of recipients per spam (section 4.4.3) and by the larger fraction of recipients shared among spammers (section 4.3.2). Again, the inherent difference between the unilateral relationship established by spam traffic and and the bilateral, socially-driven relationship established between non-spam senders and recipients may incur significantly different traffic patterns.

Number of bytes per recipient

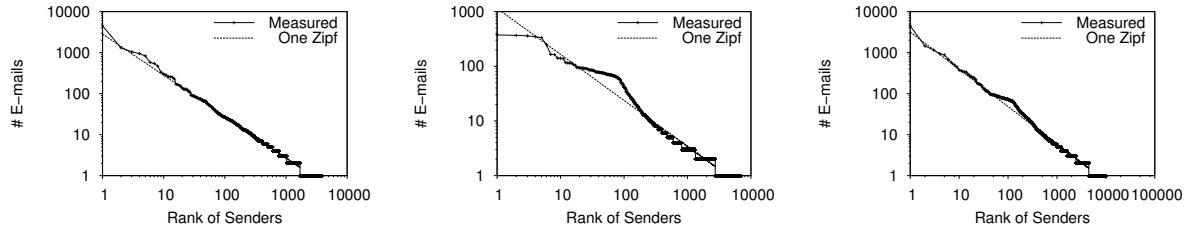
Figures 4.12-a, 4.12-b and 4.12-c show the log-log plots of the number of bytes per recipient observed on a typical day for the non-spam, spam and aggregate workloads, respectively. All three graphs show two clear linear regions, and are representative of the results found in all days analyzed. Thus, we model the number of bytes per recipient with the concatenation of two Zipf-like distributions, also shown in the graphs of Figure 4.12.

The discrepancy between these results and the number of emails per recipient distributions in the non-spam and



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.12: Distribution of Number of Bytes per Recipient



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.13: Distribution of Number of Emails per Sender

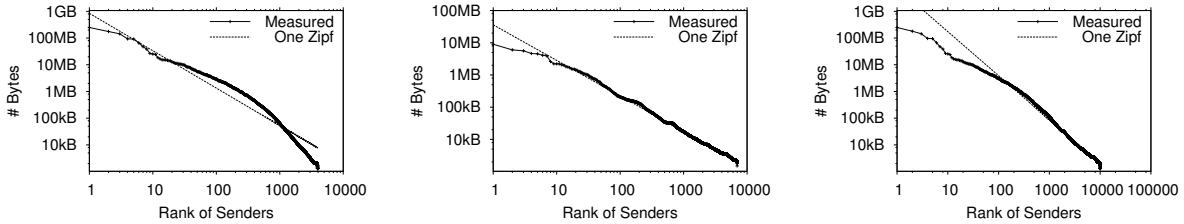
aggregate workloads may be due to the higher variability in non-spam email sizes (see Section 4.4.2). Moreover, we found that the correlation between the number of emails and the number of bytes received by each user is typically weak, ranging between 0.18-0.27, 0.50-0.66, and 0.18-0.28 for the non-spam, spam and aggregate workloads, respectively. Thus, in each workload, the users who receive the largest number of emails are not necessarily the ones who receive the largest volume of traffic.

4.5.1.2 Sender popularity

This section characterizes sender popularity. We first present the results for the number of emails per sender. Analysis of the number of bytes per sender is discussed at the end of the section.

Number of emails per sender

Figures 4.13-a, 4.13-b and 4.13-c show the log-log plots of the number of emails per sender, on a typical day, in the non-spam, spam and aggregate workloads, respectively. The three curves show similar behavior. As observed for email recipients, there is a large number of senders that send only one email on a typical day. Moreover, the portion of the curve covering the remaining senders is well approximated with a straight line. Thus, in all workloads, we model the number of emails per sender with the concatenation of a constant probability, for single-message senders, and a Zipf-like distribution (shown in the Figure 4.13(a-c)), for the remaining senders.



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.14: Distribution of Number of Bytes per Sender

Traffic	Popularity Metric	% receive/send one email/day	Zipf Model	% Data	Prob.	α
Non-Spam	# Emails	48-62	Single	100	1	0.52-0.68
	# Bytes	—	1^{st}	3-21	0.05-0.26	0.62-0.85
			2^{nd}	79-97	0.74-0.95	1.67-3.20
Spam	# Emails	29-59	1^{st}	2-3	0.05-0.08	0.22-0.30
			2^{nd}	97-98	0.92-0.95	0.47-0.64
	# Bytes	—	1^{st}	4-62	0.05-0.66	0.34-0.56
			2^{nd}	38-96	0.34-0.95	1.10-3.87
Aggregate	# Emails	29-49	Single	100	1	0.95-0.97
	# Bytes	—	1^{st}	3-85	0.05-0.88	0.64-1.32
			2^{nd}	15-94	0.13-0.95	1.85-8.62

Table 4.7: Summary of Distributions of Recipient Popularity

Note that the curve in Figure 4.13-b flattens out over a few of the most popular spammers. However, since they represent a very small fraction of all spammers, a straight line is a reasonably good fit for the curve. Nevertheless, it is interesting to note that the fitting of the single Zipf-like distribution is more accurate for the non-spam and aggregate workloads. This result is in agreement with the previous observation of a Zipf-like distribution for the number of connections per spam source in a 2004 SMTP workload [5].

Number of bytes per sender

A single Zipf-like distribution was found to be a good approximation of the number of bytes per sender, in all three workloads, as illustrated in Figure 4.14. We point out that the high variability in email sizes, which might be responsible for a noticeable flat region over the recipients with the largest number of emails (Section 4.5.1.1), is less effective here because of the larger number of emails per sender. Furthermore, unlike observed for recipients, the correlation between the number of emails and the number of bytes for each sender was typically high, in the ranges of 0.68-0.88, 0.66-0.80 and 0.70-0.87, for the non-spam, spam and aggregate workloads, respectively.

Our main conclusions with respect to email sender and recipient popularity are:

Workload	Popularity Metric	% receive/send one email/day	α
Non-Spam	# Emails	54-68	0.993-1.251
	# Bytes	—	1.72-2.08
Spam	# Emails	55-67	0.781-0.996
	# Bytes	—	0.915-1.192
Aggregate	# Emails	53-61	0.937-0.987
	# Bytes	—	1.185-1.775

Table 4.8: Summary of Distributions of Sender Popularity

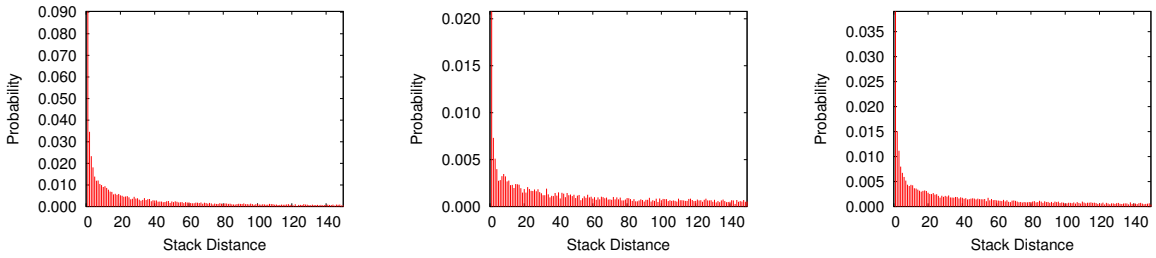
- The distributions of the number of non-spam emails per sender and recipient follow, mostly, a Zipf-like distribution. This result is consistent with previous findings that the connections between email senders and recipients are established using a distribution that follows a Power Law (e.g., a Zipf-like distribution) [5, 80, 13].

- The distribution of the number of spams per recipient does not follow a true power law, but rather, presents a flat region over the most popular recipients. This may be due to large spam recipient lists and large number of recipients shared among spammers. The number of spams per sender is reasonably well approximated by a Zipf-like distribution.

- In all three workloads, the number of bytes per recipient is most accurately modeled by two Zipf-like distributions. In the case of the non-spam and aggregate workloads, this is probably due to the high variability in email size. The distribution of the number of bytes per sender is well modeled by a single Zipf-like distribution in all three workloads.

Tables 4.7 and 4.8 summarize our findings. Table 4.7 presents the ranges of the observed percentage of recipients that receive only one email on a typical day. It also shows the range of parameter values for the Zipf-like distributions that best fit the data for the remaining recipients. For the cases where two Zipf-like distributions is the best model, it shows, for each single distribution, the total probability and percentage of recipients that fall within the corresponding region of the curve as well as the value of the α parameter. Table 4.8 shows similar data for the distributions of sender popularity.

We point out that the skewed distributions of the number of emails and bytes per sender and per recipient suggest that sender and recipient popularity could be used to improve the effectiveness of spam detection techniques. For instance, on a typical day, on average, 53% of the spams and 63% of the spam bytes originate from only 3% of all strictly spam senders. Furthermore, around 40% of these spammers are among the most active throughout the eight days covered by our log. Thus, the insertion of these popular spammers into black lists could significantly reduce the number of spams accepted by the server. Similar results are observed for the strictly non-spam senders,



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.15: Histograms of Recipient Stack Distances

suggesting the use of white lists to reduce the overhead of email scanning. Finally, the concentration of spams into a small fraction of recipients, who remain among the most popular through several days, suggests that spam detection techniques might use the email destination to improve its success rate.

4.5.2 Temporal locality

Temporal locality in an object reference stream implies that objects that have been recently referenced are more likely to be referenced again in the near future [81]. A previously proposed method to assess the temporal locality present in a reference stream is by means of the stack distances distribution [81]. A stack distance measures the number of references between two consecutive references to the same object in the stream. Short stack distances imply strong temporal locality.

This section analyzes temporal locality among recipients and among senders in our three workloads. We start by creating one email stream for each workload and day analyzed, preserving the order of email arrivals in the corresponding workload and day. To assess temporal locality among recipients, each email in a stream is replaced by its recipient list, creating, thus, a recipient stream. The distribution of stack distances in the recipient stream is then determined. Similarly, to assess temporal locality among senders, each email is replaced by its sender and the distribution of stack distances is determined.

Section 4.5.2.1 analyzes temporal locality among recipients. Temporal locality among senders is discussed in Section 4.5.2.2.

4.5.2.1 Temporal locality among recipients

Figures 4.15-a, 4.15-b and 4.15-c show histograms of recipient stack distances, for distances shorter than 150, observed on a typical day in the non-spam, spam and aggregate workloads, respectively. The complementary

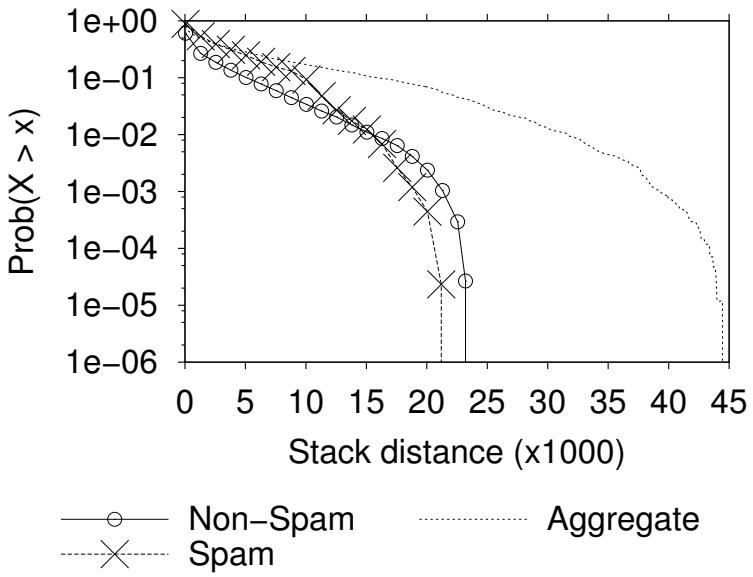


Figure 4.16: Complementary Cumulative Distributions of Recipient Stack Distances

Workload	Mean (x1000)	CV	Weibull Parameters	
			α	β
Non-Spam	0.7-1.9	2.0-2.3	0.08-0.12	0.35-0.41

Workload	Mean (x1000)	CV	Gamma Parameters	
			α	β
Spam	2.2-3.5	1.1-1.6	0.36-0.54	4877-23741
Aggregate	3.1-5.4	1.4-1.9	0.29-0.37	8465-47657

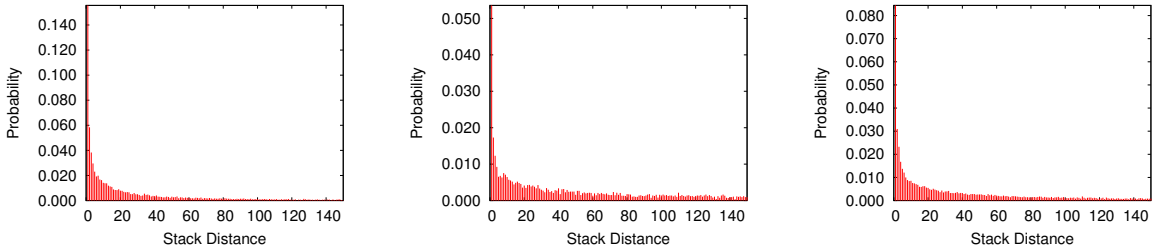
Gamma (PDF): $p_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{x}{\beta}}$;
Weibull (PDF): $p_X(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}$

Table 4.9: Summary of the Distributions of Recipient Stack Distances

cumulative distributions of recipient stack distances, measured on the same day, are shown in Figure 4.16. Note the log scale on the y-axis in Figure 4.16.

We draw the following conclusions. First, there is a higher probability of very short stack distances, and thus, stronger temporal locality, in the non-spam workload. Second, the distribution of stack distances has a slightly heavier tail for non-spam recipients than for spam recipients (see discussion below). Finally, the impact of spam on the aggregate traffic is a significant reduction on the temporal locality among recipients, evidenced by an even heavier tail in the stack distance distribution. A summary of best-fitted distributions for recipient stack distances in each workload is presented in Table 4.9. Note the small mean and large coefficient of variation in the non-spam workload.

In search for an explanation for the significantly different temporal locality observed among spam and non-spam email recipients, we defined two regions in each stack distance distribution: the head and the tail. The head



(a) Non-Spam (b) Spam (c) Aggregate

Figure 4.17: Histograms of Email Sender Stack Distances

(tail) consists of the short (large) stack distances so that the total probability of the region does not exceed $p = 0.2$. We then defined two sets of recipients: one including all recipients for which the stack distances in the head are observed, and the other containing the recipients for which the stack distances in the tail are observed. We make three observations. First, the sets are mostly disjoint, in both spam and non-spam workloads. Second, the number of distinct recipients in the head region is a significant fraction (over 30%) of the number of daily recipients, in both workloads. Third, whereas the number of recipients in the tail of the non-spam distribution is significant, the number of recipients in the tail of the spam distribution corresponds to only 4% of daily spam recipients.

These findings, jointly, imply that there are, at least, two distinct and non-negligible sets of non-spam recipients. These sets correspond to two classes of email users with very distinct behavior: those who make intense use of email for communication and, thus, receive bursts of emails from their peers, mostly during the day, when they are active, and those who send, and thus, receive email only sporadically. These sets are not clearly defined among spam recipients, since the transmission of a spam is driven by the spammer, who acts mostly independently of the recipient's intimacy with email systems.

4.5.2.2 Temporal locality among senders

The histograms of email sender stack distances observed on a typical day, for the non-spam, spam and aggregate workloads, are shown in Figures 4.17-a, 4.17-b and 4.17-c, respectively. Figure 4.18 the corresponding complementary cumulative distributions. As observed for email recipients, there is a higher probability of very short and very large stack distances for non-spam email senders. The distribution for the aggregate workload has an even heavy tail, implying a significant reduction on temporal locality among email senders due to spam. A summary of the best-fitted distributions for email sender stack distances is given in Table 4.10.

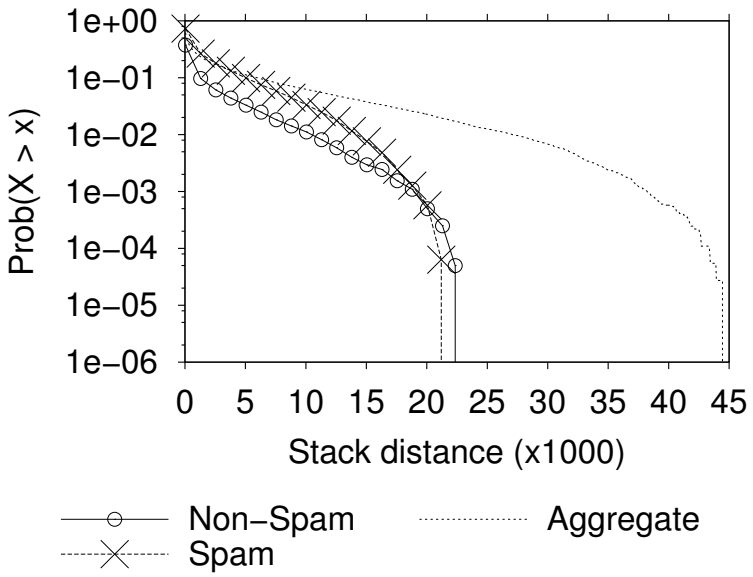


Figure 4.18: Complementary Cumulative Distributions of Email Sender Stack Distances

Workload	Mean	CV	Weibull Parameters	
			μ	σ
Non-Spam	287-644	3.35-3.78	4.34-6.63	1.58-1.70
Spam	960-1567	1.88-2.51	5.52-7.80	1.23-1.42
Aggregate	1403-2189	2.32-3.23	6.03-7.91	1.36-1.56

Table 4.10: Summary of the Distributions of Email Senders Stack Distances

4.5.3 Contact lists

This section analyzes the contacts established between senders and recipients through email in the spam, non-spam and aggregate workloads. Our basic assumption is that, in both spam and non-spam traffics and thus in the aggregate traffic as well, users (i.e., senders/recipients) have a defined list of peers they often have contact with (i.e., they send/receive an email to/from) [9]. In non-spam traffic, contact lists are the consequence of social relationships on which users' communications are based. In spam traffic, on the other hand, the lists used by spammers to distribute their solicitations are created for profit and, generally, do not reflect any form of social interaction.

In each workload, the **c**ontact list of a sender s is defined as the set of recipients to which s sent at least one email (spam, non-spam or either one, in the aggregate workload) in the period analyzed. Similarly, the contact list of a recipient r is defined as the set of senders from which r received at least one email in the period analyzed. The size of a sender/recipient contact list is given by the number of recipients/senders in its contact list.

A sender/recipient contact list is built dynamically, over time, as new emails are exchanged. Moreover, contact

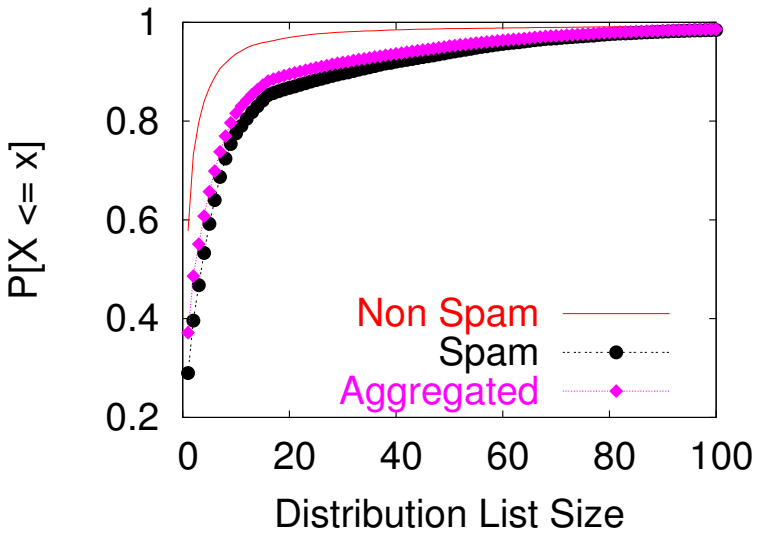


Figure 4.19: Cumulative Distribution of Size of Sender Contact Lists

Workload	Sizes of Sender Contact Lists				
	Mean	CV	Size = 1 (%)	Size < 8 (%)	Size < 400 (%)
Non-Spam	6.9	9.0	58	92	99.8
Spam	13.8	3.2	29	72	99.6
Aggregate	12.6	4.7	37	77	99.6

Table 4.11: Distribution of Sizes of Sender Contact Lists

lists certainly may change over time. However, we expect them to be much more stable than other workload aspects such as email inter-arrival times and sender popularity (see Section 4.5.1). Thus, in contrast to other analysis presented in this chapter, this section analyzes the distribution of the sizes of recipient and sender contact lists in each of our three workloads considering the entire eight-day log. The distribution of the size of sender contact lists is analyzed in Section 4.5.3.1. Section 4.5.3.2 analyzes the size of recipient contact lists.

4.5.3.1 Sender contact lists

Figure 4.19 shows the cumulative distributions of sender contact list sizes in each workload, for list sizes varying from 1 to 100, which account for more than 98% of all senders in our three workloads. Note that the distribution is much more skewed towards shorter contact lists among non-spam senders. In other words, typical spammers send emails indiscriminately to a much larger number of distinct recipients. Dictionary attacks certainly contribute to generating large contact lists. Non-spam senders, on the other hand, have, on average, a smaller number of contacts, established through some sort of social relationship. Note that the longer contact lists of spammers significantly impact the distribution in the aggregate traffic.

