DANIEL COUTINHO DE MIRANDA

CARACTERIZAÇÃO DAS REDES DE RELACIONAMENTOS EM SISTEMAS DE COMPARTILHAMENTO DE FOTOS

Belo Horizonte
03 de julho de 2008

DANIEL COUTINHO DE MIRANDA
Orientador: Jussara Marques de Almeida

CARACTERIZAÇÃO DAS REDES DE RELACIONAMENTOS EM SISTEMAS DE COMPARTILHAMENTO DE FOTOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Belo Horizonte
03 de julho de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Caracterização das Redes de Relacionamentos em Sistemas de Compartilhamento de Fotos

DANIEL COUTINHO DE MIRANDA

Dissertação defendida e aprovada pela banca examinadora constituída por:

Doutora Jussara Marques de Almeida – Orientador
Universidade Federal de Minas Gerais

Doutor Virgílio Augusto Fernandes Almeida
Universidade Federal de Minas Gerais

Doutor Marcos André Gonçalves
Universidade Federal de Minas Gerais

Belo Horizonte, 03 de julho de 2008

Agradecimentos

Agradeço primeiramente a Deus, pela oportunidade de desenvolver meus estudos e a força para superar todas as dificuldades.

Agradeço aos meus pais, minha avó e meus irmãos pelo amor, carinho e constante apoio durante todos esses anos.

À professora Jussara Almeida, minha orientadora, pelos tantos ensinamentos, idéias, compreensão e apoio que tornaram possível o desenvolvimento do mestrado.

A todos os professores que contribuíram para que eu chegasse até onde cheguei. Em especial aos professores Virgílio Almeida e Marcos André pelas diversas horas de dedicação, contribuições e orientações em projeto desenvolvido durante o mestrado.

A todos do VoD, Speed e demais colegas do mestrado, pela amizade, experiências e idéias compartilhadas. Aos meus grandes amigos da Sta. Rita, pelo apoio e incentivo.

Por fim, agradeço também ao CNPq pelo suporte financeiro.

Resumo

Com o advento da Web 2.0, uma gama variada de aplicações têm incentivado maior interação entre usuários, acarretando a formação de redes de relacionamentos complexas. As características dessas redes de relacionamentos podem ser melhor entendidas de forma a identificar, por exemplo, como usuários fazem uso do sistema, permitindo assim futuras propostas de melhorias ou novas funcionalidades. Um exemplo destes tipos de aplicação são os sistemas de compartilhamento de fotos, cada vez mais populares graças à facilidade de criação e armazenamento de fotos e imagens digitais, reforçando a necessidade de uma maior investigação desses sistemas.

Esta dissertação apresenta uma caracterização de dois tipos de redes de relacionamentos extraídas do Flickr, um sistema Web 2.0 muito popular direcionado ao armazenamento, organização, busca e compartilhamento de fotos. A primeira rede consiste de usuários do Flickr interligados por meio de relações de contato ou amizade entre eles, enquanto a segunda inclui usuários do Flickr interligados por meio de relações de testemunho. A caracterização das redes foi realizada em duas vertentes. A primeira vertente engloba a análise de características individuais dos usuários presentes em cada uma dessas redes, visando identificar e contrastar o padrão de uso do sistema visto em cada rede. Já a segunda vertente engloba a análise social de cada uma das redes e tem como objetivo entender as interações existentes entre os usuários em cada uma delas.

Nossos resultados indicam que usuários que fazem uso da funcionalidade de testemunhos são, em geral, usuários mais ativos e com maior compromisso com a utilização do sistema. Além disso, um estudo de tráfego das fotos dos usuários presentes nas duas redes mostrou, entre outras coisas, que a popularidade das fotos dos usuários é fortemente correlacionada com o número de testemunhos recebidos pelos usuário e com o número de vezes que o usuário foi adicionado em listas de contatos. Finalmente, verificou-se que, enquanto a rede de contatos possui fortes características sociais, a rede de testemunhos é menos social e provavelmente reflete interações mais relacionadas ao interesse de um usuário pelo conteúdo de outros. Estes resultados se manifestaram estáveis em três coletas diferentes, realizadas em um intervalo de seis meses.

Abstract

With the increase of the popularity of the Web 2.0, several applications have encouraged more interaction between users, resulting in the formation of complex networks of relationships. The characteristics of these networks can be better understood in order to identify, for example, how users make use of the system, allowing future improvement and the proposal of new features. Example of this kind of application are photo sharing systems, which are becoming more popular due to the easiness of creating and storing photos and digital images, reinforcing the need of a more deep investigation of these systems.

This dissertation presents a characterization of two types of networks of interactions between users in Flickr, a very popular Web 2.0 system to store, organize, search and share photos. The first network consists of Flickr's users interconnected by contact or friendship relations while the second network includes users interconnected by the testimonial relations. The networks were characterized in two different levels. The first level is related to the individual characteristics of the users existing in each network, aiming to identify and contrast how the system is used by the users in each network. The second level consists of the analysis of social characteristics of each network and has the goal of providing a more solid understanding of the existing interactions among the users of the networks.

Our results show that users who use the testimonial feature are in general more active and more committed to using the system. Moreover, an analysis of the traffic of the photos owned by users present in both networks pointed out, among other results, that the popularity of a user photos is strongly correlated with the number of testimonials received by users and with the number of times that the user was added into contact lists. Finally, we verified that, while the contact network has strong social characteristics, the testimonial network is less social and probably reflects interactions that are based on the interest of a user by the content of others. These results were consistent across three different data sets, collected over a six month period.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Sistemas de compartilhamento de fotos têm sido cada vez mais utilizados por fotógrafos profissionais e amadores. Esses sistemas administram um número muito grande de usuários e mantêm coleções gigantescas de dados e fotos. Muitos desses sistemas permitem que usuários se relacionem formando redes sociais de afinidades, que podem ser melhor entendidas e exploradas a fim de auxiliar a criação de algoritmos mais interessantes para, por exemplo, permitir melhor navegação e descoberta de conteúdo, além de soluções que possam tornar esse tipo de sistema cada vez mais escalável e eficiente.

1.1 Motivação

Web 2.0 é uma evolução da WWW (*World Wide Web*) que tem ganhado muita atenção recentemente [14]. Aplicações da Web 2.0 têm como uma das principais características oferecer maior integração e participação do usuário. Diferentemente do que acontecia em sistemas Web tradicionais (também denominados Web 1.0), onde usuários agiam mais como consumidores passivos da informação, esses novos sistemas fornecem um ferramental que permite que usuários facilmente sejam também criadores e publicadores de informações em diferentes mídias (ex: texto, imagem