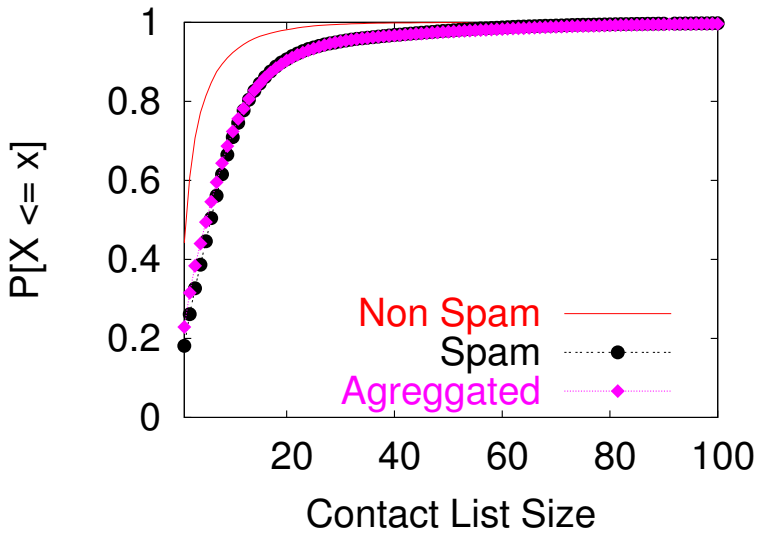


Figure 4.20: Cumulative Distribution of Size of Recipient Contact Lists

Table 4.11 shows the mean and coefficient of variation of sender contact list sizes in the non-spam, spam and aggregate workloads. It also shows the percentage of senders that fall into different ranges of list sizes. On average, spammers have contact to twice as many distinct recipients than non-spam senders. Furthermore, the fraction of sender contact lists with only one recipient over the eight-day period analyzed is significantly higher in the non-spam traffic. In fact only 8% of non-spam sender contact lists have size greater than eight, whereas 28% of spammer contact lists have size above that mark. Nevertheless, it is interesting to note the high variability in the contact list sizes among non-spam senders. This reflects the natural variability in the behavior of legitimate email users: some have (email) contact to a much larger number of people than others.

Therefore, spammer contact lists are longer than hammer contact lists, on average, but much less variable. Furthermore, the clear distinction between the distributions of contact list sizes among spammers and non-spam senders significantly impact the distribution in the aggregate traffic.

4.5.3.2 Recipient contact lists

Figure 4.20 shows the cumulative distribution of recipient contact lists sizes in each workload, for list sizes varying from 1 to 100, which account for more than 99% of all recipients. As observed among senders, the distribution of contact list sizes is much more skewed towards shorter lists among non-spam recipients. The impact of spam on the aggregate traffic is clear: the aggregate distribution is much more heavy tailed than the non-spam distribution.

Table 4.12 shows the mean and coefficient of variation of recipient contact list sizes in the non-spam, spam and

Workload	Sizes of Recipient Contact Lists				
	Mean	CV	Size = 1 (%)	Size < 6 (%)	Size < 30 (%)
Non-Spam	3.7	1.4	44	81	99.5
Spam	9.6	1.3	18	45	95.2
Aggregate	9.5	1.5	23	49	95.1

Table 4.12: Distribution of Sizes of Recipient Contact Lists

aggregate workloads. It also shows the percentage of recipients that fall into different ranges of list sizes. Note that, on average, the number of distinct spammers with whom recipients have contact is almost three times the number of contacts of non-spam recipients. Furthermore, the fraction of recipient contact lists with only one sender is significantly higher in the non-spam traffic. In fact, only 19% of non-spam recipient contact lists have size greater than 5, whereas 55% of spam recipient contact lists have size above that mark.

In summary, the analysis presented in this section shows that the sizes of contact lists can strongly distinguish non-spam and spam traffic and that they strongly impact the distributions observed in the aggregate traffic. An attempt to exploit this distinction for clustering senders and recipients into communities that share similar contact lists, and use such communities for detecting spam is presented in [11].

4.6 Conclusion

This chapter provides an extensive analysis of a spam traffic, uncovering characteristics that significantly distinguish it from legitimate email traffic and assessing how the aggregate traffic is affected by the presence of a large number of spam messages.

Our characterization, based on the information available on the email headers, revealed that email arrival processes, email sizes, number of recipients per email, popularity and temporal locality among recipients and the sizes of sender and recipient contact lists are some key workload aspects where spam traffic significantly deviates from traditional non-spam traffic. We believe that such discrepancies are consequence of the inherently different nature of email senders in each traffic. Traditional email senders are usually human beings who use emails to interact socially with their peers. Spammers, instead, typically use automatic tools to generate and send their emails to a multitude of “potential", mostly unknown, users. A further analysis of the relationship between spammers and their recipients and of spammer behavior, in general, is presented in the Chapters 5 and 7.

Chapter 5

Characterization of Graphs of Legitimate and Spam Email

5.1 Introduction

In this chapter we present a characterization of the way spammers, legitimate senders and their peers develop their relationship by exchanging emails. In order to do so, we model legitimate and spam email traffic as graphs and use these structures to identify theoretical structural metrics that can be used to differentiate them. We are also interested in finding a set of metrics that can be used, in the future, in a predictive model of spam dissemination.

Our study, which was published in a preliminary version in [9], goes beyond several other recent analyzes [6, 7] on the graphical nature of spam traffic. We deal with a different database, involving a much larger number of users and messages, and analyze a wider set of metrics, both static and dynamic. We will show that there is no single graphical metric that unequivocally distinguishes between legitimate and spam email. There are, however, several graph theoretical measures that can be combined into a probabilistic spam detection framework. These are then identified as candidates for the construction of a future spam filtering algorithm.

Our key findings are:

- Out-degree distribution of external senders, in the user and domain graphs, are well modeled by a power law. However, in the user graph, out-degrees lower than 20 are much more probable for spammers than for non-spam senders. On the other hand, with out-degree greater than 400 there are almost no spam senders. In the domain graph, the out-degree distribution shows a much higher probability for nodes with low out-degree in spam than in non-spam traffic.

- The analysis of the communication reciprocity suggests that a strong signature of spammers is its structural

imbalance between the set of senders and receivers associated with a spam sender. Even in the domain graph the difference is very clear.

- The asymmetry set shows the number of spam addresses in the difference of the in and out sets of each graph node. We found that, in both user and domain graphs, there is a strong statistical correlation between the size of the asymmetry set and the number of spammers in it. These previous two results together show that spam messages are almost never replied to.

- Our results for the clustering coefficient show that spam users have much lower cohesion in their communication than non-spam users. Moreover, it shows that spammers send emails, generally, to uncorrelated recipients.

- As a result of the communication asymmetry between spam nodes, we found that, in both graphs, there is much less probability of visiting a spam than a non-spam node during a random walk.

- The spam subgraph is a much more rapidly growing structure. Due to the size of our log, one week long, we do not find a saturation point in any of the subgraphs analyzed.

- We found that temporal locality among pairs (sender-recipient) of users is stronger in non-spam traffic than in spam traffic, showing concentration of legitimate email communications in time.

- We analyzed the entropy of the flow of messages/bytes of the communication between spammers and legitimate senders and their peers. We found that legitimate senders communicate in a much more variable way with their peers than spammers.

The remaining of this chapter is organized as follows. In section 5.2 we introduce the modeling of email traffic in terms of two graph classes and present the metrics to be studied. Section 5.3 presents the email workload used in this work. We present several theoretical graph metrics and evaluate this workload according to them in Section 5.4. Finally, we present our conclusions in section 5.5.

5.2 Graph-Based modeling of email workloads

In order to characterize spam email traffic versus non-spam we define two types of graphs [1]: a *user graph* and a *domain graph*. The vertices of the *user graph* are email senders and recipients of our log. An email sent by A to receiver B results in a link between A and B. The *domain graph* has as vertices the domains of the external senders being analyzed, and users inside the local domain. In this case, if an user B of the local domain receives an email

[1] In Chapter 3 we present the definition of graph and the set structural graph properties we analyze in this Chapter

from any user in an external domain D, we define a link between D and B. Note that, sets of users external to the local domain who share a domain are aggregated into a single node. Note, also, that the domain graph is a simply bipartite graph and not all characteristics studied will apply to it.

The edges of both graphs can take one of four forms: directed or undirected; binary (or unweighted) or weighted (e.g., by the number of emails exchanged or by the total size of the emails exchanged in bytes). These options cover most of the possibilities for direct graphical construction out of the email logs at our disposal (described in Section 5.3).

The user graph is, in principle, the most useful to identify the individual nature of users as spam or non-spam senders. In some cases these characteristics extend to the whole external domain (particularly if the spammer changes his name[2] more often than its domain) and the domain graph produces a useful aggregation of the user data. We believe that user graphs will be more effective in identifying senders of non-spam since spam senders very frequently tend to change their full email address.

The user or domain graphs can be constructed exclusively out of spam traffic, non-spam traffic, or the aggregate set of all emails. Some of the graph theoretical properties studied below will be analyzed in terms of the graphs constructed when considering the different traffics separately while others will be evaluated on selected nodes from the aggregated traffic. The selected nodes represent senders in the aggregated graph and can be divided into two classes - spam and non-spam - based on the type of emails they send. These classes do not form disjoint sets, see Table 5.2.

Given these two graph types we will analyze two types of properties: (i) structural and (ii) dynamical. The former captures the structure of social relationships between users exchanging emails, while the latter relates to how graphical properties evolve over time. As we shall show below there are distinct independent signatures of spam traffic in both structural and dynamical properties. As a consequence they should be taken together to generate a better detection procedure.

5.3 Email workloads

Measure	**N**on-Spam	**S**pam	**A**ggregate
# emails	336,580	278,522	615,102
Size of emails	11.00 GB	1.70 GB	12.71 GB
# sender users	94,985	170,664	263,144
# sender domains	20,414	48,087	59,971
# recipients	26,450	12,867	35,471

Table 5.1: Workload summary

[2]The first part of the address, located before the @.

The construction of the graphs introduced in Section 5.2 is subject to several practical constraints. Our knowledge of email traffic comes from Postfix logs of the central SMTP incoming/outgoing servers of an academic department from a large University in Brazil. Incoming emails only contain the recipients internal to the department's domain. Outgoing emails contain the full list of recipients. Moreover our data set does not contain information about emails exchanged between users external to the domain.

The logs were collected between 11/18/2004 and 12/31/2004 and contain the following data for each email: (i) received time and date; (ii) a reject flag, indicating whether connection was rejected during email acceptance (iii) Size of email[3]; (iv) sender address; (v) list of recipients and (vi) a spam flag, indicating if it was classified as spam or not by SpamAssassin [51], see Section 4.2.1 in Chapter 4 for more information about it.

The logs were sanitized and anonymized to protect the users' privacy. Statistical characteristics of the workload are in agreement with previous email traffic analysis [10, 4, 3]. Table 5.1 summarizes the data set.

5.4 Spam networks vs. legitimate email networks

Type	**E**xternal		**I**nternal	
Spam	169931	(277535)	733	(987)
Non-Spam	93666	(186607)	1319	(151973)
Spam & Non-Spam	2366	(-)	139	(-)
Total	263231	(462142)	1913	(152960)

Table 5.2: Number of unique email addresses by origin (internal or external to the domain) and classified as spam, non-spam or both. Numbers in parentheses indicate the total number of emails sent by each class.

Although spam emails originate mainly from users outside the local domain spam senders use several techniques to forge or steal local addresses (e.g., crawling the web for email addresses available at web pages, network sniffing, name dictionaries). As a result spam email does originate from the local domain both from real users and from forged ones. This mixing between regular email users and spam senders can lead to more complex email networks than might have been naively expected and poses a challenging problem for detection.

Table 5.2 summarizes the number of addresses and emails by node classes and by internal or external origin. Node classes are as defined in Section 5.2 plus a third category - Spam & Non-Spam - which is the intersection of the former two. The size of this overlap shows the impact of email address spoofing.

Most emails originate outside the domain. In our log most outside users are spam senders and account for the majority of the emails. Because it is very easy for a spammer to forge an address spam senders use many addresses simultaneously and/or frequently switch between them. This strategy is visible in our database as non-spam internal

[3] Only for the accepted emails.

users send many more emails per user than spam internal users. We expect that this is a general feature of spam versus non-spam traffic.

The number of spam senders that are internal is very small. Among these, the fraction of the internal spam senders that send exclusively spam is 81%. These addresses correspond presumably to internal emails that have been forged and do not actually exist[4]. The remaining addresses send both spam and non-spam and are probably genuine users whose addresses have been spoofed.

Section 5.4.1 presents the structural analysis of both graphs and its relation with spammers and non-spam sender behaviors. The last Section 5.4.2 presents results for the dynamical characteristics of both graphs for each traffic.

5.4.1 Structural analysis of spam vs. non-spam email graphs

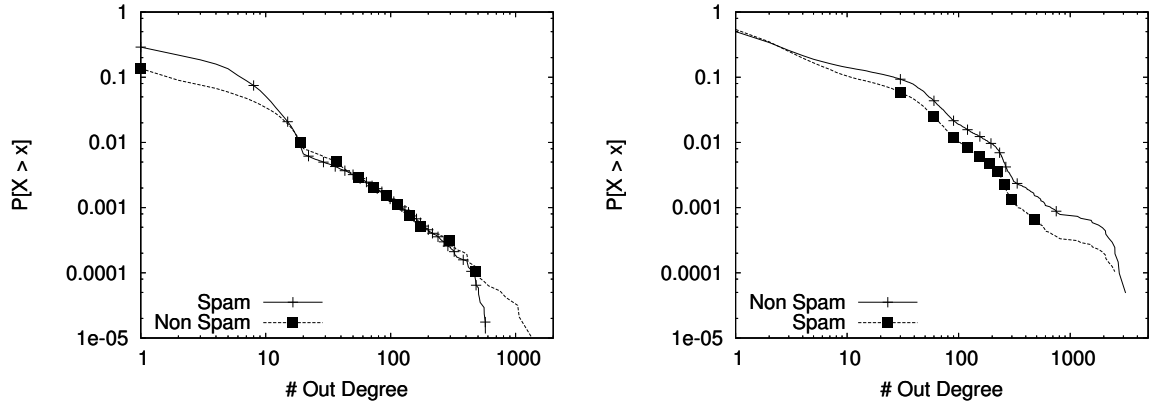
This Section presents the results of the analysis of how spammers and legitimate senders behaviors reflects in the structural properties of the user and domain graphs. Section 5.4.1.1 presents the distribution of the sender out-degrees for each traffic. The communication reciprocity of the spam and legitimate senders is presented in Section 5.4.1.2. In Section 5.4.1.3 we discuss the correlation between the asymmetry set of each sender and the number of spammers in it. Results for the communication cohesion of each group of users is presented in Section 5.4.1.4. Finally, in Section 5.4.1.5 is presented results for the probability of a user being part of a random walk in each graph.

5.4.1.1 Sender out-degree distributions

One of the most common structural measures analyzed in complex networks is the distribution of the number of the incoming and outgoing node connections, or degree [57, 13, 62]. Figure 5.1 shows the distribution of the out-degrees, the out-degree of a node in a graph is formaly defined in Chapter 3, of the different sender classes for the user and domain graphs.

The out degree distributions approximately follow a power law ($Const/x^{\alpha}$). By using a simple statistical linear regression we estimated the exponent α that best models the data. For the user graph we obtained α = 1.497 (with R^2 = 0.965.) for spam senders and α = 1.359 (R^2 = 0.981) for non-spam senders. We conclude that the spam senders' out degree distribution is slightly more skewed. We conjecture that this is because spammers have a limited knowledge of the set of users in each specific domain. Since in our analysis we only observe a fraction of the spammers' lists (the one composed by the messages sent to the domain studied) there are no spammers with recipients' lists as large as those found for non-spam senders.

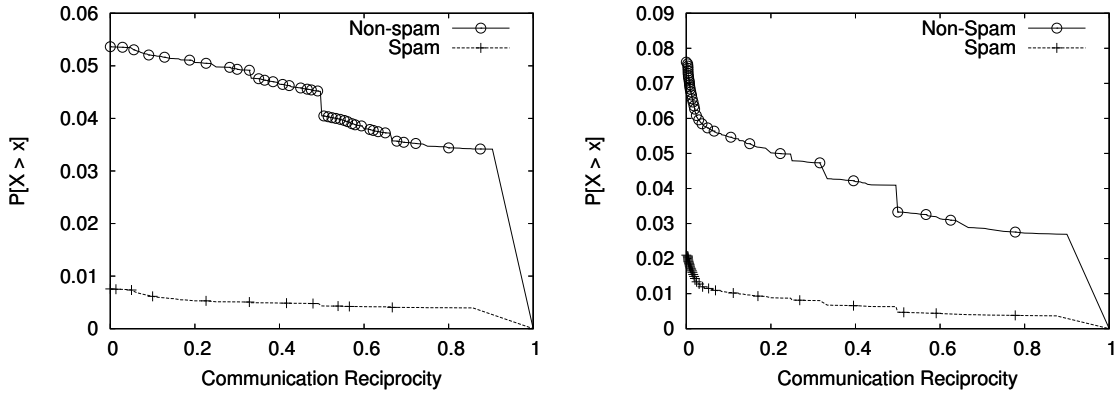
[4]This suggests that a simple effective way to filter out spam originating from internal domain addresses is to verify that they correspond to an existing user.



(a) User Graph

(b) Domain Graph

Figure 5.1: Distribution of the node degrees for sender classes in the aggregated graphs



(a) User Graph

(b) Domain Graph

Figure 5.2: Distribution of Communication Reciprocity

Degrees from 1 to near 20 are much more probable for spam senders than for non-spam senders, while very large degrees are more likely in non-spam. There are only small differences between the two sender classes in the body of the distribution, for degrees from about 20 to 400. The mean out-degrees, are 3.56 and 1.63 for non-spam and spam, respectively (see Table 5.2).

In the domain graph the out-degree distribution shows a much higher probability for nodes with low out-degree in spam traffic than in non-spam.

5.4.1.2 Communication reciprocity

In order to evaluate discrepancies between in and out sets of addresses for a given node we create a simple metric called Communication Reciprocity (CR) of x, a more formal definition of CR is given in Chapter 3, as:

$$CR(x) = \frac{|OS(x) \cap IS(x)|}{|OS(x)|}, \tag{5.1}$$

where $OS(x)$ is the set of nodes that receive a message from node x and $IS(x)$ is the set of addresses that send messages to x. With our choice of normalization this metric measures the probability of a node receiving a response from each one of his addresses.

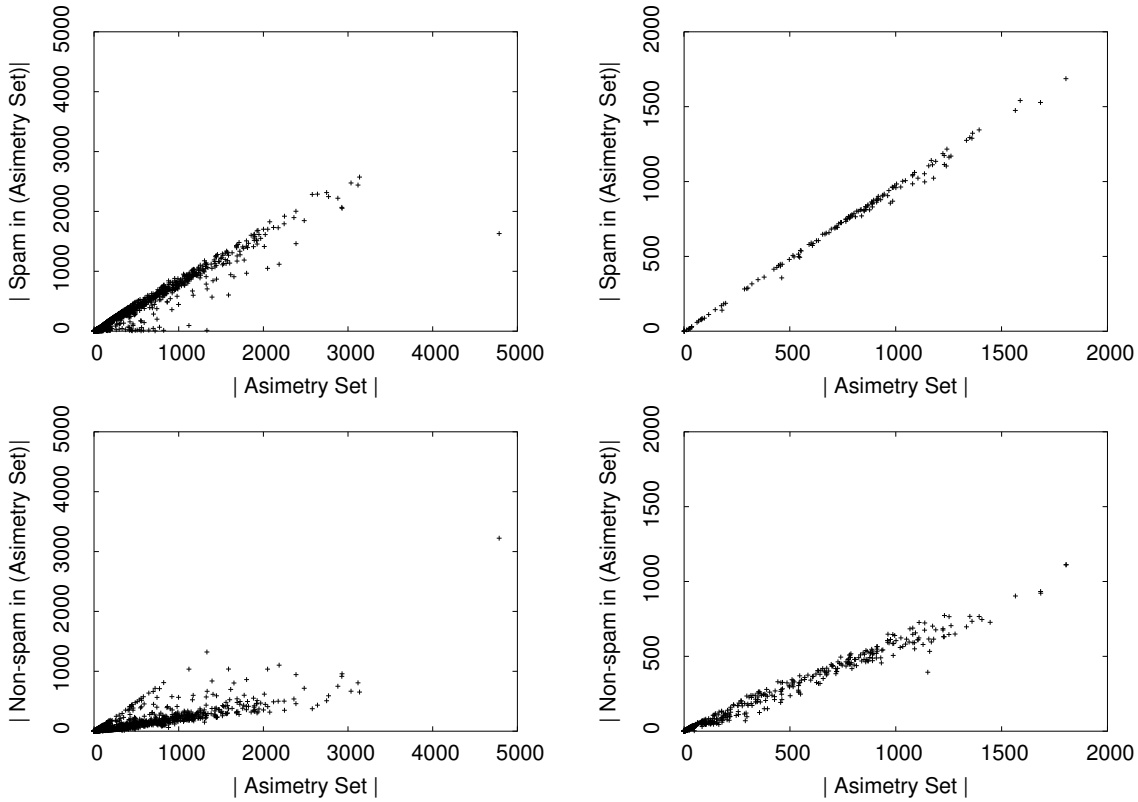
Figure 5.2 shows the distribution of the Communication Reciprocity. This metric is able to effectively differentiate users associated with spam from non-spam. The grouping of users in the domain graph makes this differentiation more difficult. However, even in the domain graph the difference is very clear.

The analysis of the communication reciprocity suggests that a strong signature of spam is its structural imbalance between the set of senders and receivers associated with a spam sender. However whenever there is an imbalance, how many of the unmatched addresses correspond to spam senders? A first approach to answering this question is presented in the next section.

5.4.1.3 Asymmetry set

Let the asymmetry set for a node be the difference of its in and out sets. Figure 5.3 shows the number of spammers addresses in the asymmetry set versus the size of the asymmetry set itself. The resulting relation is very well fit by a straight line at 45^o, showing a strong correlation between the two numbers. The statistical correlation is $C = 0.979$ for user graph and $C = 0.998$ for the domain graph. So, almost all senders in the asymmetry sets are spammers indifferently of the graph analyzed. The non-spam data is not very well modeled by a 45^o straight line. These correspond to the non-spam senders that were not answered (or to whom we could not see an answer in our log). The correlation is $C = 0.8723$ and $C = 0.9932$ for the user and domain graphs respectively. As expected from the result of the spam data the non-spam data has a higher correlation for the domain graph.

This result can be made sharper if we analyze the correlation between the number of spammers in the incoming set of a node and spammers in its asymmetry set. We find $C = 0.999$ and $C = 0.994$ for the user and domain graphs, respectively. There is a slightly worse correlation in the domain graph. We conjecture this is due to the external reliable domains used by spammers (e.g., through spoofing and forging techniques). These may not be counted in the asymmetry set since they are replied through their legitimate emails but are part of the incoming set as spammers.



(a) User Graph (b) Domain Graph

Figure 5.3: Number of spammers/non-spam senders in the asymmetry set vs. the number of nodes in it

These results show that spam messages are almost never replied to, except in cases of spoofed or forged domains or users' ids and rarely, we assume, intentionally.

Asymmetry sets can, in principle, be used as a component in a probabilistic spam detection mechanism. The arrival of an email from a sender that has already been contacted by an internal recipient is an indication that it has high probability of being a non-spam.

5.4.1.4 Clustering coefficient

Another common characteristic of social networks is a high average clustering coefficient (CC) [84]. As formaly defined in Chapter 3, the CC of a node n, denoted C_n, is defined as the probability of any two of its neighbors being neighbors themselves. This metric is associated to the number of triangles that contain a node n. For an undirected graph, the maximum number of triangles connecting the N_n neighbors of n is $N_n \times (N_n - 1)/2$. Thus, the CC measures the ratio between actual triangles and their maximal value. During clustering coefficient analysis we only consider the nodes with $N_n > 1$, since this is a necessary condition for the CC to be nonzero.

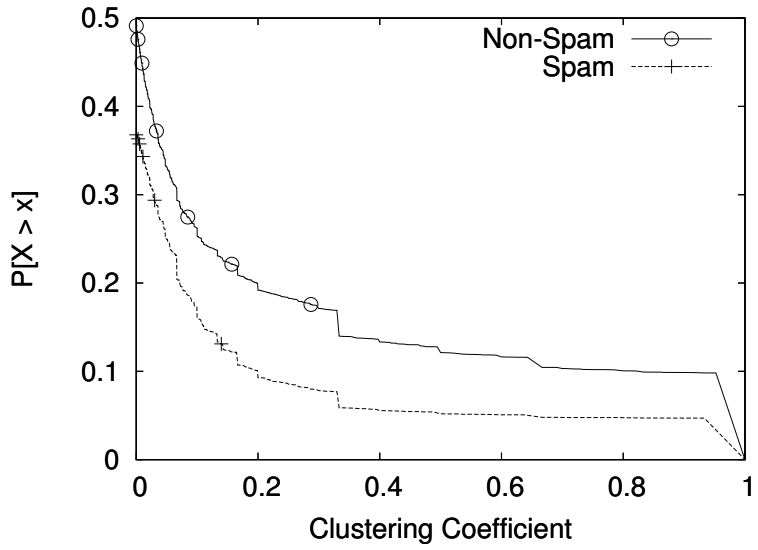
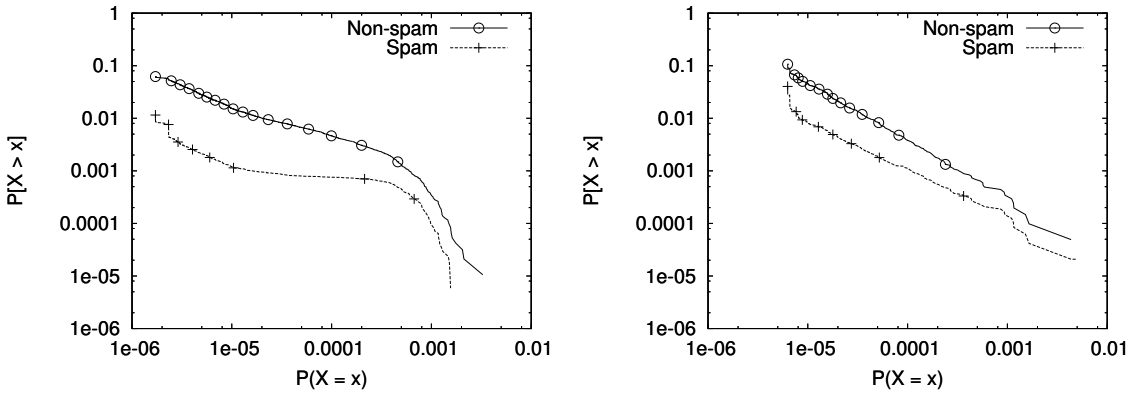


Figure 5.4: Distribution of the clustering coefficient for the non-spam and spam user classes in the aggregated user graph.

Figure 5.4 shows the distribution for the CC of nodes in the aggregated graphs. The clustering coefficient measures cohesion of communication, not only between two users but among *friends of friends*. This is a pervasive characteristic of social relations that tends to be absent in spam sender-receiver connections. As a result legitimate email users have on average higher CC than spam senders, 0.16 against 0.08, respectively.

Some recent studies [6] have analyzed graphical metrics of the connected components of email graphs. A connected component is a subset of the nodes of a graph, so that one node can be reached from any other node in the set following edges between them. A complementary measure to the CC and size of the connected component is the average path length between two nodes. The connected component and average path length properties are generally related to so-called small world networks, which display high CC (higher than a random graph with the same connectivity) and short path length, usually comparable to $\log N$, where N is the number of nodes in the graph.

In our experiments both the connected component and the average path length have not been able to convincingly differentiate spam from legitimate traffic. All of the graphs studied are small world networks. Also all of the graphs have giant connected components. Other studies have used the clustering coefficient of connected components to identify spam in networks constructed from the correspondence of a single user [6]. However for data from servers that aggregate the communication between different senders and recipients we find that these metrics do not suffice to perform a clear identification of spam.



(a) User Graph

(b) Domain Graph

Figure 5.5: Distribution of the probability of finding a node during a random walk.

5.4.1.5 Random walk

Another interesting structural characteristic of graphs is the probability of visiting a node during a random walk, as defined in [71] and in Chapter 3, through the graphs. At each step of the random walk we need to select the next node to be visited. This can be done in two ways. The next node can be randomly selected from the out set of the current node or we can perform a jump. For a jump, one of the nodes of the graph is selected randomly as the next node. Note that, this measure is related to node betweenness[5] since higher node betweenness tends to generate a higher probability of visitation. Nevertheless this probability is much easier to compute than node betweenness for large graphs. The probability $P(x)$ of finding a node x in a random walk is computed iteratively as follows:

$$P(x) = \frac{d}{N} + (1-d) * \sum_{z \in IS(x)} \frac{P(z)}{|OS(z)|}, \tag{5.2}$$

where d is the probability of performing a jump during a random walk, N is the number of nodes in the graph. The parameter d is a dumping factor that can be varied. A value usually used in the literature is 0.15 [71], that is also the value we use in our measurements.

The results are shown in Figure 5.5. The difference between spam and non-spam behavior is less noticeable in the domain graph than in the user graph. Spam nodes show generally lower probabilities of being visited, as might have been expected because of the asymmetry of their communication. Visiting probabilities for spam nodes in the user graph are localized to the initial and final parts of the distribution and are less pronounced in the middle range.

The node visitation probability distributions can be modeled by a power law. We estimate the corresponding

[5]The number of shortest paths between any two connected nodes that pass through a node.

exponent at $\alpha = 0.694,\ 1.097$ and 0.975 for the non-spam component of the user graph, and for the non-spam and spam components in the domain graph, respectively. The R^2 associated with the fits varies between 0.959 and 0.998. The R^2 for the spam curve of the user graph is 0.853, showing that it is not well modeled by a power law, as visual inspection suggests.

5.4.2 Dynamical analysis

Beyond the structural characteristics of the graphs of spam and non-spam email other metrics related to the dynamics of communication and graph evolution may help model spam traffic. In this section we will analyze some dynamical characteristics of the user and domain graphs defined.

In the Section 5.4.2.1 we present results for the growth of the graphs in number of nodes and edges. Section 5.4.2.2 analyzes temporal locality among peers on each traffic. Finally, entropy of the flow of incoming messages and bytes from each sender is studied in Section 5.4.2.3.

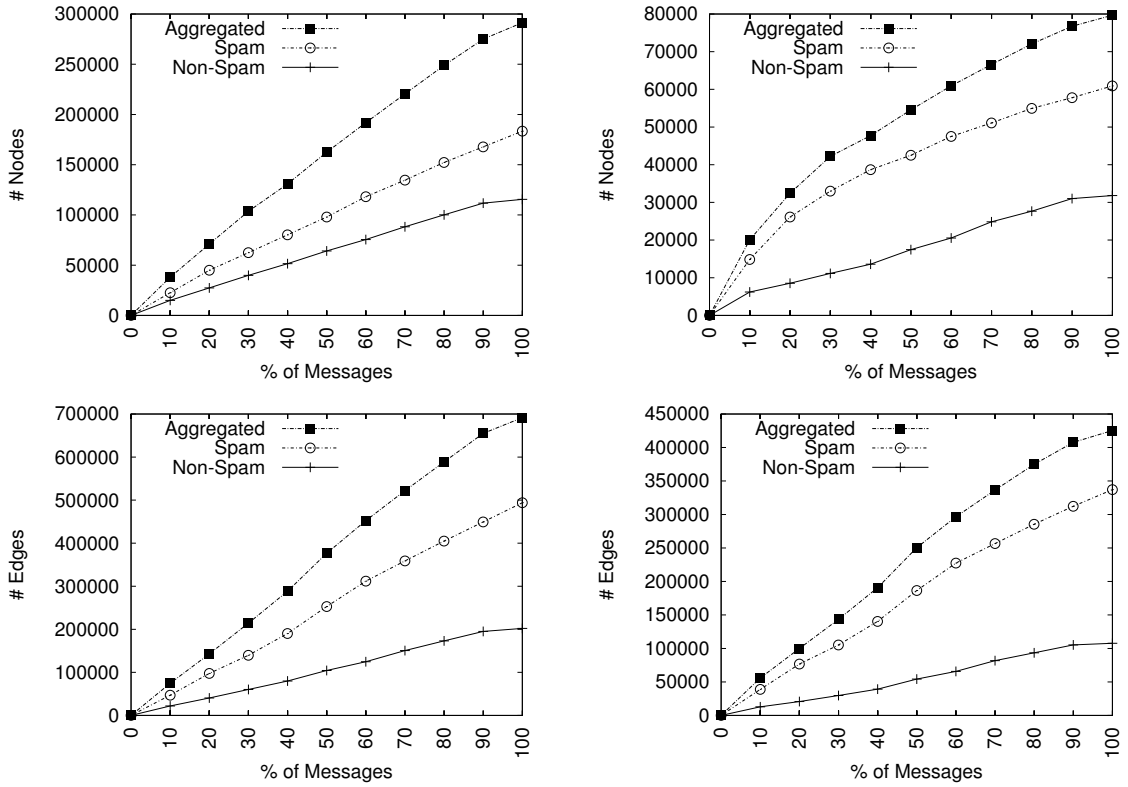
5.4.2.1 Growing of the graphs

A large amount of effort has been devoted recently to creating realistic growth models for complex networks. One of the key characteristics of such models is the number of nodes and edges evolution, as well as the probabilistic connection rules for the new nodes to those already in the graph. Figure 5.6 shows the evolution of the graph in terms of number of nodes and edges. We plot these quantities against percentage of messages evaluated for each graph, to avoid the rate of message arrival influence, which varies with time depending on the type of the traffic being considered (e.g., the bell shaped behavior for the non-spam traffic against the almost constant rate for spam traffic [10, 1, 4]).

The growth of the aggregated graph (a composition of the spam and the non-spam graphs) results from the growth in both the spam and non-spam components. The spam subgraph is a much more rapidly growing structure. Over the time of the log we find no saturation effect in these numbers. Instead the number of addresses and edges grows almost linearly with the number of emails. An eventual saturation in the non-spam component might be expected for longer times.

5.4.2.2 Temporal locality among pairs sender-recipient

Another important dynamical graph characteristic is how the weights of edges evolve. An interesting metric that can be used to measure this is the *stack distance* [81] of connected pairs in terms of the emails they exchange over time. The stack distance measures the number of distinct references between two consecutive instances of the same object in a stream. We take the total email log as the stream and each pair (sender,receiver) as the object,



(a) User graph (b) Domain graph

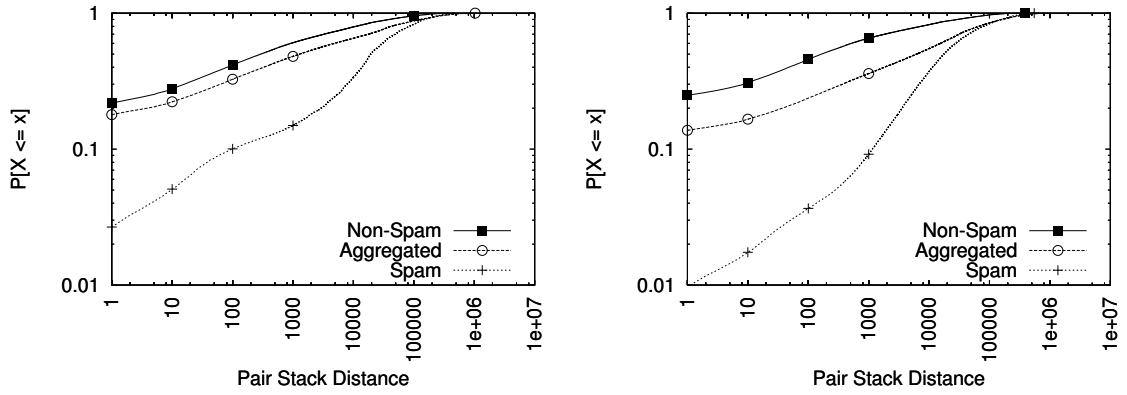
Figure 5.6: Graph evolution by percentage of messages.

disregarding the order. Figure 5.7 shows the pairs' stack distance distributions. We see that temporal locality is much stronger in non-spam traffic. This means in practice that legitimate users exchange emails over small concentrations of time.

5.4.2.3 Entropy of the flow of messages/bytes per sender

We were also interested in studying how the nodes communicate with their peers in terms of number of messages. Because of the impersonal nature of spam we expect that spam senders communicate in a more predictable way with their recipients. Not only will legitimate senders show more variation in the number of messages they send to each person in their out sets, but they will also show variability of the messages themselves in terms of their sizes. In order to quantify these effects we evaluated the normalized entropy of the in and out flows for each node, defined as:

$$H(x) = \frac{\sum_{y \in OS(x)} -p(y) * log(p(y))}{log(|S(x)|)}, \tag{5.3}$$



(a) User Graph

(b) Domain Graph

Figure 5.7: Distribution of stack distances for the pairs (sender, recipient) in the distinct traffic.

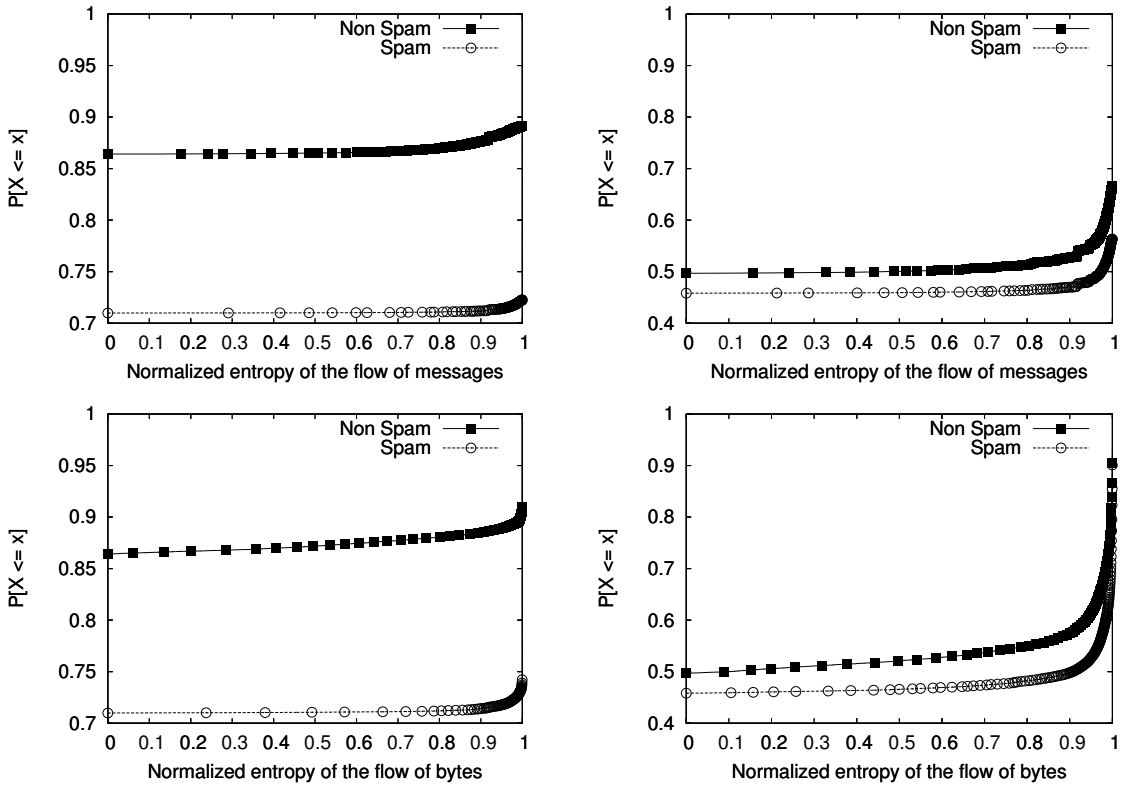
where $p(y)$ is the probability of y receiving a message from x and and $|S(x)|$ is the number of distinct elements in the set being considered.

Figure 5.8 shows the normalized entropy for the out flow of the nodes in the different sender classes for the aggregated graphs. As expected, spammers communicate with their recipients with much less variability (higher entropy). A similar analysis was conducted considering the bytes that each node sends with similar results. Due to the great aggregation of nodes in the domain graphs, the entropy of the bytes/messages flow is not capable of distinguishing the two traffics as it is in the user graphs.

5.5 Conclusions

In this chapter, we have shown that legitimate and spam email graphs differ in two fundamental classes of characteristics: structural, which capture the skeleton architecture of the graphs, and dynamical, concerning node communication and graph evolution.

We have shown that the spam and non-spam subgraphs are structurally characterized by different distributions of their node clustering coefficient. Legitimate email users display on average higher clustering coefficients than spam senders. Node visitation probability is a measure of the centrality of a node relative to other nodes in the graph. Legitimate email nodes have higher visitation probability than spam nodes. We also defined a new metric called communication reciprocity. It measures the probability that a node receives a response from any of its addressees. There is a strong difference in the probability distributions of the communication reciprocity in the legitimate and spam graphs; legitimate nodes have a much higher probability of being responded to. Another



(a) User graph (b) Domain graph

Figure 5.8: Distribution of entropy of the number of messages/bytes in the flow of emails for the aggregated graph.

metric introduced in this chapter is the email asymmetry set, which represents the difference between the sets of in and out edges of a node. We showed that there is a strong correlation between the size of asymmetry sets and the number of spammers in the set. Dynamically the spam graph grows much faster than the legitimate email graph. The legitimate email graph grows more slowly both in the number of nodes and edges, manifesting the higher stability of relations in a social group. Two other dynamical metrics, entropy and stack distance, are used to reveal the temporal characteristics of communication among nodes. Spam nodes display a much higher entropy than legitimate email users, and a much longer stack distance.

We have shown that differences in both classes of graph characteristics can be explained by the same hypothesis, namely that legitimate email graphs reflect real social networks, while spam graphs are technological networks, devoid of a sense of community. Although no single metric can unequivocally differentiate legitimate emails from spam, the combination of several graphical measures paint a clear picture of the processes whereby legitimate and spam email are created. For this reason they can be used to augment the effectiveness of mechanisms to detect illegitimate emails.

Chapter 6

Using Structural Similarity for Improving Spam Detection

6.1 Introduction

None of the existing spam filtering mechanisms is infallible [3, 8]. Their major problems are false positives and wrong mail classification. In addition, filters must be continuously updated to capture the multitude of mechanism constantly introduced by spammers to avoid filtering actions.

In this Chapter, we propose and evaluate two algorithms for spam detection that use structural relationships between senders and recipients as the basis for the detection of spam messages. The algorithms work in conjunction with another spam classifier (hereafter called auxiliary algorithm), necessary to produce spam or legitimate mail tags on past senders and receivers, which in turn are used to infer new ones, through structural similarity. The key idea is that the lists of distinct recipients that spammers and legitimate users send messages to, as well as the lists of distinct senders from which users receive messages [1], can be used as the identifiers of senders and recipients in email traffic [3, 9, 10]. We show the application of our structural based algorithms over the results of the auxiliary classifier leads to the correction of a number of misclassification. A preliminary version of the results presented in this chapter where published in [11].

The algorithms presented in this Chapter aim at improving the effectiveness of spam filtering mechanisms, by reducing false positives and by providing information to tune their collection of rules. Differently from most of the alternatives for spam filtering, we focus our analysis on the characteristics that we conjecture are the most difficult for the spammers to change, i.e., their structural relationships with email users, as a way to improve the

[1] The contact lists as defined in Chapter 4

classification provided by other filters. Other recent studies have focused on spam combat techniques based on characteristics of graph models of email traffic [6, 7, 11].

This Chapter is organized as follows: Section 6.2 presents the methodology used to handle email data. The structural algorithm proposed is described in Section 6.3. We describe the workloads and the classification results of our algorithms for each workload in Section 6.4. In Section 6.4.3 we discuss the use of weighted representations for email users. Finally, conclusions are presented Section 6.5.

6.2 Modeling similarity among email senders and recipients

Our proposed spam detection algorithms exploit the structural similarities that exist in groups of senders and recipients of email. This section introduces a unifying modeling framework of individual email users and proposes a metric to capture the similarity between them. It then shows how to obtain clusters of users who share great similarity.

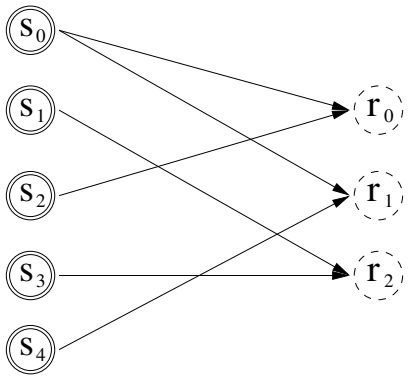
In Chapter 3 we present all the definitions related to the modeling presented in this Section.

Our basic assumption is that, in both legitimate and spam traffics, users have a defined list of peers they often have contact with (i.e., they send/receive an email to/from, see Section 6.4). In legitimate email traffic, contact lists are the consequence of social relationships. On the other hand, the lists created by spammers to distribute their solicitations are guided by business opportunities and, generally, do not reflect any form of social interaction. A user's contact list certainly may change over time. However, we expect it to be much less variable than other characteristics commonly used for spam detection, such as the presence of certain keywords in the email content or its size and encoding. In other words, we expect contact lists to be an effective basis for detecting spam.

We start by representing an email user as a vector in a multi-dimensional vector space created out of all possible contacts. We represent email senders and recipients separately. We then use vector operations (the inner product in contact space) to express the similarity among multiple senders/recipients, and use this metric for clustering them. Our approach can be used with several different email user identification (e.g., email address, domain name, SMTP relay, etc); therefore, the term email user is used throughout this work to denote any means of identification of an email sender/recipient that one may want to use.

Let N_r be the number of distinct recipients. We represent a sender s_i as a N_r dimensional vector $\vec{s}_i$, defined in the vector space of email recipients being considered. The n-th dimension (representing recipient r_n) of $\vec{s}_i$ is defined as:

$$\vec{s}_i[n] = \begin{cases} 1, & \text{if } s_i \rightarrow r_n \\ 0, & \text{otherwise} \end{cases}, \tag{6.1}$$



(a) Messages

$\vec{s}_0 = (\ 1,\ 1,\ 0)$
$\vec{s}_1 = (\ 0,\ 0,\ 1)$
...
$\vec{s}_4 = (\ 0,\ 1,\ 0)$

(b) Senders

$\vec{r}_0 = (\ 1,\ 0,\ 1,\ 0,\ 0)$
$\vec{r}_1 = (\ 1,\ 0,\ 0,\ 0,\ 1)$
$\vec{r}_2 = (\ 0,\ 1,\ 0,\ 1,\ 0)$

(c) Recipients

Figure 6.1: Message Exchange Example: (a) Shows a Graph-representation of Email Data, (b) Shows the Vector Representation of the Senders and (c) Shows the Vector Representation of the Recipients

where $s_i \rightarrow r_n$ indicates that sender s_i has sent at least one email to recipient r_n.

Similarly, we define $\vec{r}_i$ as a N_s dimensional vector representing recipient r_i, where N_s is the number of distinct senders being considered. The n-th dimension of this vector is analogously set to 1 if and only if recipient r_i has received at least one email from s_n.

Figure 6.1-a shows a graph that represents the communication between senders and recipients. Nodes represent email users and the edges connecting them imply that a user has sent at least one message to another email user. The bipartite graph representation is adequate since we separate the two facades of each user, i.e. their sending and receiving behaviors, creating a separate representation for each. Figures 6.1-b and 6.1-c show the vector representations for senders and recipients, respectively.

We next define the similarity between two senders s_i and s_j as the cosine of the angle (the normalized inner product) between their vector representation ($\vec{s}_i$ and $\vec{s}_j$). The cosine is a well known metric that has been successfully employed in several application areas, including document similarity in information retrieval systems [85, 86] and intrusion detection [87]. This similarity metric is computed as:

$$sim(s_i, s_j) = \frac{\vec{s}_i \bullet \vec{s}_j}{|\vec{s}_i||\vec{s}_j|} = \cos(\vec{s}_i, \vec{s}_j), \tag{6.2}$$

where $\vec{s}_i \bullet \vec{s}_j$ represents the inner product of the vectors and $|\vec{s}_i|$ represents the norm of $\vec{s}_i$. Note that, since the vector representation of senders are found in the first quadrant of the respective conceptual spaces, this metric varies from 0, when senders do not share any recipient in their contact lists, to 1, when senders have identical

contact lists and thus have the same representation. The similarity between two recipients is analogously defined.

We note that our similarity metric has different interpretations in legitimate and spam traffics. In legitimate email traffic, it tends to represent social interaction with the same group of users, whereas in the spam traffic, a great similarity probably represents the use of different identifiers by the same spammer or the sharing of distribution lists by distinct spammers.

Finally, we can use our vector modeling approach to represent a cluster of senders or recipients. A sender cluster sc_i, represented by vector $\vec{sc}_i$, is computed as the sum of its elements, that is:

$$\vec{sc}_i = \sum_{s \in sc_i} \vec{s}. \tag{6.3}$$

The similarity between sender s and an existing cluster sc_i can then be directly assessed by extending Equation 6.2 as follows:

$$sim(sc_i, s) = \begin{cases} \cos(\vec{sc}_i - \vec{s}, \vec{s}), & \text{if } s \in sc_i \\ \cos(\vec{sc}_i, \vec{s}), & \text{otherwise} \end{cases} \tag{6.4}$$

We note that a sender $\vec{s}$ vectorial representation and the sender cluster to which it belongs may change over time as new emails are considered. In order to accurately estimate the similarity between a sender $\vec{s}$ and a sender cluster $\vec{sc}_i$ to which $\vec{s}$ currently belongs to, we first remove $\vec{s}$ from $\vec{sc}_i$, and then take the cosine between the two vectors ($\vec{sc}_i - \vec{s}$ and $\vec{s}$). This is performed so that the previous classification of a user does not influence its actual classification. Recipient clusters and the similarity between a recipient and a given recipient cluster are defined analogously.

6.3 Structural similarity algorithms

This section introduces our email classification approach, which exploits the similarities between email senders and recipients to group them into clusters and uses historical information to improve spam detection accuracy. We present two algorithms, each one designed to work together with any existing spam detection or filtering technique. Our goal is to provide a significant reduction of false positives (i.e., legitimate emails wrongly classified as spam), which can be as high as 15% in current filters [36].

The architecture proposed in this chapter is shown in Figure 6.2. A message arrives at the spam detection system and is directed to a structural similarity based algorithm. This algorithm first sends the message to the

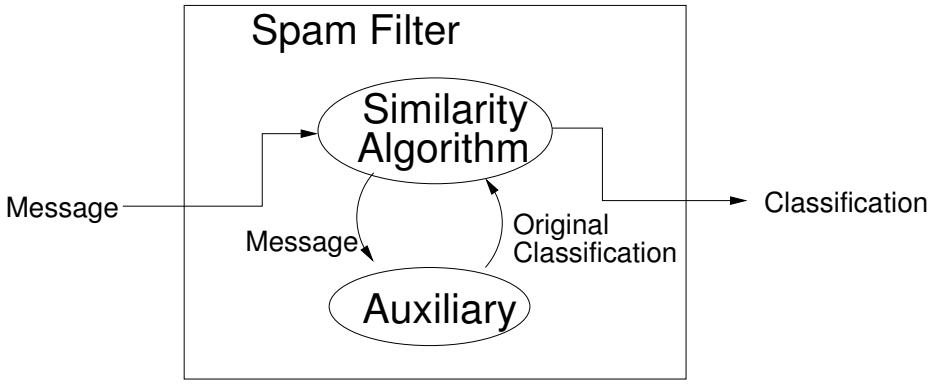


Figure 6.2: The Spam Fight Architecture

auxiliary algorithm in order to retrieve a first classification attempt of that message. Based on this classification, on the cluster formed by senders and recipients, and, on historical information, our algorithms generate a new classification, which can coincide or not with the original classification provided by the auxiliary algorithm. The idea is to use the classification performed by the auxiliary method to build an incremental historical knowledge base that gets more representative as more messages are processed.

The two algorithms presented below are similar in their inner workings but differ in the way they use historical information derived from the clusters to provide the final classification of the messages. Whereas the cluster-based algorithm derives its classification from historical information of the clusters formed by the senders/recipients of each message, the communication-based algorithm uses historical information of the interactions between clusters in order to classify messages. The cluster-based and the communication-based algorithms are presented in detail in Section 6.3.1 and Section 6.3.2 respectively.

6.3.1 Cluster-based algorithm

This algorithm, in Algorithm 1, maintains sets of sender and recipient clusters, created by the structural similarity, as defined in Equation (6.4). A sender/recipient of an incoming email is added to the sender/recipient cluster that is most similar to it provided that their similarity exceeds a given threshold τ. Thus, τ defines the minimum similarity a sender/recipient must have with a cluster to be assigned to it. Varying τ allows us to create more tightly or loosely knit clusters. If no cluster can be found, a new single-user cluster is created. In this case, this sender/recipient is used as seed for populating the new cluster.

The sets of sender and recipient clusters are updated at each new email arrival based on the email sender and on the list of recipients. Recall that to determine the cluster of a previous classified user we first remove the user from its current cluster and then assess its similarity to each existing cluster. Thus, single-user clusters tend to be

(1) **foreach** arriving message m
(2) $mClass$ =classification of m by auxiliary detection method;
(3) sc =find cluster for $m.sender$;
(4) Update spam probability for sc using $mClass$;
(5) $P_s(m)$ =spam probability for sc;
(6) $P_r(m) = 0$;
(7) **foreach** recipient $r \in m.recipients$
(8) rc =find cluster for r;
(9) Update spam probability for rc using $mClass$;
(10) $P_r(m) = P_r(m)$+spam probability for rc;
(11) $P_r(m) = P_r(m)/size(m.recipients)$;
(12) $SR(m)$ = compute spam rank based on $P_s(m)$ and $P_r(m)$;
(13) **if** $SR(m) > \omega$
(14) classify m as spam;
(15) **else if** $SR(m) < 1 - \omega$
(16) classify m as legitimate;
(17) **else**
(18) classify m as $mClass$;

Algorithm 1: Cluster-based Algorithm for Email Classification.

reduced as more emails are processed, except for users that appear only very sporadically.

A probability of sending/receiving spam messages is assigned to each sender/recipient cluster. We refer to this measure simply as the *cluster spam probability*. We calculate the spam probability of a sender/recipient cluster as the average spam probability of its elements, which, in turn, is estimated using the frequency of spams sent/received by each of them in the past. Therefore, the cluster-based algorithm uses the result of the email classification performed by the auxiliary algorithm on each arriving email m ($mClass$ in Algorithm 1) to continuously update cluster spam probabilities.

Let us define the probability of an email message m being sent by a spammer, $P_s(m)$, as the spam probability of its sender's cluster. Similarly, let the probability of an email m being addressed to users that receive spam, $P_r(m)$, as the average spam probability of all of its recipients' clusters. The cluster-based algorithm uses $P_s(m)$ and $P_r(m)$ to compute a number that expresses the chance of email m being spam. We call this number the *spam rank* of email m, denoted by $SR(m)$. The idea is that emails with large values of $P_s(m)$ and $P_r(m)$ should have high spam ranks and thus should be classified as spam messages. Similarly, emails with small values of $P_s(m)$ and $P_r(m)$ should receive low spam rank and be classified as legitimate email (see Section 6.4.2 for experimental support of this hypothesis).

Figure 6.3 shows a graphical representation of the computation of the spam rank for a message. We first normalize the probabilities $P_s(m)$ and $P_r(m)$ by a factor of $\sqrt{2}$, so that the diagonal of the square region defined in the bi-dimensional space be sized as 1 (see Figure 6.3-left and equation 6.5). Each email m can be represented as a

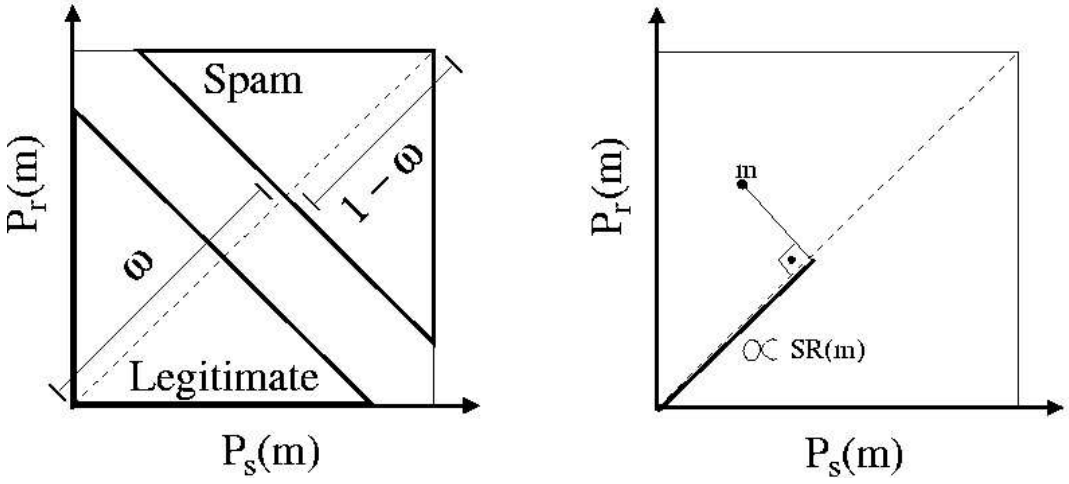


Figure 6.3: Spam Rank Computation and Email Classification for the Cluster-based Algorithm.

point in this square. The spam rank of m, $SR(m)$, is then defined as the length of the segment starting at the origin (0,0) and ending at the projection of m on the diagonal of the square (see Figure 6.3-right). With these definitions the spam rank varies between 0 and 1.

$$SR(m) = \frac{|\langle P_r(m), P_s(m)\rangle| \cos\theta}{\sqrt{2}}, \tag{6.5}$$

where θ is angle between the vector $\langle P_r(m), P_s(m)\rangle$ and the vector $\langle 1, 1\rangle$, $|\langle P_r(m), P_s(m)\rangle|$ means the norm of the vector and $\sqrt{2}$ is a normalization factor in order to obtain an $SR(m) \leq 1$.

The spam rank $SR(m)$ is then used to classify m as follows: if it is greater than a given threshold ω, the email is classified as spam; if it is smaller than $1 - \omega$, it is classified as legitimate email. Otherwise, we can not precisely classify the message, and we rely on the initial classification provided by the auxiliary detection algorithm. The parameter ω can be tuned to determine the precision of our classification. Graphically, emails are classified according to the marked regions shown in Figure 6.3-left. The two identical triangles represent the regions where our algorithm is able to classify emails as either spam (upper right) or legitimate email (lower left).

6.3.2 Communication-based algorithm

The algorithm, in Algorithm 2, works similarly to the cluster-based algorithm. The threshold τ is used to separate sender and recipients of emails in clusters that are then used to classify the messages.

The difference between the approaches is the way we compute the spam rank of a message. While the cluster-based algorithm uses aggregated information about the clusters, the communication-based approach uses information of the exchange of messages between clusters. This algorithm tries achieve more accuracy by exploiting the

(1) **foreach** arriving message m
(2) $mClass$ =classification of m by auxiliary detection method;
(3) sc =find cluster for $m.sender$;
(4) Update spam probability for sc using $mClass$;
(5) $SR(m) = 0$;
(6) **foreach** recipient $r \in m.recipients$
(7) rc =find cluster for r;
(8) Update spam probability for rc using $mClass$;
(9) $SR(m) = SR(m)+$ probability of sc and rc exchange a spam;
(10) $SR(m) = SR(m)/size(m.recipients)$
(11) **if** $SR(m) > \omega$
(12) classify m as spam;
(13) **else if** $SR(m) < 1 - \omega$
(14) classify m as legitimate;
(15) **else**
(16) classify m as $mClass$;

Algorithm 2: Communication-based Algorithm for Email Classification.

facades of each cluster's relationships with others. As can be seen from Algorithm 2, the spam rank of a message in this case is computed as the average probability of the sender cluster exchanging a spam message with the cluster of their recipients. The parameter ω is still used in order to determine the final classification of the message.

6.4 Experimental results

In this section we describe the experimental results. Section 6.4.1 presents results obtained by our algorithms using actual data collected from an SMTP server. Results derived from synthetic workloads generated to evaluate the general properties of our approaches are presented in Section 6.4.2. Finally, Section 6.4.3 discusses the use of a weighted representation for email users as opposed to the binary representation showed in Section 6.2.

6.4.1 Actual workload

Our email workload consists of anonymized SMTP logs of incoming emails to a large university, with around 22000 students, in Brazil. The logs are collected at the university central Internet email server of the. This server handles all emails coming from the outside addressed to the vast majority of students, faculty and staff who have email addresses under the major university domain name.

The central email server runs the Exim email software [75], the AMaViS virus scanner [76] and the Trendmicro Vscan anti-virus tool [77]. It also runs a set of pre-acceptance spam filters, including local black listing and local heuristics for detecting suspicious senders. These filters block on average 50% of all daily SMTP connection

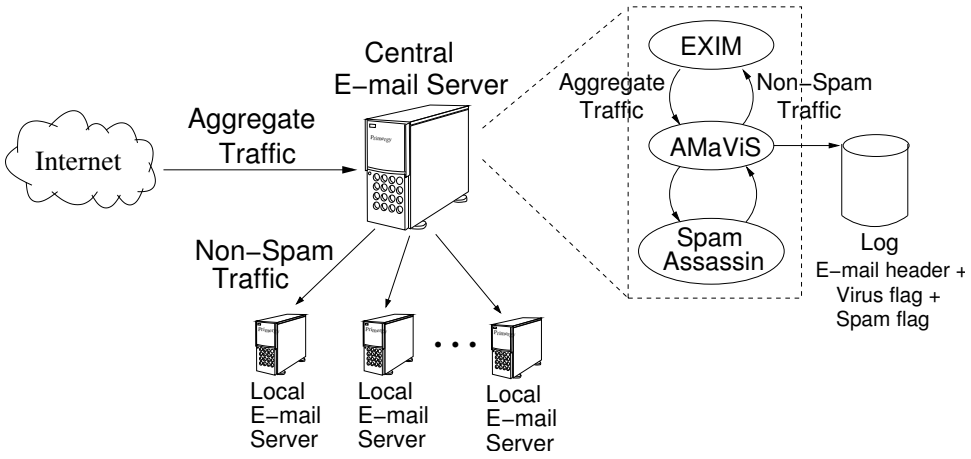


Figure 6.4: Central Email Server Architecture.

arrivals. The server also runs SpamAssassin [51], see Section 4.2.1 in Chapter 4 for more information about it, on all email messages that are accepted.

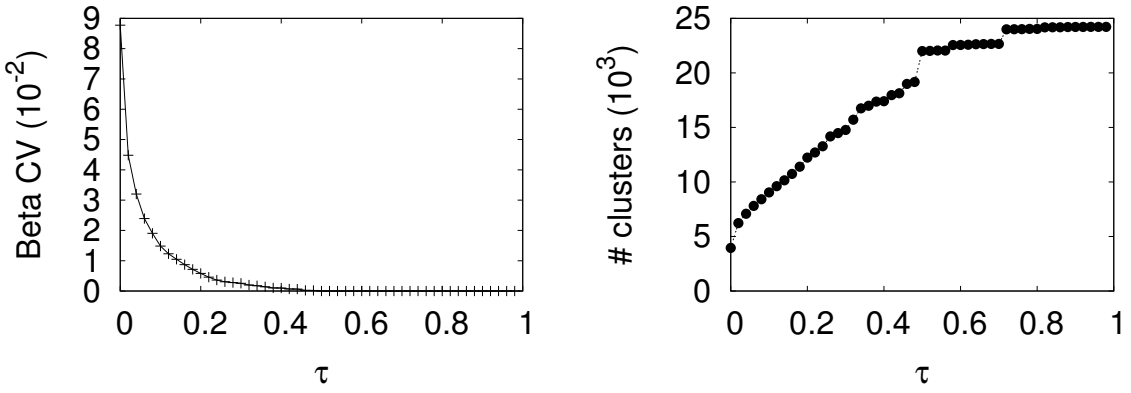
We analyze an eight-day log, between 01/19/2004 to 01/26/2004, collected by the AMaViS software at the central email server, during the university's academic year. For each email that is accepted by the server the logs register the following: the arrival time, the message size, the sender email address, a list of recipient email addresses, flags indicating whether the email was classified as spam or detected to be infected with a virus and its content in these cases. We also have the full body of the messages that were classified as spam by SpamAssassin. Figure 6.4 shows the overall data collection architecture at the central email server.

Emails that are flagged with virus or addressed to recipients in a domain name outside the university, for which the central email server is a published relay, are *not* included in our analysis. These emails correspond to only 0.8% of all logged data. Table 6.1 summarizes our workload.

Measure	**N**on-Spam	**S**pam	**A**ggregate
# of emails	191,417	173,584	365,001
Size of emails	11.3 GB	1.2 GB	12.5 GB
# of distinct senders	12,338	19,567	27,734
# of distinct recipients	22,762	27,926	38,875

Table 6.1: Summary of the Workload.

Note that the central server does not perform any test on the existence of the recipient addresses of the accepted emails. Such tests are performed by the local servers. Thus, some of the recipient email addresses in our logs may not actually exist. These recipient addresses could be result of honest mistakes or the consequence of dictionary attacks. A dictionary attack is a spamming technique where spammers submit thousands or millions of email messages with random *potential* addresses, such as joe@domain.com, john@domain.com, etc.



(a) Beta CV

(b) Clusters

Figure 6.5: Number of Email User Clusters and Beta CV vs. τ.

By visually inspecting the fraction of the list of senders *user names* [2] of spam messages in our workload, we found that a large number of them corresponded to a seemingly random sequence of characters, suggesting that spammers tend to change user names as an evasion technique. Therefore, for the experiments presented below we identified the sender of a message by his/her domain while recipients were identified by their full addresses, including both domain and user name.

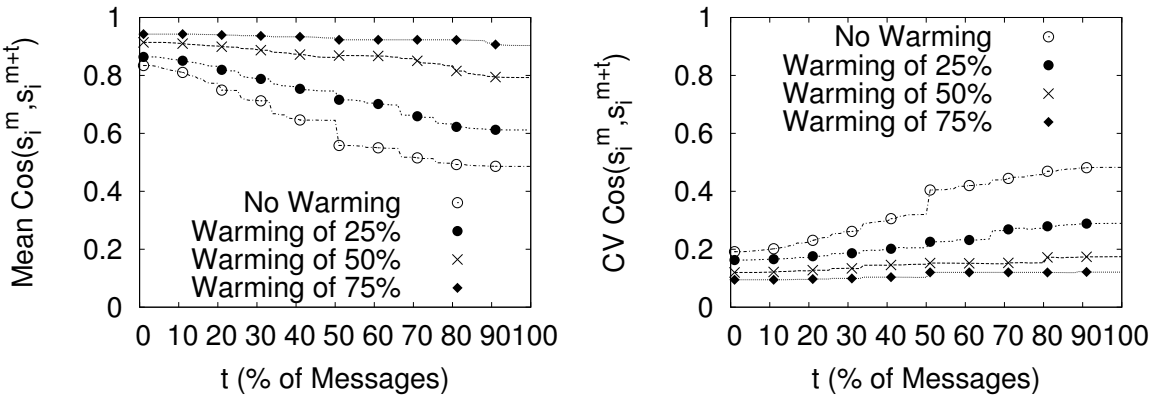
6.4.1.1 Classification results

The number and quality of the clusters generated through our similarity measure are the direct result of the chosen value for the threshold τ (see Section 6.3). In order to determine the best parameter value the simulation was executed several times for varying τ. The most appropriate value for τ is the same for both algorithms, due to the fact that they use the same clustering scheme.

$$\beta_{CV} = \frac{intraClusters_{CV}}{interClusters_{CV}} \tag{6.6}$$

Figure 6.5 shows how the beta CV (Coefficient of Variation) for the clusters and the number of clusters vs. τ. Beta CV denotes the ratio of intra CV over the inter CV for the clusters, see equation 6.6, whereas the intra CV measures the coefficient of variation for the similarities intra-cluster, the inter CV measures the coefficient of variation for similarities between different clusters. Thus, beta CV is a measure of the quality of the clusters generated. The more stable the beta CV the better quality in terms of the grouping is obtained [74]. There is one

[2] The part before @ in email addresses.



(a) Mean Cosine

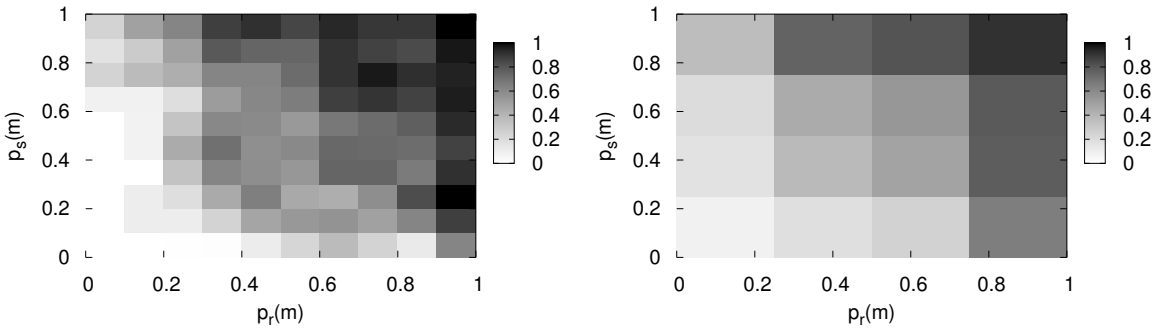
(b) CV of Cosine

Figure 6.6: Spam Senders Identification Stabilization

clear point of stability in the curve of Figure 6.5-b at $\tau = 0.5$. Moreover, the value of beta CV also stabilize at $\tau = 0.5$ (Figure 6.5-a). This is the value we adopt for the remaining of the chapter.

Our algorithms rely on three hypothesis. First, contact lists of email users provide an effective mean for identifying them. Second, for the cluster-based algorithm, messages can be more accurately classified as spam or not based on the probabilities of sending/receiving spams of the cluster that their sender/recipients belong to. Finally, for the communication-based algorithm, we hypothesize that historical information about communication between a sender and a recipient cluster can be used to detect spam.

In order to show that contact lists provide an effective means of user identification, we analyze how sender /recipient vectorial representations change over time, as new messages arrive in the system. In this analysis, we consider a warm-up period containing a certain fraction of the messages, during which sender/recipients are updated. We define $\vec{s}_{i,j}$ as the vectorial representation of sender i at a point in time when it had sent $j\%$ of its messages (j is larger than the warm-up period). We use the similarity between the representations of sender i in two points in time, spaced by a certain factor t of messages sent, as a measure of how stable the sender identification is over that period. We then analyze how this similarity, given by $\cos(\vec{s}_{i,j}, \vec{s}_{i,j+t})$ evolves as the step t increases. Figure 6.6-a shows average similarity measures for the senders in the spam traffic, varying the step from 1% to 100%, for different warm-up periods. Figure 6.6-b shows, for each average, the corresponding coefficient of variation (CV). Note that, as expected, when there is no warm-up period ("No Warming"), fluctuations in the early messages dominate, and as a consequence stabilization is not reached by the end of our workload. In other words, a large set of messages would be required for stabilization to be reached. As the warm-up period increases, stabilization is reached with a smaller number of messages. In particular, when we consider only the last 25% of the messages



(a) Bin size = 0.10

(b) Bin size = 0.25

Figure 6.7: Number of Spam Messages by Varying Message Spam Probabilities for Different Bin Sizes.

for each sender ("Warming of 75%"), sender representation remains, on average, very stable with CV approaching zero. Similar patterns were observed for legitimate email senders as well as legitimate and spam recipients.

Next, we investigate the second hypothesis. Figure 6.7 shows the fraction of spam messages in our workload for different values of $P_s(m)$ and $P_r(m)$ grouped based on a discretization of the full space represented in the plot. This space is subdivided into smaller squares of the same size called bins, the darker the gray scale the greater the number of spams in each bin. Clearly, spam and legitimate messages are located on the top-right and bottom-left regions of the spectrum as we have hypothesized in Section 6.3. There is, however, an intermediate region in the middle where we cannot satisfactorily determine the classification. This is why it becomes necessary to vary ω. One should adjust ω based on the level of confidence it has on the auxiliary algorithm.

Figure 6.7 shows that messages addressed to recipients that have high $P_r(m)$ tend to be spam more frequently than messages with the same value of $P_s(m)$. Analogously, messages with low $P_s(m)$ have higher probability of being legitimate messages.

The third hypothesis is that we are able to classify messages based on the previous patterns of communication between their sender and recipients' clusters. Figure 6.8 shows the probability of a message being spam as a function of its spam rank computed using the communication-based algorithm (see Algorithm 2). The plot clearly shows that there is a strong correlation between the chance of a message being spam and the communication between its sender and recipient clusters.

Both of our algorithms make use of an auxiliary spam detection algorithm, such as SpamAssassin. Therefore, we need to evaluate how frequently we repeat the same classification of that algorithm. Figure 6.9 shows the the percentage of messages that received the same classification and the total number of classified messages in our

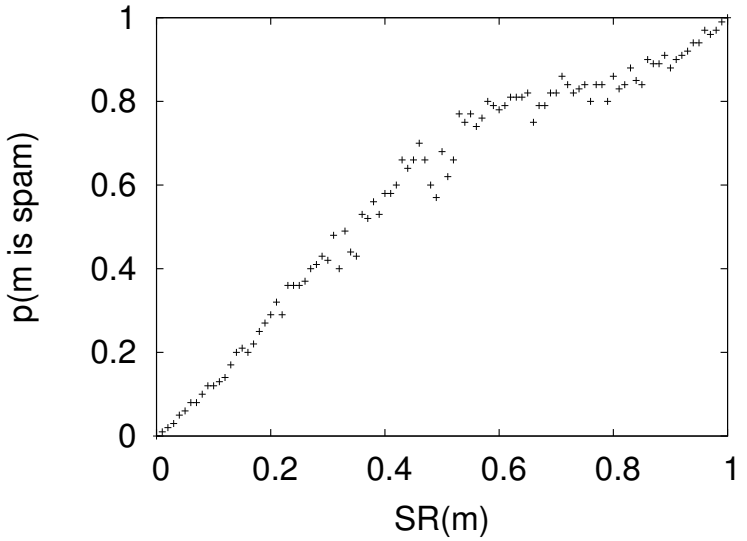


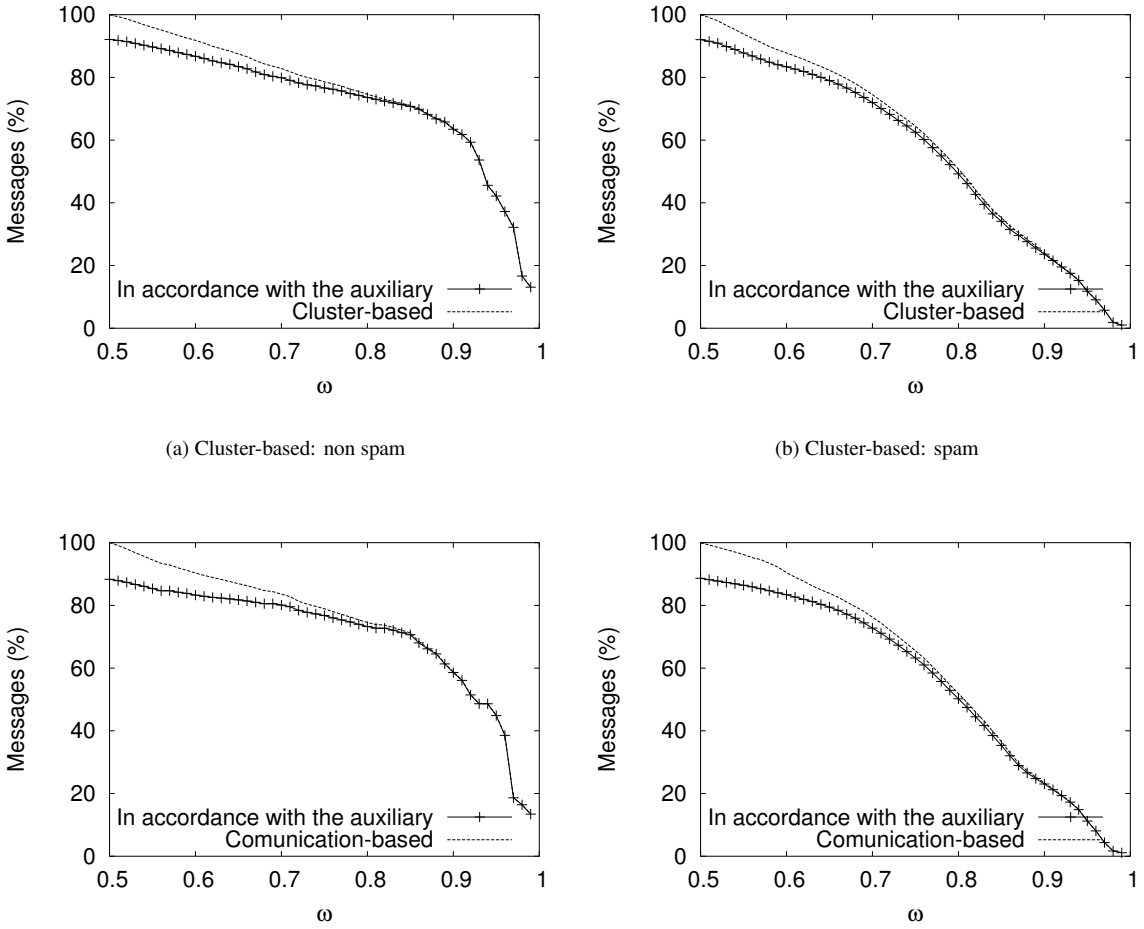
Figure 6.8: Probability of a Message Being Spam as A Function of its Communication-based Spam Rank.

simulation by varying ω, considering only messages classified by the auxiliary as legitimate (Figure 6.9 a-c) and spam (Figure 6.9 b-d). The difference between these curves is the set of messages that were classified differently from the original classification.

It can be observed from the plots that the communication-based (Figure 6.9 c-d) algorithm is able to classify more messages than the cluster-based (Figure 6.9 a-b) one. We conjecture the reason is that each cluster has a set of properties that are explicitly recognized by the communication-based approach and not by the cluster-based one, which instead needs to rely more often on the classification provided by the auxiliary algorithm.

In another experiment, we simulated a different algorithm that also makes use of historical information provided by an auxiliary spam detector described in [3]. The main differences are that it uses historical information of each sender separately and it does not use recipients information. We built a simulator for this algorithm and executed it against our data set. The results show that it was able to classify 85.11% of the messages in accordance to the auxiliary algorithm, while our approach classified more than 95% with $\omega = 0.85$. It is important to note that our algorithms can be tuned by the proper set of threshold ω. The higher the parameter ω the more our algorithms classify in accordance with the auxiliary classification, but less messages are classified.

We believe that the differences between the original classification and the classification found for high ω values using both algorithms are generally due to misclassification by the auxiliary algorithm. In our data set we have access to the full body of the messages that were originally classified as spam. Therefore, we can evaluate the fraction of the total amount of false positives (messages that the auxiliary algorithm classified as spam and that one of our algorithms classify as legitimate) that were generated by the auxiliary algorithm. This is important so that



(a) Cluster-based: non spam

(b) Cluster-based: spam

(c) Communication-based: non spam

(d) Communication-based: spam

Figure 6.9: Messages Classified in Accordance With the Auxiliary Algorithm and the Total Number of Messages Classified by Varying ω.

the common belief that the cost of false positives is much higher than the cost of false negatives [8].

Each of the possible false positives (for $\omega = 0.85$, resulting in 1,850 emails) were manually evaluated by three people so as to determine whether such a message was indeed spam. Figure 6.10 shows the percentage of messages correctly classified by our two algorithms and by the auxiliary one in this subset. This graphs show the evolution of the success rate, after a warm-up phase of 50% of the messages, as messages were processed and more information about email users was gathered. Both algorithms outperform the auxiliary since they generate less false positives. Note that the cluster-based algorithm is clearly the best option until nearly 75% of the data is included, showing that it needs less training examples than the communication-based algorithm. After that, the two algorithms have a similar behavior.

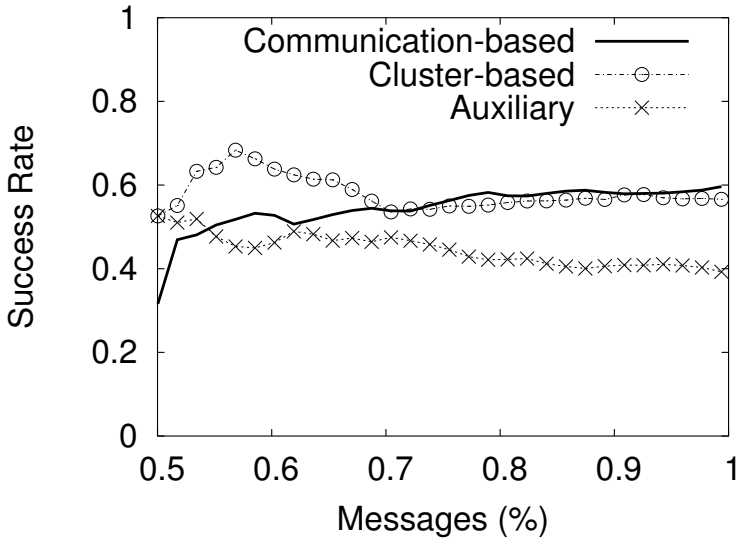


Figure 6.10: Classification Effectiveness.

Due to the cost of manually checking messages we can not afford to classify all of the messages classified as spam by the auxiliary algorithm. However, we evaluate a randomly chosen fraction of the messages classified as spam by the auxiliary and by our algorithms, which represents the total data with a confidence level of 99% and a confidence interval of 3% [65]. With $\omega = 0.85$, we found that 15% and 37% of the total messages are in this group for the cluster-based and communication-based algorithms, respectively. Moreover, we analyzed manually a sample of 3.50% (1,708 emails) and 1.52% (1,821 emails) for the cluster-based and communication-based algorithms, respectively. We found that 99.9% and 99.5% of the analyzed messages were correctly classified using cluster-based and communication-based algorithms, respectively, showing the high precision of both proposed algorithms.

Finally, we emphasize that we can not determine the legitimate classified messages classification quality, since we do not have access to the full body of those messages.

6.4.2 Synthetic workload

In this section we will use a set of synthetic data to evaluate our approach. In Section 6.4.1.1 we showed that our approach is able to correct false positives generated by the auxiliary algorithm but the synthetic data used in this section allows us to better estimate the resilience of our approach against errors of the auxiliary algorithm and also against email address forging.

The synthetic workload generator developed does not make use of most of the characteristics of email traffic as presented in some recent papers [9, 10, 5, 4, 2]. Instead, we are interested in creating a well controlled data set where senders and recipients can be clustered so that our algorithms can be evaluated under different spammer

Parameter	**I**nterpretation
Spam senders fraction	Used to tune the number of spammers that are in the workload. Therefore, this parameter is indirectly responsible by the total number of spam messages generated.
Legitimate senders fraction	Used to tune the number of legitimate senders in the workload. Similar to the *spam fraction*, this parameter is indirectly responsible by the total number of legitimate messages generated.
Mixed-behavior senders fraction	Used to determine the percentage of senders in the workload that send both spams and legitimate messages. For each sender that falls in the mixed class, a probability of it sending a spam is uniformly selected.
# senders	The total number of senders in the workload.
# recipients	Total number of recipients in the workload.
# distribution lists	Each sender is assigned to a specific distribution list that represents the users to which they send messages. This parameter regulates the amount of distribution lists that are generated, in order to have a clustering of the senders this parameter must be lower than *# senders*.
Distribution list size	Size of each distribution list.
# messages	Total number of messages generated.
Error rate	This parameter is used to tune the quality of the auxiliary algorithm. If it is set to 0 it is generated a log stream with 0% of miss-classification, while a 0.5 value of this parameter means a log stream with 50% of miss-classification.

Table 6.2: Synthetic Workload Generator Parameters.

strategies, different auxiliary algorithms and amount of misclassifying errors. Table 6.2 summarizes the parameters considered.

In order to generate the synthetic log stream we first create a set of recipients according to the parameter *# recipients* and separate them in *# distribution lists*. These lists are assigned to senders to distribute their messages. Each list has the same number of recipients determined by the parameter *distribution list size*. Recipients are randomly [3] selected and inserted into the lists. Note that there may be intersections between the distinct lists, this is an important feature, valuable to evaluate the behavior of our clustering approach.

Each of the distribution lists is then classified as spam, legitimate or mixed according to the probabilities, denoted *spam senders fraction*, *legitimate senders fraction* and *mixed-behavior senders fraction* parameters. Note that this doesn't mean that a recipient will receive only spam or legitimate messages since there may be intersections between the different lists.

Second, we create a set of senders following the *# senders* parameter. Similarly to what has been done to distribution lists, the senders are separated into classes according to *spam senders fraction*, *legitimate senders fraction* and *mixed-behavior senders fraction* parameters. Each sender is assigned to an specific distribution list

[3] Hereafter, whenever we mention random in the synthetic workload generator we mean that elements are selected following a uniform distribution.

with the same classification.

Finally, we generate the *# messages*. For each message, we randomly select a sender. Note that, as explained in Table 6.2, the parameters *spam senders fraction* and *legitimate senders fraction* somewhat indirectly determine the number of messages generated for each class. The only exception is when a sender is a mixed-behavior sender. These senders have an associated probability of sending spams that is randomly selected from 0 to 1 during its creation.

The classification for each message is determined by its sender. If it is a spammer every message coming from that sender is classified as spam, and if it is a legitimate sender each message coming from him is considered as legitimate. If the sender is a mixed-behavior sender the type of the message is selected based on his probability of sending spams. After this procedure the message goes through a classification procedure that creates a misclassification or not, according to the *error rate* parameter.

Parameter	**V**alue
Spam senders fraction	(100% - *mixed-behavior senders fraction*)/2
Legitimate senders fraction	(100% - *mixed-behavior senders fraction*)/2
# senders	2,000
# recipients	3,000
# distribution lists	1,000
Distribution list size	10
# messages	25,200

Table 6.3: Basic Parameters for Workload Generation.

Table 6.3 shows the basic parameter values used in the generation of all synthetic workloads of this chapter. These parameters were chosen so that the workloads generated resemble the actual workload described in Section 6.4.1. The size of the synthetic workloads is approximately 7% of the size of the original workload. We believe this provides us with a controlled and sound set of workloads to test our algorithms.

Workload type	Mixed-behavior fraction	Error rate
Well-defined	0%	0 - 1
Mixed	0% - 100%	0.20

Table 6.4: Workload types analyzed.

The parameters we vary in our experiments are the *error rate* and the *mixed-behavior senders fraction*. They allow us to analyze the capability of our algorithm to correct misclassification and the potential effect of email forging by spammers. In Table 6.4 we show how we vary these parameters to create the two types of synthetic workloads analyzed here. In one of these types, that we call *Well-Defined Workload*, the *Mixed-behavior senders i/fraction* was set to 0 and the *error rate* is varied. In the other type, that we call *Mixed Workload*, we vary the *error*

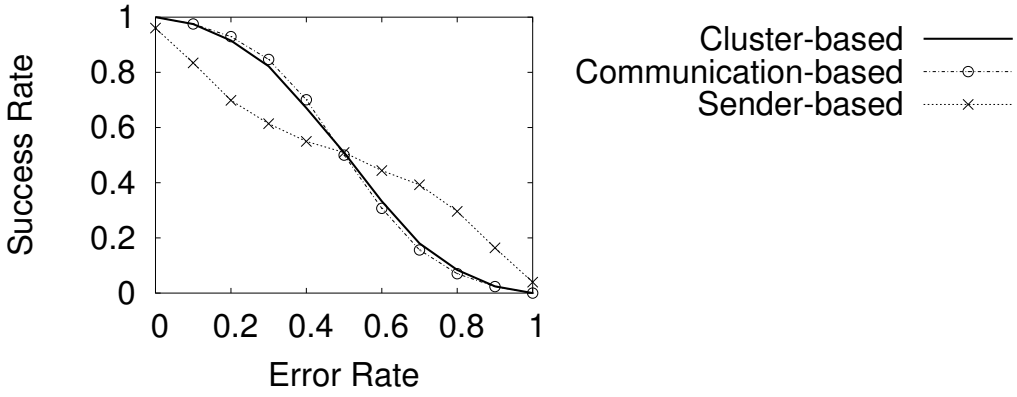


Figure 6.11: Actual Success Rate in Terms of the Success Rate of the Auxiliary Algorithm for the *Well-defined* Workload. The Parameter ω is Set to 0.85.

rate and the *Mixed-behavior senders fraction* parameters. Note that in the latter, the percentage of both spam and legitimate is (100 - *Mixed-behavior senders fraction*)%.

6.4.2.1 Classification results

This section presents the results of simulations of our two algorithms and of the sender-based approach algorithm (see Section 6.4.1.1) using our two types of synthetic workloads. We are interested in testing the resilience of the algorithms in terms of the *error rate* of the auxiliary algorithm. Figure 6.11 shows the classification success rate by varying values of the *error rate* parameter using the *Well-defined* workload.

A general conclusion drawn from Figure 6.11, is that both our algorithms perform better than the sender-based algorithm. This is mostly because of two factors: (i) The sender-based algorithm initially classifies messages as spam if they don't have historical information about their senders; (ii) There is a parameter in this algorithm that is similar to ω in our approach and it is set to 0.5 according to [3]. The reason for the sender-based algorithm performing better than the auxiliary [4] for *error rates* higher than 0.5 and worse for lower rates is its pessimistic assumption of considering each first message from each sender as spam. Also from Figure 6.11, we can see that the two proposed algorithms have similar behavior. Nevertheless, the communication-based algorithm has performed slightly better than the cluster-based one.

We were also interested in the resilience of the algorithms in terms of the possibility of users sending both spam and legitimate messages at the same time. Figure 6.12 shows the results obtained by the three algorithms for the *Mixed* workload. Our algorithms are still better than the sender-based and it is clear that if the amount of addresses that are used both for sending spams and legitimate messages is small our algorithms are still capable of producing

[4]The success rate for the auxiliary algorithm in Figure 6.11 can be computed as 1 - *Error Rate*

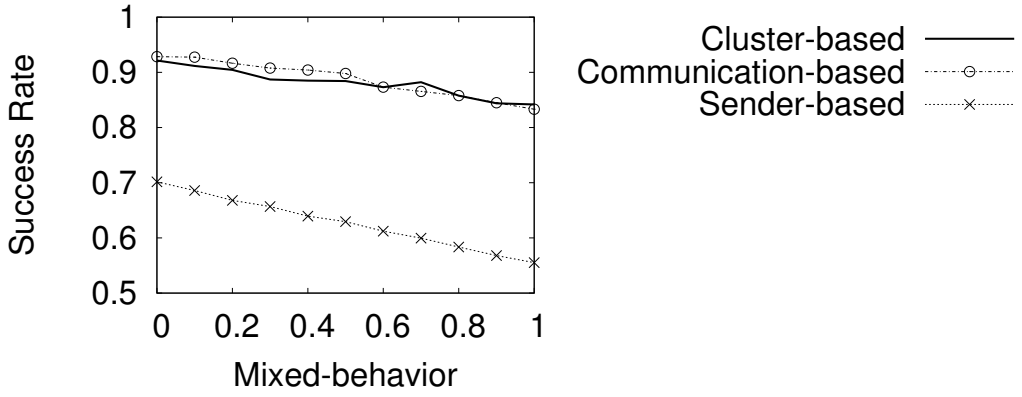


Figure 6.12: Actual Success Rate in Terms of the Success Rate of the Auxiliary for the *Mixed* Workload. The *error rate* for this Experiment was Set to 0.20. The Parameter ω is Set to 0.85.

better results than the auxiliary algorithm alone. Besides, the communication-based algorithm is more sensitive to users sending both spam and legitimate messages at the same time. For the workload described in Section 6.4.1 we have computed the percentage of users that send both spam and legitimate messages is approximately 23%. Other studies show that if we consider the full address of the user as its identification, this number is approximately 0.94% [9].

6.4.3 Weighted vector representation

One question that arises is whether a weighted vector representation for email users would be better than a binary representation. In order to answer this question we simulated our two proposed algorithms considering the vector identification of each user defined as follows.

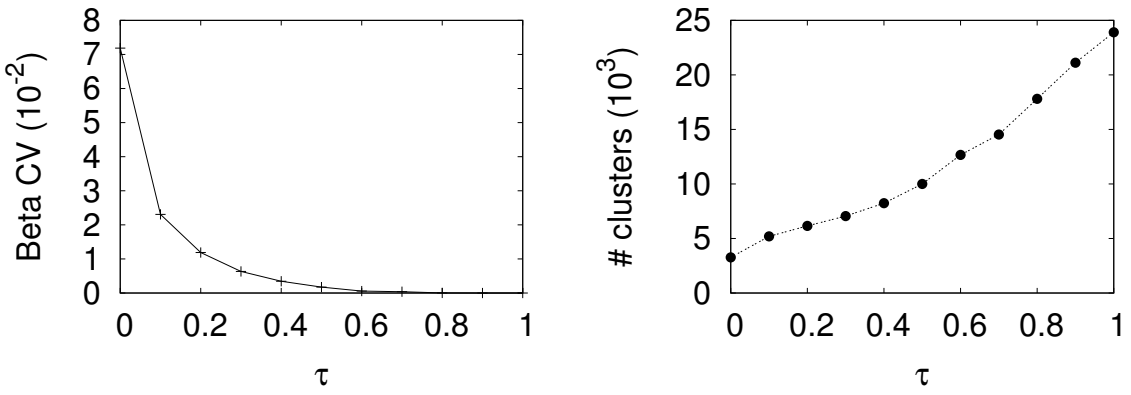
Let N_r be the number of distinct recipients. We represent sender s_i as a N_r dimensional vector, $\vec{s_i}$, defined in the space created by the email recipients being considered. The n-th dimension (representing recipient r_n) of $\vec{s_i}$ is defined as:

$$\vec{s_i}[n] = \begin{cases} m_{i,n}, & \text{if } s_i \rightarrow r_n \\ 0, & \text{otherwise} \end{cases}, \tag{6.7}$$

where $m_{i,n}$ indicates the number of messages that have been sent by s_i to r_n.

Similarly, we define $\vec{r_i}$ as an N_s dimensional vector representation for the recipient r_i, where N_s is the number of distinct senders being considered. The n-th dimension of this vector is set to $m_{i,n}$ if recipient r_i has received exactly $m_{i,n}$ emails from s_n.

Figure 6.13 shows the variation of beta CV and number of cluster by varying τ while considering the weighted



(a) Beta CV

(b) Clusters

Figure 6.13: Number of Email User Clusters and Beta CV vs. τ for the Weighted Representation.

representation for email users. As can be noted, our algorithm is not able to cluster the users as properly in this scenario as it was able to do in the binary case (see Figure 6.5). We conjecture that this is due to the greater variability of users' communication patterns that have been presented in other studies [9].

6.5 Conclusions

In this chapter we propose and evaluate two algorithms to improve spam detection mechanisms based on the structural similarity between contact lists of email users. The idea is that contact lists are much more stable identifiers of email users than message contents, sender domain or user names which all can be changed quickly and widely. The proposed algorithms work in tandem with an auxiliary classifier. We have shown that we can successfully group spam and legitimate email users into clusters and use them to improve the quality of spam detection.

Specifically we have implemented a simulator based on data collected from the main SMTP server for a large university that uses SpamAssassin. We have shown that our algorithm can be tuned to produce classifications similar to those of the original classifier algorithm and that, for a certain set of parameters, it was capable of correcting false positives.

Moreover, we have also experimented with synthetic workloads. The results show that our algorithms are able to correct misclassification both when the auxiliary classifies the messages incorrectly and when a single user identifier is used to send both spam and legitimate email.

Whereas we expect the communication-based approach to have a better performance in the long-term, it is more sensitive to the quality of the clusters generated or the time period used to train it. On the other hand, the cluster-

based approach works better for small training data and is less sensitive to the quality of the clusters generated.

The use of weights in the vector representation to cluster email users has been studied to some extent but has shown worse results than the simple binary approach. We believe this happened because of two facts: the greater variability in the way legitimate users exchange messages and the simplicity of the clustering algorithm that we use in our approach.

Chapter 7

Social versus Anti-Social Behavior in Email Traffic

7.1 Introduction

The characterization of email traffic as a complex social network [1] has been the main topic of several recent studies. To the best of our knowledge, all of these characterizations have dealt with the fact that email traffic represents not only social interaction, but also other types of relationships between email users with simple strategies. Examples of such strategies are the restriction of email traffic internal to single organizations [13, 14, 15, 16, 17], restriction of email users in such a way that only users that participate in two-way communications are considered [15, 17, 16], elimination of emails users with a high volume of messages [15, 16, 17], and elimination of links connecting email users that represent a low number of messages exchanged [14].

On the other hand, there has been growing interest in uncovering evidence of *antisocial* behavior in online networks. Recent work addresses topics such as uninhibited remarks, hostile flaming, non-conforming behavior, group polarization, and spurious traffic [18, 19]. Email as a mean of possible mass distribution is particularly associated with the dissemination of computer viruses as well as spam traffic [20], that flood the Internet with unwanted messages usually containing commercial propositions or, more recently, a variety of other scams. This behavior, which we call generically *antisocial*, displays different characteristics from social relations for which social networks have been extensively constructed and analyzed.

What is conceptually interesting about this antisocial behavior is that it generates quantitative graph theoretical and dynamical properties that are nontrivial and reflect a certain type of interaction that we quantitatively characterize and contrast to the general properties of social networks.

[1] According to Newman, a *social network* is a set of people or groups of people connected through patterns of social interaction, which can be represented as nodes and links, respectively, in a graph [12].

In this chapter it is presented a characterization of relations in email traffic tackling the problem from a distinct point of view. First, most of the previous restrictions adopted in previous studies are lifted. Second, we consider a single aggregated email traffic as formed from two different components: a legitimate component that is generated by legitimate users during their social interaction through email exchange and a non-legitimate component that is generated by users that use their email addresses to send unsolicited bulk messages generally called spam. Whereas, the legitimate component of our study dominated by social behavior, the non-legitimate component posses a behavior that we call "antisocial behavior".

Our key findings are:

- We find a power law degree distribution for all nodes of the four networks analyzed in social and antisocial traffics. However, the power law model is a better model of the degree distribution in social traffic than in antisocial traffic.

- We found that users in the antisocial traffic have significant lower CC than in the social traffic, mainly in the networks that include external users.

- We found uncorrelated neighbor's degree distributions, for the social and antisocial networks that include the external users. On the other hand, due to the unusual behavior of sending spams to forged internal users, in one of our workloads, we found that the internal antisocial network has assortative mixing. Finally, all internal social networks present positive degree correlation between neighbors.

- We find a power law distribution of community sizes for all defined networks, in the two workloads. Nevertheless, we find that the antisocial communities in the networks that include external users have on average near 30% more nodes than social communities. Besides, we find that the biggest community in each social network is almost twice as larger as the corresponding community in the antisocial network, for the two workloads.

- We find that nodes in social traffic have a significantly greater coefficient of preferential exchange than its pairs in the antisocial traffic. Moreover, nodes in the antisocial network that includes external users have a low, but sometimes non zero, coefficient of preferential exchange.

- Finally, the results show that social email traffic has lower entropy (higher structural information) than antisocial traffic for both work and non-work periods. The larger the time interval under analysis, the more noticeable this difference becomes, thus capturing longer patterns of communication and the presence of time correlations.

Therefore, our characterization reveals significant differences between the social and antisocial traffics. These differences are possibly due to the inherent distinct relationship nature of the email senders and their connections with email recipients in each group. Whereas, a legitimate email transmission is the result of a bilateral relationship, typically initiated by a human being, driven by some social relationship, a spam transmission is basically a unilateral action, typically performed by automatic tools and driven by the spammers' antisocial will to reach as many targets as possible.

The remaining of this chapter is organized as follows. Our characterization methodology and the email workloads are described in Sections 7.2 and 7.3, respectively. Section 7.4 analyzes the structural characteristics of the defined networks. The dynamical characteristics of the social and antisocial networks are studied in Section 7.5. Finally, in Section 7.6 we present our conclusions.

7.2 Methodology

In order to characterize the social and antisocial behaviors and show how they reflect in the structural and dynamical properties of email networks we define two types of graphs [2]: the *user graph* and the *aggregated graph*. In both graphs we can distinguish two groups of nodes: the first composed of users that are under the domain of the workload, see Section 7.3, under analysis and, another group composed of outside users.

The vertices of the *user graph* are email senders and recipients. An email sent by A to receiver B results in an edge between A and B. The *aggregated graph* is a graph with only two vertices, one of them aggregating the internal to the domain recipients and the other one aggregating external to the domain senders.

The unlabeled edges of the *user graphs* can be directed or undirected, while the edges of the *aggregated graph* are directed and labeled with the flow of messages between the connected nodes.

The *user graph* is used for identifying the structural characteristics of the networks and in the analysis of the dynamical properties of the communication between pairs of users. On the other hand, the study of the dynamics of the incoming flow of messages is done by using the *aggregated graph*.

We constructed four *user graphs* representing different email networks that emerge from the two workloads described in Section 7.3. A *social network (SN)* is built from the legitimate messages exchanged between all the users, including those external users that send/receive emails to/from internal users. Similarly, an *antisocial network (AN)* is built from the spam messages exchanged between all users. An *internal social network (ISN)* is built considering internal users exclusively involved in legitimate email communication. Finally, the internal spam

[2] In Chapter 3 we present the definition of graph and a set structural graph properties we analyze in this Chapter

traffic [3] is used to build an *internal antisocial network (IAN)*. In general these networks are undirected, unless otherwise noted.

We also note that messages exchanged through mailing lists, which also involve bulk email traffic, may exhibit antisocial characteristics. As in [60], aiming at minimizing the impact of such communication patterns on our analysis, we remove from our internal social network users who exchange emails with 40 or more other users.

To study the incoming flow of messages we constructed two *aggregated graphs*. The first uses only outside domains, legitimate messages income, and the other one with the incoming flow of spam messages.

Finally, in order to extract the communities of our networks we define, for each network, a similarity graph. A similarity graph is a complete graph [4], where users are the vertices and each pair of vertices is linked by a labeled edge. The edge linking two users is labeled with the number of common neighbors the linked users have. We constructed one similarity graph for each network in each workload.

7.3 Experimental data analysis

Log	**Period**	**Social** #emails (#users)	**Antisocial** #emails (#users)
1^{st}-log	11/18/04 to 12/31/04	292173 (107016)	270491 (178360)
2^{nd}-log	05/16/05 to 08/23/05	857596 (285005)	361955 (149391)

Table 7.1: Workload Summary

Our email workload consists of two anonymized SMTP logs of emails from a department, with at about 1,400 users including staff, and administrators, of a large university in Brazil. The logs were collected at the central Internet email server of the department. The server handles all emails exchanged inside the department as well as all the emails exchanged between external and internal department users. Statistical characteristics of the workloads are in agreement with previous email traffic analysis [4, 10, 3].

Table 7.1 summarizes the data sets. The logs were collected from a department that runs its own set of pre-acceptance spam filters, including local black lists, local heuristics and, more recently, a GreyList scheme [46, 29, 3]. Besides, for all external incoming emails the server runs a public spam filter, SpamAssassin, see Section 4.2.1 in Chapter 4 for more information about it, which classifies the emails as spam or not [51].

We have 48% and 30% of all emails classified as spam in the first and second logs, respectively. The difference in the fraction of spams in the logs is due to the adoption of distinct pre-acceptance schemes of spam detection by the department between the two log collections. Note yet that the collection time of the 2^{nd}-log is greater

[3] Originating from and addressed to an internal user. These are usually the result of forged identifiers.

[4] A complete graph is a graph where all pair of nodes are connected by a link.

than the collection time of the 1^{st}-log by, at about, half an year, as can be seen in Table 7.1. Therefore, these logs will allow us to show that the structural and dynamical properties of the social and antisocial networks are somehow stable over time. Furthermore, as the rules of the pre and pos-acceptance spam detection used in the logs are independently defined, we intended to show that the properties studied here are more related to users behavior aspects than to the particular characteristics of the log.

	Network - #Vertices (#Undirected Edges)			
Workload	Internal Social	Social	Internal Antisocial	Antisocial
1^{st}-log	2211 (12630)	107016 (176454)	1274 (2607)	178360 (480226)
2^{nd}-log	3147 (19450)	285005 (383601)	1129 (3542)	149391 (368718)

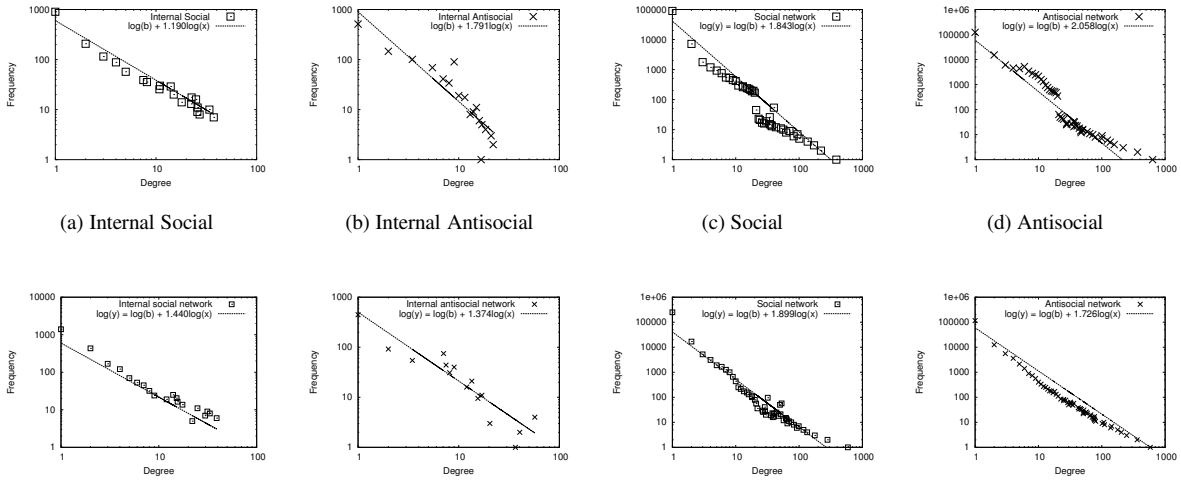
Table 7.2: Networks summary

The construction of the networks is subject to at least two important constraints. The first is that our knowledge of the external users behavior is limited to their contacts inside our log domains, i.e., we do not have emails exchanged among external users. Second, as our antisocial traffic is composed of spam emails, we deal with a large number of forged user names. Particularly, we get access to the user names of the internal antisocial traffic of the 1^{st}-log and find out that almost all users are not real users.

Table 7.2 shows the basic statistics of the constructed networks. Due to the use of forged user names by spammers the size of the internal antisocial networks is not negligible. Moreover, they have very distinct structural properties as we show in the next Sections.

7.4 Network structural characteristics

In this Section we study the structural characteristics, refer to the Chapter 3 for the formal definition of the metrics, of the four email networks defined. In order to show that the differences in the social and antisocial use of emails reflect in very distinct structural properties of the networks, we study a set of graph properties commonly used to distinguish social networks from other complex networks in the literature. Section 7.4.1 presents results for the degree distribution of the networks. Results for the distribution of clustering coefficient are presented in Section 7.4.2. In Section 7.4.3 is analyzed the size of the connected component of the networks. Results for the correlation between neighbor node degrees of the networks are presented in Section 7.4.4. Finally, Section 7.4.5 presents a study of the distribution of the social and antisocial community sizes in our workloads.



(a) Internal Social (b) Internal Antisocial (c) Social (d) Antisocial

(e) Internal Social (f) Internal Antisocial (g) Social (h) Antisocial

Figure 7.1: Distribution of node degree - a-d 1^{st}-log and e-h 2^{nd}-log

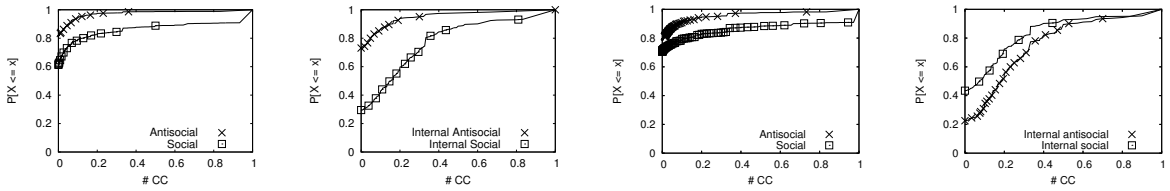
	Network - α (R^2)			
Workload	Internal Social	Social	Internal Antisocial	Antisocial
1^{st}-log	1.190 (0.964)	1.843 (0.933)	1.791 (0.831)	2.058 (0.910)
2^{nd}-log	1.440 (0.957)	1.899 (0.935)	1.374 (0.861)	1.726 (0.918)

Table 7.3: Parameters of the degree distribution models

7.4.1 Degree distribution

We find power law degree distributions for the undirected versions of our networks, as in a diversity of previous studies [80, 57, 13]. The exponents (α) and error estimations R^2 are shown in table 7.3. Note that, as expected, all the social networks are, in the two logs, very well modeled by a power law distribution with $R^2 > 0.93$. On the other hand, the internal antisocial networks are not well modeled by a scale-free distribution with $R^2 < 0.87$. Moreover, the proposed antisocial network fits are not as good as the proposed models for any of the social networks.

Figures 7.1 a-h show the log-log plots of the degree distribution of the internal social, internal antisocial, social and antisocial networks, respectively, for the two logs. The asymptotic decay of the head and tail of the internal antisocial degree distribution plots, Figures 7.1-b and 7.1-f, have a strong influence in the weakness of the internal antisocial proposed models. Note that, these results are in accordance with results showed in Chapter 4.



(a) 1^{st}-log (b) 1^{st}-log (c) 2^{nd}-log (d) 2^{nd}-log

Figure 7.2: Distribution of the clustering coefficient

In spite of these results, the degree distribution is a weak discriminator, mainly for the networks that include external users, between social and antisocial behavior. Moreover, it is clearly affected by the incomplete knowledge of part of the networks, which happens whenever external users are considered. Such lack of knowledge clearly leads to an underestimation of the degree of external users, which in turn results in the incorrect shifting of these users to lower degree.

7.4.2 Clustering coefficient

	Network - CC (σ^2)			
Workload	Internal Social	Social	Internal Antisocial	Antisocial
1^{st}-log	0.226 (0.073)	0.137 (0.091)	0.052 (0.023)	0.026 (0.014)
2^{nd}-log	0.166 (0.059)	0.163 (0.109)	0.244 (0.064)	0.042 (0.023)

Table 7.4: Mean Clustering Coefficient of the Networks

According to Newman and Park, a high clustering coefficient is a theoretical graph quantity typical of social networks [80, 57]. Therefore, we investigate whether this structural property of our undirected email user networks can distinguish the social imprint of legitimate email communication from the antisocial characteristics of spam. The CC of a node n, denoted C_n, is defined as the normalized probability of any two of its neighbors being neighbors themselves. This metric is associated to the number of triangles that contain a node n. For an undirected graph, the maximum number of triangles connecting the N_n neighbors of n is $N_n \times (N_n - 1)/2$. Thus, the CC measures the ratio between actual triangles and their maximum value.

Table 7.4 shows the mean and variance of CC for all the networks in the two logs. Note that, the mean CC of the internal social networks is higher and comparable with that found by other authors [80]. On the other hand, as generally expected, the mean CC of the antisocial networks that include external users is almost zero, in the two logs. Nevertheless, the spammers unusual behavior of sending messages to forged internal user names makes the mean CC of the internal antisocial network in the 2^{nd}-log be higher than the CC of its internal social network.

Moreover, note that the internal antisocial network in the 2^{nd}-log is 13% smaller, in number of users, but 36% greater, in number of directed links, than its pair in the 1^{st}-log.

Figure 7.2 a-d shows the distribution of the CC of nodes of all networks, in the two workloads. As expected, in the 1^{st}-log, Figure 7.2 a and b, users of the antisocial traffic have significantly lower CC than in the social traffic. Similarly, in the 2^{nd}-log, Figure 7.2 c and d users in the social network have greater CC than in the antisocial network. However, also in the 2^{nd}-log, against our expectations, users of the internal antisocial network have greater CC than users of the internal social network.

On the other hand, all four networks in the two logs contain a significant fraction of their nodes with zero clustering coefficient, but this proportion is much higher for graphs that include external users and/or antisocial components. Specifically, 61% and 68% of all nodes in the social networks of the 1^{st} and 2^{nd} logs have $CC = 0$, respectively. On the other side, this becomes more than 81% in the 1^{st}-log and more than 78% in the 2^{nd}-log for the nodes in the antisocial component.

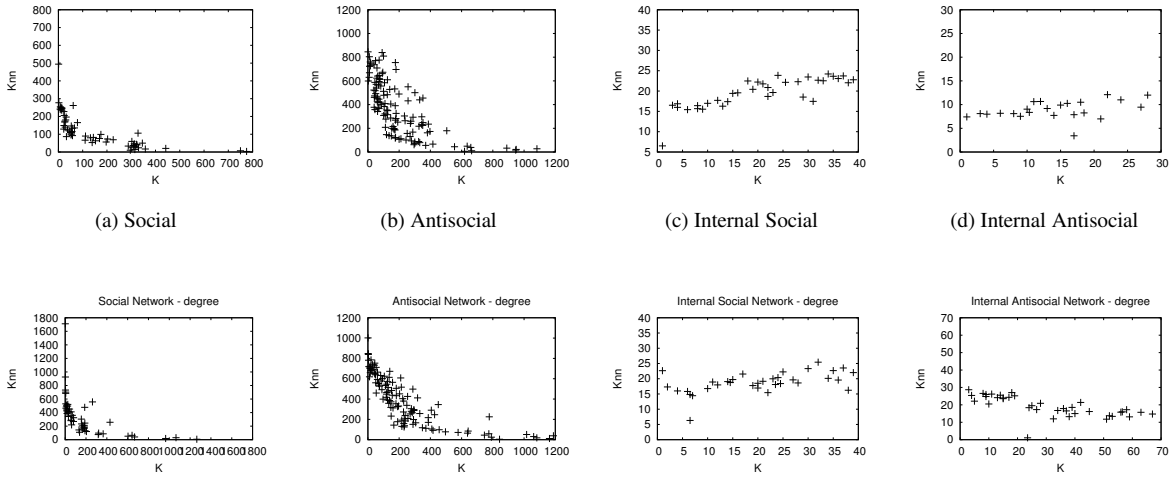
The internal social network, in the 1^{st}-log, has only 29% of its nodes with $CC = 0$ compared to 73% for the internal antisocial network. Due to the spammers behavior of sending emails to forged internal users, observed in the 2^{nd}-log, we found that, while 23% of the internal antisocial users have $CC = 0$, in the internal social network 43% of the users have zero clustering coefficient.

These results show that there are clear differences on the transitivity of the social and antisocial components of our email networks, but also that low clustering is not a sufficient condition for a node to be associated with antisocial behavior. Similarly to our analysis of degree, these results also indicate that the separation of the two traffics is important in order to generate a truly social component. Failure to do so will result in underestimation of the average social network transitivity.

7.4.3 Connected components

Some recent studies [6] have analyzed graphical metrics of the giant connected component of email graphs. A connected component is a subset of the vertices of a graph, such that one node can be reached from any other node following the set edges between them [66].

We found, for undirected versions of our networks that in all the social networks there is a giant connected component with more than 92% of the vertices. On the other hand, the antisocial networks also have a giant connected component, in this case with more than 88% of the network vertices. We conjecture that, in the social networks, the giant connected component is mainly due to the exchange of messages between users. Otherwise, the great overlap of the distribution lists of the spammers as showed in Chapter 4, seem to be responsible for the giant connected component found in the antisocial networks.



(a) Social (b) Antisocial (c) Internal Social (d) Internal Antisocial

(e) Social (f) Antisocial (g) Internal Social (h) Internal Antisocial

Figure 7.3: Variations of the neighbors degree with degree k of a node - a-d 1^{st}-log and e-h 2^{nd}-log

The giant connected components shows no clear difference between social and antisocial network components. Therefore, it is not a good measure to use in distinguishing the truly antisocial/social component.

7.4.4 Assortative mixing

A social network is usually characterized by a positive degree correlation between neighbors [88, 72, 89, 57]. On the other hand, non-social networks, generally, have negative degree correlation as shown in [89]. Here we analyze the nature of degree correlations between neighbors in undirected versions of our email networks, by looking at $K_{nn}(k)$ (the mean degree of the neighbors of nodes with degree k as defined in [72]).

The $K_{nn}(k)$ is an increasing function of k when the network is said to show assortative mixing. Otherwise, it is a decreasing function when the network shows dis-assortative mixing [72]).

The common use of email easily creates imbalances of degree between the senders and recipients of messages, due to the easy and cost less way an email can be addressed to number of recipients, this tends to generate negative correlation between degree of neighbors of at least some legitimate users. Particularly, spam senders invariably follow the strategy of increasing their degree indiscriminately and maximally, and consequently reach, on average, a population of senders with much lower degree, which are statistically much more abundant for a scale free degree distribution.

Figures 7.3 a-h show the K_{nn} as a function of k for our four networks, of the 1^{st} a-d and e-h 2^{nd} logs, respectively.

The social and antisocial networks shows dis-assortative mixing, as seen in Figures 7.3 a and b for the 1^{st}-log and e and f for the 2^{nd}-log. As expected, the antisocial networks have a stronger disassortativity when compared with social networks, in the two traffics. The disassortativity of the antisocial networks is due, we suppose, to the strategy of spammers of trying indiscriminately to reach as much recipients as possible.

In order to understand the unexpected disassortativity of the social networks, we calculate, the statistical correlation between external sender degrees and its corresponding internal recipient degrees and find $C = -0.103$ and $C = -0.153$, in the 1^{st} and 2^{nd} logs, respectively. These results show that the nature of the dis-assortative mixing of the social networks is a result of the relationship between these groups of users. Therefore, in the social network, we suppose, the negative degree correlation is the result of the common subscription by many users to external distribution lists, such as those related to news, promotions, etc [5].

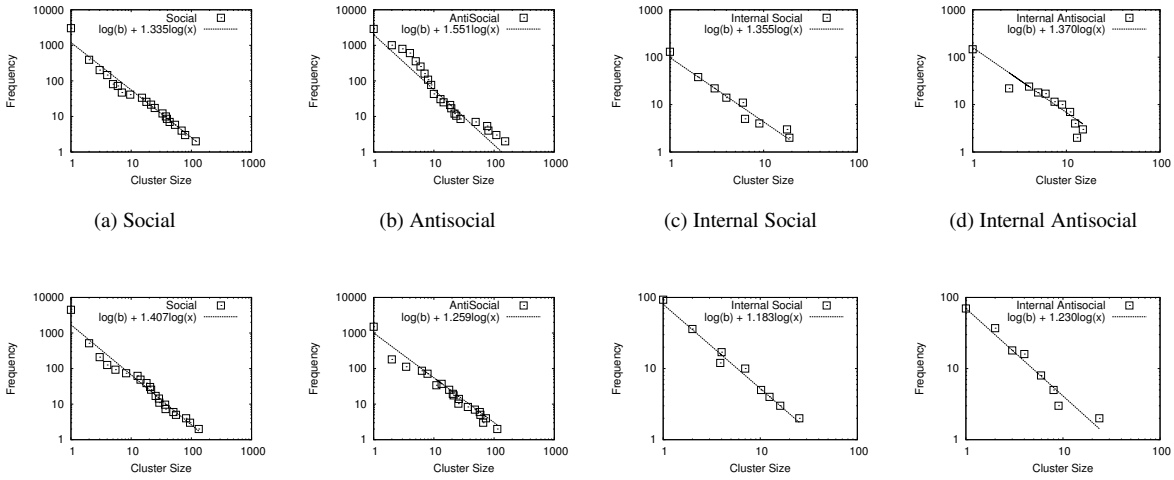
On the other hand, the internal social networks of the two logs, as expected and in accordance to previous analysis of other social networks [72], have a clearly positive assortative mixing, as it can be seen in Figures 7.3 c and g. Besides, the internal antisocial network, for the 2^{nd}-log, as expected, Figure 7.3-h, presents a clear disassortativity. In another way, against our expectations, the internal antisocial network of the 1^{st}-log presents a non negligible positive correlation ($C = 0.38$, for the points in the Figure 7.3-d) between the neighbors degrees. We conjecture, again, it's due to the behavior of spammers of exchanging spam messages among forged internal users.

In spite of these properties, the antisocial network built from the exchange of spam messages has definite properties, showing negligible transitivity and strong dis-assortative mixing. Moreover, our analysis shows, in contrast to previous expectations [57], that social email networks involving users that are external to the departmental domain may present a negative degree correlation, presumably reflecting the technological properties of email as a mean of distribution of legitimate information, rather than very asymmetric but reciprocal social exchanges.

7.4.5 Community sizes

Recent studies [57, 15, 14, 17, 90] suggest that social networks can be largely understood in terms of the organization of nodes into communities, refer to Chapter 3 for a formal definition of community, a feature that can explain, to some extent, the observed values for the clustering coefficient (as seen in Section 7.4.2), degree correlations (as seen in Section 7.4.4) and size of the connected component (as seen in Section 7.4.3). This observation has

[5] Recall that, unlike the internal social network, node degrees in our entire social network are not constrained, and thus, may represent distribution lists.



(a) Social (b) Antisocial (c) Internal Social (d) Internal Antisocial

(e) Social (f) Antisocial (g) Internal Social (h) Internal Antisocial

Figure 7.4: Distribution of cluster size- 1^{st}-log a-d and e-h 2^{nd}-log

	Network - α (R^2)			
Workload	Internal Social	Social	Internal. Antisocial	Antisocial
1^{st}-log	1.355 (0.959)	1.335 (0.968)	1.370 (0.900)	1.551 (0.951)
2^{nd}-log	1.183 (0.986)	1.407 (0.963)	1.230 (0.961)	1.259 (0.972)

Table 7.5: Parameters of the cluster size distribution models

indeed led to the interesting suggestion that email networks can be used to infer informal communities of practice within organizations [14], as well as their hierarchical structure [15, 14, 16, 17], features that can in principle be useful for the efficient management of human collective behavior and for learning about information flow in the organizations [59].

One of the key characteristics of social communities studied by various of previous works is the distribution of the community size [17, 15, 16, 91]. The size of a community is defined as the number of users or nodes of our networks inside it. In this Section we present a study of the size of the communities extracted from our networks.

Using the defined similarity graphs, differently from [17, 15, 16, 91] that used the clustering algorithm proposed by Girvan and Newman in [73], we applied a traditional hierarchical clustering algorithm, called Minimum Spanning Tree [74], to extract the communities.

Given a set of N nodes to be clustered, and an NxN similarity (the similarity between two nodes is defined in Section 7.2) matrix, the basic process of hierarchical clustering used is this:

1. Start by assigning each node its own cluster, so that if you have N nodes, you now have N clusters, each containing just one node.

2. Let the distances between the clusters equal the distances between the nodes they contain.

3. Find the closest (or most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

4. Compute distances between the new cluster and each of the old clusters (as the number common neighbors the users inside the new cluster and the users inside each old cluster have).

5. Repeat steps 1 to 4 until the happening of any of the previously chosen stop condition.

In [74] is proposed a method to stop the process of clusterings based on the quality of the clusters generated. The process consists of calculating the statistical CV of the distances (similarities) inter-clusters and intra-clusters at the end of each iteration. Note that, in order to have more knit tightly clusters we must minimize the intra-cluster distances CV and, as well as, maximize the inter-cluster's CV. The authors of [74] propose to use the number $intra-clustersCV/inter-clustersCV$ as a parameter, when this number stabilizes, i. e., when the aggregation of a node to a cluster do not mean significant change in this ratio, we are with the "best"possible number of communities.

Figures 7.4 a-h show the log-log plots of the frequency distribution function of the size of the communities extracted from each defined network, for the two logs, respectively. The proposed models are represented by the line in the plots.

Although, we use a clustering algorithm distinct from those used in [17, 15, 16, 91], our results are consistent with theirs. In accordance with their results, we found a power law distribution for the sizes of the social communities. Moreover, Figures 7.4 a-h show, clearly, that community sizes in the social traffic are more tightly modeled by a power law than communities sizes in the antisocial traffic.

We also find that unitary communities are much more probable in social networks ($P(X = 1) > 0.60$) than in antisocial networks ($P(X = 1) < 0.50$). Moreover, the mean size of antisocial communities is 29% greater than the mean size of social communities. These results suggest that sharing contact lists is a behavior more characteristic of antisocial users than of the social users.

On the other hand, while the greatest community in the antisocial networks are 979 and 1412 in the two logs, in the social networks we found 1578 and 2897, respectively, showing that social relationships result in larger community sizes when compared with antisocial relationships.

These results also indicate that the study of the communities of interest without the separation of the two traffics will result in wrong estimation of communities sizes.

So far we have examined metrics related to the topological skeleton of email networks and the dynamical exchange of emails between peers in the networks, that revealed differences between social and antisocial behavior.

These differences suggest mechanisms to differentiate legitimate human collaboration from opportunistic behavior on the basis of network structure, and have indeed been proposed as the basis for spam detection algorithms [11, 6]. However, much remains unsatisfactory about the transitivity and assortative mixing measures as means of characterizing patterns of human communication.

7.5 Graph dynamical characteristics

In this Section we study the dynamical characteristics of the communication in our networks. We expect to find out very different patterns of communications, in order to show the strong differences in the social and antisocial uses of email. By one side, social relationships are driven by the reciprocity in the communication as shown in [60, 9] and have a clear pattern of ask/answer as shown by [60, 61]. On the other side, antisocial communication, represented here by spam traffic, is characterized by the unilaterality and automation of the communication [9, 10].

In the next Section 7.5.1 we study the dynamics of communication links between senders and recipients directly, without reference to third parties. In the last Section 7.5.2 we consider means for quantifying the differences in the incoming flow of spam and legitimate messages, in order to do this we use the *aggregated graphs* as defined in Section 7.2.

7.5.1 Preferential exchange

	Network - $\langle E \rangle$ (σ^2)			
Workload	Internal Social	Social	Internal Antisocial	Antisocial
1^{st}-log	0.258 (0.140)	0.033 (0.025)	0.063 (0.040)	0.0001 (0.0001)
2^{nd}-log	0.158 (0.107)	0.023 (0.017)	0.160 (0.075)	0.0002 (0.0001)

Table 7.6: Coefficient of Preferential Exchange in the Networks

Here we investigate the simplest measure of communication between two users: reciprocity. We build a simple coefficient of preferential exchange E_i for user i as:

$$E_i = 1 - \left| \frac{\sum_{j \in C_i} [k(j \to i) - k(i \to j)]}{\sum_{j \in C_i} [k(j \to i) + k(i \to j)]} \right| \tag{7.1}$$

where C_i is the set of all users that have contact with user i within a given time period, and $k(j \to i)$ is the number of messages sent by user j to i. Therefore, $0 \leq E_i \leq 1$, with the lower end corresponding to no message replied, and the upper end with every message obtaining a response. This can be further averaged over all users to generate network averages $\langle E \rangle$.

Table 7.6 shows the mean $\langle E \rangle$ for each network of the two logs. Due to the partial knowledge of the complete social relations of the external senders, the internal social networks have coefficient of preferential exchange 5 times greater than the social networks that include external users in the two cases. Therefore, antisocial networks are naturally associated with small (but sometimes non-zero!) reciprocity, whereas social networks, in particular those containing legitimate users whose behavior we know most completely, are associated with the highest reciprocity in the two logs. Surprisingly, the internal antisocial networks have a higher $\langle E \rangle$ than the internal social network in the 2^{nd}-log and a non negligible value in 1^{st}-log. This is due, we conjecture, to the spammers behavior of sending spam to their own internal forged users names used to spread spam, what can be seen as a kind of simulation of a social behavior.

7.5.2 Traffic entropy

Up to this point we concentrated on the structure of the network of interactions mediated by email messages. In its construction as a graph we have not paid attention to the detailed temporal structure of message exchanges. An interesting question then is whether the dynamical properties of email traffic can distinguish social and antisocial relations.

This question has recently become a subject of interest. Eckmann, Moses and Sergi [60] have shown that coherent structures emerge from the temporal correlations between time series expressing short periods of intense message exchange between groups of users. Barabasi [61], on the other hand, has shown that the distribution of time intervals between email messages sent by a single user may be well described by a power law distribution, with bursts of activity alternating with long silences.

Both these characterizations identify properties of legitimate email traffic - temporal correlations between users and inter-message power law time statistics - that are thought to be exclusively social and thus not shared by the antisocial traffic component. In fact intense email exchanges between small groups of users are to be expected in patterns of human communication, creating the correlations observed by Eckmann, Moses and Sergi [60]. Barabasi in turn suggests that the power law statistics he observed can be explained in terms of a queueing model which encodes prioritization of tasks driven by human decision making.

Although suggestive, these interesting results were obtained for selected senders and receivers of email. Consequently, it remains unclear whether they hold for the general user or for aggregated groups of users. To this end, we investigated the statistics of our social and antisocial traffics by averaging over the behavior of all users. The first obvious temporal property of email traffic is its non stationarity. This creates difficulties for any attempt at statistical estimation. Social email traffic in particular shows large temporal variations, from night to day, working

days to weekends, and for our data set, strong seasonality associated with the academic calendar. As we show below, antisocial traffic displays weaker non-stationarity.

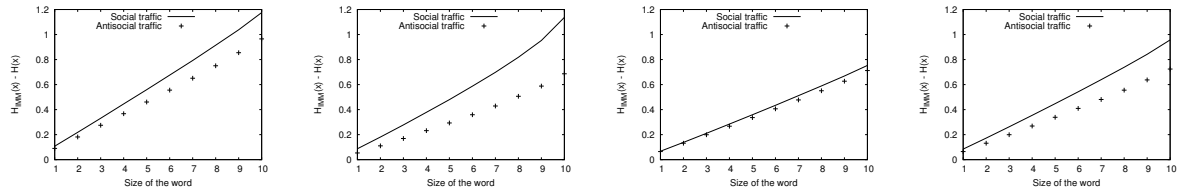
The second temporal feature of email traffic is an immediate result of the power law degree distributions described above. The majority of users do not communicate often with many others, but have instead low degree associated with an infrequent and often irregular usage of email. This means that the typical email user in our data - and, we believe, in most other large email networks - does not show time coherence with others, nor is he/she utilizing email under the temporal optimization pattern suggested by Barabasi.

To circumvent some of these difficulties, we attempted to identify statistical temporal patterns of communication that are characteristic of the social vs. antisocial aggregated traffics, using the *aggregated graphs* defined in Section 7.2. In so doing, we average over the behaviors of many users. Specifically, we represent temporal patterns of message arrival through the definition of a communication *word* of size d. The dimension d is the number of time intervals, or letters, in the communication word. Hence, a word is represented by a vector $W = (i_1, i_2, \ldots, i_d)$. The simplest representation of the traffic is through a binary assignment, where the value of i_j is set to 1 if one or more messages were exchanged in the corresponding time interval, or $i_j = 0$ otherwise. We estimate the probability for a given word to occur out of N realizations obtained from the measurement data through simple word frequencies. The entropy of the traffic is defined as usual as

$$H(W^d) = -\sum_{i=1}^{N} p({w^d}_i)\log_2 p({w^d}_i), \tag{7.2}$$

which is a function of word size d.

To illustrate these statements up consider the simplest statistical model that generates a binary time series subject to a given message arrival rate p. Then p can be written as the probability to obtain a 1 at each letter. If we further assume that bits corresponding to different letters are uncorrelated then the bit value at each letter can be regarded as the result of an independent Bernoulli trial. It follows that the probability for all words of length d with a given number n of 1s is given by the binomial distribution $P(p; n, d) = \binom{d}{n}\, p^n\, (1-p)^{d-n}$. Because all words with a given number n of 1s are equally likely, their probability is $p_w(p; n, d) = p^n\, (1-p)^{d-n}$. The corresponding entropy is also easy to compute as $H(W^d) = d\, m$, where $m = -(1-p)\log_2(1-p) - p\log_2 p > 0$ is the entropy per letter. The fact that the entropy is proportional to the word length d is a direct consequence of the assumed lack of temporal correlations. These expressions become especially simple if the temporal bin for each letter is chosen such that $p = 1/2$, in which case $m = 1$ is maximal. This independent message model (IMM) is the maximal entropy distribution for a traffic characterized by a message arrival p. As such real traffics must display



(a) Work hours (b) Non-work hours (c) Work hours (d) Non-work hours

Figure 7.5: Variation of the difference between the independent message model entropy and the entropy of the legitimate and spam traffics with the word size.

lower entropy relative to it.

In order to study these variations patterns we aggregated the data into two temporal periods: *work hours* (i. e. the period from 8AM to 8PM of the weekdays, except holidays, in the log) and remaining time which we aggregated as *non-work hours*. The difference between the maximal entropy model and the entropy of the real time series can be interpreted as the temporal structural information of each traffic.

Figure 7.5 a-d shows the variation with word size d of the difference between the independent message model entropy and the entropy of the legitimate and spam traffics during work 7.5-a and 7.5-c, and 7.5-b and 7.5-d non-work periods, for the two logs, respectively. All word probability distributions were constructed by normalizing the time bin for each letter word so that $p = 1/2$. As a result the time bin for each letter of the social traffic during work hours was set to 4s, and 11s for the corresponding non-work period. Time bins for the antisocial traffic were set at 4s during work hours and 5s otherwise. The excess curvature for large d is the result of poorer estimation of rare words.

The results show that social email traffic has lower entropy (higher structural information) than antisocial traffic for both working and non-working periods in the two workloads. The larger the work the more noticeable this difference becomes, thus capturing longer patterns of communication and the presence of time correlations. The difference between the independent message model, where for $p = 1/2$ all words are equally likely, and the real traffics is that in the latter words with many 1s are suppressed while the probability of words with two to three 1s separated by one to three 0s is enhanced. The difference between social and antisocial traffics is more subtle, with social email traffic displaying a greater probability for words with an isolated message in a long stream of silence. These structures are reminiscent to those found by Barabasi, but display less definitive statistical signatures. We see in general though that both social and antisocial traffics are not random, and that social email shows stronger temporal structure with a high probability for long silences and bursts of a few messages.

7.6 Conclusions

In summary, we have shown that the richness of behaviors in human communication - both symbiotic and opportunistic or antisocial - is present in the structure of networks of email communication and can be quantified via graph theoretical and time series analysis. Opportunistic nodes display antisocial behavior that can be captured graphically. Perhaps even more directly, antisocial email traffic can be identified by a greater statistical simplicity (higher entropy) in temporal patterns of communication, typical of the fact that each sender/recipient relationship is not developed to be unique and the same schemes are used to reach many recipients indiscriminately. Moreover, the ease to exchange email messages that leads to these opportunistic behaviors also has consequences for the truly social component of the network, which exhibits a power law degree distribution with a small exponent and, in some cases, small or negative assortative mixing by degree. We believe that the quantitative characteristics of antisocial communication patterns observed here for email networks are probably general to other opportunistic behaviors, bound to be present in other networks of human interaction.

Furthermore, we found that low network transitivity and preferential exchange are a strong imprint of antisocial behavior. So, in order to generate the truly social component we must separate the two components or we incur in misestimation of these characteristics.

Finally, although we found power law distribution for sizes of the extracted communities for all the networks, there are important differences in the way users can be clustered in communities in the two components. While communities made of only one node are much more common in the social components, the antisocial communities are on average two times bigger than social ones.

Chapter 8

Conclusions and Future Work

This thesis provides an extensive analysis of a spam and legitimate email users behavior, uncovering characteristics that significantly distinguish them.

Our statistical characterization of the email workload, presented in Chapter 4, based on the information available on the email headers, revealed that:

- While spam traffic is roughly insensitive to the period of measurement, legitimate traffic exhibits clear daily and weekly patterns;
- Spam email inter-arrival times are roughly stable over time, while legitimate email inter-arrival time is characterized by occurrence of bursts during working hours;
- Spam email messages are not only significantly smaller on average than legitimate emails, but also show much less size variability;
- Spam emails are addressed on average to a greater number of recipients than legitimate emails;
- Legitimate sender and recipient popularity and spam sender popularity, defined in terms of number of emails, follow a Zipf-like distribution, while spam recipient popularity does not;
- There are two distinct and non-negligible sets of legitimate recipients: those with very strong temporal locality and those who receive emails only sporadically. These two sets are not present in our spam workload and;
- Spam senders and recipients deal with many more contacts than legitimate senders and recipients.

However, we found that spam traffic significantly biases aggregated traffic from traditional legitimate traffic.

Chapter 5 shows that legitimate and spam email graphs differ in two fundamental classes of characteristics: structural, which capture the graphs' skeleton architecture, and dynamical, concerning node communication and graph evolution. We have shown that the spam and non spam subgraphs are structurally characterized by:

- Different distributions of the clustering coefficient of their nodes. Legitimate users present a higher clusterings coefficient.;
- Higher visitation probability of legitimate email nodes than spam nodes;
- Higher probability of being replied to for legitimate nodes when compared with nodes that receive spam and;
- A strong correlation between the size of asymmetry sets and the number of spammers in the set.

On the other hand, concerning graph dynamical properties, we have shown that:

- The spam graph grows much faster, both in number of nodes and edges, than the legitimate email graph, manifesting the higher stability of relations in a social group and;
- The dynamic of the communication between pairs of sender-recipients in spam traffic displays a much higher entropy and a much longer stack distance than legitimate email nodes communication.

In Chapter 6 we propose and evaluate two algorithms to improve spam detection mechanisms based on the structural similarity among contact lists of email users. We have shown that we can successfully group spam and legitimate email users into clusters and use them to improve the quality of spam detection. In one algorithm, we use historical information of the clusters, while in the other, we use historical information of the communication between clusters of users for spam detection. In both cases, we were able to reduce significantly (approximately $\approx 60\%$) the rate of false positives.

Characteristics of the networks of legitimate and spam emails were analyzed in Chapter 7. We have shown that:

- The richness of behaviors in human communication - both symbiotic and opportunistic or antisocial, here represented by spammer behaviors - is present in the structure of networks of email communication and can be quantified via graph theoretical and time series analysis;
- Low network transitivity and preferential exchange are a strong imprint of spam or spammers behavior;
- Although we found power law distribution model for sizes of the extracted communities for all the networks, there are significative differences in the way users can be clustered in communities in the two components. While unitary communities are much more common in the social components, the antisocial, or spammer, communities are on average bigger than social ones.

In summary, we have shown that spam and legitimate user behaviors exhibit strong differences that reflect in several structural and dynamical distinctions. Secondly, we use these knowledge to propose and analyze a new spam detection algorithm based on the structural similarity among contact lists of email users. Finally, we studied structural and dynamical characteristics of networks of spam and legitimate email users in order to uncover the antisocial behavior of spammers.

Possible directions for future work include:

- Verification of our results over distinct time and workloads, in order to be sure that the results are not specific of the workloads analyzed;
- Design and evaluation of new spam detection and filtering techniques that exploit the distinctions between spam and legitimate traffics and, user behavior uncovered in this thesis;
- Design and implementation of a spam-infected email synthetic workload generator to be used in the experimental evaluation of new spam detection strategies;
- Use weighted graphs (e.g., with number of messages as edge weight) in the analysis of social an antisocial network characteristics, as we showed that traffic intensity is one of the main distinctions between antisocial and social behaviors in email networks;
- Characterize other antisocial workloads like virus/worm attacks, in order to try to uncover general signatures of antisocial behavior;
- Study how large volumes of spam in email networks influence the dissemination of virus and other malicious content in email networks, in order to evaluate if the immunization techniques actually applied in email networks continue to be effective in spam-infected email networks.

Bibliography

[1] L. Bertolotti and M. C. Calzarossa, "Workload Characterization of Mail Servers," in *Proceedings of the SPECTS 2000*, Vancouver, Canada, July 2000, pp. 301–307, Elsevier Science Publishers B. V.

[2] L. Bertolotti and M. C. Calzarossa, "Models of Mail Server Workloads," *Performance Evaluation*, vol. 46, no. 2-3, pp. 65–76, 2001.

[3] R. D. Twining, M. M. Willianson, M. Mowbray, and M. Rahmouni, "Email Prioritization: Reducing Delays on Legitimate Mail Caused by Junk Mail," in *Proceedings of the Usenix Annual Technical Conference*, Boston, MA, June 2004, pp. 45–58.

[4] L. F. Cranor and B. A. LaMacchia, "Spam!," *Communications of the ACM*, vol. 41-8, pp. 74–83, August 1998.

[5] J. Jung and E. Sit, "An Empirical Study of Spam Traffic and the Use of DNS Black Lists," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, Taormina, Italy, 2004, pp. 370–375, ACM Press.

[6] P. Oscar Boykin and Vwani Roychowdhury, "Leveraging Social Networks to Fight Spam," *IEEE Computer*, vol. 38, no. 4, pp. 61–68, April 2005.

[7] P. Desikan and J. Srivastava, "Analyzing Network Traffic to Detect E-Mail Spamming Machines," in *Proceedings of the ICDM Workshop on Privacy and Security Aspects of Data Mining*, Brighton UK, November 2004, pp. 67–76.

[8] V. G. Cerf, "Spam, Spim, and Spit," *Commun. ACM*, vol. 48, no. 4, pp. 39–43, 2005.

[9] L. H. Gomes, R. B. Almeida, L. M. A. Bettencourt, V. A. F. Almeida, and J. M. Almeida, "Comparative Graph Theoretical Characterization of Networks of Spam and Regular Email," Second Conference on Email and Anti-Spam, July 2005.

[10] L. H. Gomes, C. Cazita, J. Almeida, V. A. F. Almeida, and W. Meira Jr., "Characterizing a Spam Traffic," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, Taormina, Italy, 2004, pp. 356–369, ACM Press.

[11] L. H. Gomes, F. D. O. Castro, R. B. Almeida, L. M. A. Bettencourt, V. A. F. Almeida, and J. M. Almeida, "Improving Spam Detection Based on Structural Similarity," Steps to Reducing Unwanted Traffic on the Interne Workshop, July 2005.

[12] M. Newman, D. Watts, and S. Strogatz, "Random Graph Models of Social Networks," *Proceedings of the Natl. Acad. Sci, USA*, vol. 99-2566-2572, 2002.

[13] M. E. Newman, S. Forrest S., and J. Balthrop, "Email Networks and the Spread of Computer Viruses," *Physical Review E*, vol. 66-035101(R), pp. 1–4, September 2002.

[14] J. R. Tyler and B. A. Huberman D. M. Wilkinson, "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations," in *Proceedings of the International Communities and Technologies*, September 2003.

[15] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar Community Structure in a Network of Human Iteration," *Physical Revew E*, vol. 68-065103, 2003.

[16] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "The Real Communication Network Behind the Formal Chart: Community Structure in Organizations," *Revista Hispana de Para El Analisis de Redes Sociales*, vol. 6, September 2004.

[17] A. Arenas, L. Danon, A. Díaz-Guilera, P. M. Gleiser, and R. Guimerà, "Community Analysis in Social Networks," *Eur. Phys. J. B.*, vol. 38, pp. 373–380, 2004.

[18] B. Wellman andJ. Salaff andD. Dimitrova andL. Garton andM. Gulia andC. Haythornthwaite, "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community," *Annual Review of Sociology*, vol. 22 213-238, August 1996.

[19] J. G. Kossinets and D. J. Watts, "Empirical Analysis of an Evolving Social Network," *Science*, vol. 311, no. 5757, pp. 88–90, January 2006.

[20] "Spam Haus Project," http://www.spamhaus.org.

[21] W. Gansterer, M. Ilger, P. Lechner, R. Newmayer, and J. Straub, "Anti-Spam Methods - State-of-the-Art," Tech. Rep. FA 384108 - Spam-Abwehr, Institute of Distributed and Multimedia Systems. Faculty of Computer Science. University of Vienna, Austria, March 2005.

[22] "Homel Food Corporation - Home Page," http://www.hormel.com.

[23] P. Hofmann, "Unsolicited Bulk E-mail: Definitions and Problems," Tech. Rep. IMCR-004, Internet Mail Consortium Santa Cruz, CA, October 1997.

[24] "Message Labs Home Page," http://www.messagelabs.co.uk/.

[25] M. Nelson, "Anti-Spam for Business and ISPs: Market Size 2003-2008," Tech. Rep., HP Laboratories Bristol, UK, April 2003.

[26] D. Fallows, "Spam: How It Is Hurting Email and Degrading Life on the Internet," Tech. Rep., Pew and Internet American Life Project 1100 Conncticut Avenue, NW - Suite 710 Washington, D.C. 20036, October 2003.

[27] L. Rainie and D. Fallows, "The Can-Spam Act has Not Helped most Email Users so Far," Tech. Rep., Pew and Internet American Life Project 1100 Conncticut Avenue, NW - Suite 710 Washington, D.C. 20036, March 2004.

[28] "MAPS - Mail Abuse Prevention System Home Page," http://mail-abuse.org/rbl/getoff.html.

[29] E. Harris, "The Next Step in the Spam Control War: Greylisting," http://projects.puremagic.com/greylisting.

[30] H. P. Baker, "Authentication Approaches," in *Proceedings of the 56th IETF Meeting*, San Francisco, California, March 2003.

[31] B. Krishnamurthy, "SHRED: Spam Harassment Reduction via Economic Disincentives," in *Proceedings of the 56th IETF Meeting*, San Francisco, California, March 2003.

[32] H. P. Brandmo, "Solving Spam by Establishing a Platform for Sender Accountability," in *Proceedings of the 56th IETF Meeting*, San Francisco, California, March 2003.

[33] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," Tech. Rep. WS-98-05, AAAI Workshop on Learning for Text Categorization, Madison, Wisconsin, July 1998.

[34] T. Fawcett, "In Vivo Spam Filtering: A Challenge Problem for KDD," *SIGKDD Explor. Newsl.*, vol. 5, no. 2, pp. 140–148, December 2003.

[35] F. Zhou, L. Zhuang, B. Zhao, L. Huang, A. Joseph, and J. Kubiatowicz, "Approximate Object Location and Spam Filtering on Peer-to-Peer Systems," in *Proceedings of the Middleware*, Rio de Janeiro, Brazil, January 2003, pp. 1–20.

[36] S. Atkins, "Size and Cost of the Problem," in *Proceedings of the 56th IETF Meeting*, San Francisco, California, March 2003.

[37] Maria Calzarossa and Giuseppe Serazzi, "Workload characterization: A survey," *Proceedings of the IEEE*, vol. 81, no. 8, pp. 1136–1150, 1993.

[38] M. Arlitt and C. Williamson, "Web Server Workload Characterization: The Search of Invariants," in *Proceedings of the 1996 Sigmetrics Conference on Measurement of Computer Systems*, Philadelphia, PA, May 1996, pp. 126–137, ACM Press.

[39] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin, "A Hierarchical Characterization of a Live Streaming Media Workload," in *Proceedings of the Second ACM SIGCOMM Workshop on Internet Measurement*, Marseille, France, November 2002, pp. 117–130, ACM Press.

[40] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon, "Analysis of Educational Media Server Workloads," in *Proceedings of the 11th Int'l, Workshop on Network and Operating System Support for Digital Audio and Video*, Port Jefferson, New York, June 2001, pp. 21–30, ACM Press.

[41] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto, "Analyzing Client Interativity in Streaming Media," in *Proceedings of the 13th World Wide Web Conference - WWW 2004*, New York City, May 2004, pp. 534–543, ACM Press.

[42] K. P. Gummadi, R. J. Dunn, S. S., Steven D. Gribble, H. M. Levy, and J. Zahorjan, "Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload," in *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, Bolton Landing, N.Y., 2003, pp. 151–160, ACM Press.

[43] C. Dewes, A. Wichmann, and A. Feldmann, "An Analysis of Internet Chat Systems," in *Proceedings of the 2003 ACM SIGCOMM Conference on Internet Measurement*, Miami Beach, FL, 2003, pp. 51–64, ACM Press.

[44] A. Jacobsson and B. Carlsson, "Privacy and Spam: Empirical Studies of Unsolicited Commercial E-Mail," in *Proceedings of the 2nd IFIP Summer School on Risks & Challenges of the Network Society*, Karlstad, Sweden, August 2003.

[45] M. Paganini, "ASK: Active Spam Killer," in *Proceedings of the 2003 Usenix Annual Technical Conference*, San Antonio, Texas, June 2003.

[46] "Spam Blockers Home Page," http://www.spam-blockers.com.

[47] J. B. Postel, "RFC 821: Simple Mail Transfer Protocol," http://www.ietf.org/rfc/rfc821.txt, August 1982.

[48] "Lutz Donnerhacke and IKS GmbH Jena - Teergrubing FAQ," http://www.iks-jena.de/mitarb/lutz/usenet/teergrube.en.html.

[49] Marty Lamb and Martian Software, "TarProxy: Lessons Learned and What's Ahead," in *Proceedings of the 2004 Spam Conference*, Cambridge, MA, January 2004.

[50] W. W. Cohen, "Learning to Classify English Text with ILP Methods," in *Proceedings of the 5th International Workshop on Inductive Logic Programming*, L. De Raedt, Ed., Department of Computer Science, Katholieke Universiteit Leuven, 1995, pp. 3–24, IOS Press.

[51] "SpamAssassin Home Page," http://www.spamassassin.org.

[52] B. Massey, M. Thomure, and R. B. S. Long, "Learning Spam: Simple Techniques for Freely-Available Software," in *Proceedings of the 2003 Usenix Annual Technical Conference*, San Antonio, Texas, June 2003, pp. 63–76, USENIX.

[53] MERL B. Yerazunis, "The Plateau at 99.9% Accuracy, and How to Get Past It," in *Proceedings of the 2004 Spam Conference*, Cambridge, MA, January 2004.

[54] J. D. M. Rennie and T. Jaakkola, "Automatic Feature Induction for Text Classification," www.ai.mit.edu/research/abstracts/abstracts2002/machine-learning/13rennie.pdf, 2003, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

[55] P. Graham, "Better Bayesian Filtering," in *Proceedings of the 2003 Spam Conference*, Cambridge, MA, January 2003.

[56] B. Burton, "SpamProbe - Bayesian Spam Filtering Tweaks," http://spamprobe.sourceforge.net.

[57] M. E. Newman, "Mixing Patterns in Networks," *Physical Review E*, vol. 67-026126, February 2002.

[58] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen, "Hierarchy Measures in Complex Networks," *Physical Revew Letters*, vol. 92-178702, 2005.

[59] F. Wu, B. A. Huberman, L. A. Adamic, and J. R. Tyler, "Information Flow in Social Groups," in *Proceedings of the 23rd CNLS Conference on Networks*, Santa Fe, NM, May 2003.

[60] J. P. Eckmann, E. Moses, and D. Sergi, "Entropy of Dialogs Creates Coerent Structures in E-Mail Traffic," *PNAS - Proceedings of the National Academy of Sciences*, vol. 101-14333, no. 40, pp. 333–337, October 2004.

[61] A.-L. Barabási, "The Origin of Bursts and Heavy Tails in Human Dynamics," *Nature*, vol. 435-207, 2005.

[62] J. Balthrop, S. Forrest, M. E. J. Newman, and M. M. Williamson, "Technological Networks and the Spread of Computer Viruses," *Computer Science*, vol. 304, April 2004.

[63] C. C. Zou, D. Towsley, and W. Gong, "Email Virus Propagation Modeling and Analysis," Tech. Rep. TR-CSE-03-04, Departament of Computer Science. Univ. Massachusetts, Amherst, 2004.

[64] R. Albert, H. Jeong, and A.-L. Barabási, "Error and Attack Tolerance of Complex Networks," *Nature*, vol. 406-378, July 2000.

[65] K. S. Trivedi, *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, John Wiley & Sons, New York, NY, 2001.

[66] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, "Characterization of Complex Networks: A Survey of Measurements," http://www.arxiv.org/abs/cond-mat/0505185, August 2006.

[67] G. K. Zipf, *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge. M.A. U.S.A., 1949.

[68] R. J. Wilson, *Introduction to Graph Theory*, Academic Press, New York, NY, 1979.

[69] G. Schmidt and T. Strohlein, *Relations and Graphs*, Spring-Verlag, New York, NY, 1993.

[70] D. Garlaschelli and M. I. Loffredo, "Patterns of link reciprocity in directed networks," *Phys. Rev. Letters*, vol. 93-268701, 2004.

[71] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, April 1998.

[72] M. Catanzaro, G. Caldarelli, and L. Pietronero, "Assortative Mixing for Social in Networks," *Physical Review E*, vol. 70-037101, September 2004.

[73] M. Girvan and M. E. J. Newman, "Community Structure in Social and Biological Networks," *Proceedings of the Natl. Acad. Sci, USA*, vol. 99-8271-8276, 2002.

[74] D. Menascé and V. Almeida, *Capacity Planning for Web Services: Metrics, Models and Methods*, Prentice Hall Inc., USA, September 2001.

[75] "Exim Internet Mailer Home Page," http://www.exim.org.

[76] "AMaViS - Home Page," http://www.amavis.org.

[77] "Trend Micro Home Page," http://www.trendmicro.com.

[78] "What is a Dictionary Attack?," http://www.filterpoint.com/help/dictionary.html.

[79] V. Paxson and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[80] H. Ebel, L.-I. Mielsch, and S. Bornhold, "Scale-free Topology of E-mail Networks," *Phisical Review E*, vol. 66-035103(R), pp. 1–3, September 2002.

[81] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira, "Characterizing Reference Locality in the WWW," in *Proceedings of the IEEE Conference on Parallel and Distributed Information Systems (PDIS)*, Miami Beach, FL, December 1996, pp. 92–103.

[82] "The List Guy Home Page," http://www.listguy.com/email-sent.html.

[83] L. Cherkasova and G. Ciardo, "Characterizing Temporal Locality and its Impact on Web Server Performance," in *Proceedings of the Int'l Conf. on Computer Communications and Networks*, Las Vegas, NV, October 2000, pp. 16–18, IEEE.

[84] S. N. Dorogvtsev and J. F. F. Mendes, "Evolution of Networks," *Advances in Physics*, vol. 51, no. 66, pp. 1079–1187, 2002.

[85] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley Longman Publishing Co. Inc., 1999.

[86] I. H. Witten, T. C. Bell, and A. Moffat, *Managing Gigabytes: Compressing and Indexing Documents and Images*, John Wiley & Sons, Inc., New York, NY, 1994.

[87] Y. Xie, H. Kim, D. O'Hallaron, M. Reiter, , and H. Zhang, "Seurat: A Pointillist Approach to Anomaly Detection," in *Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection (RAID2004)*, Sophia Antipolis, French Riviera, France, September 2004, pp. 238 – 257, Springer Verlag.

[88] L. adamic and E. Adar, "How to Search a Social Network," electronic arXiv, cond-mat/0310120., 2004.

[89] M. E. Newman, "Assortative Mixing in Network," *Physical Review E*, vol. 89-208701, 2002.

[90] J. M. Casado, T. Garfinkel, W. Cui, V. Paxson, and S. Savage, "Opportunistic Measurement: Extracting Insight from Spurious Traffic," in *Proceedings of the 4th Workshop on Hot Topics in Networks*, Maryland, MD, November 2005.

[91] W. Aiello, C. Kalmanek, P. McDaniel, S. Sen, O. Spatscheck, and J. V. der Merwe, "Analysis of communities of interest in data networks," in *Proceedings of the Passive and Active Measurement Workshop - PAM'05*, March 2005.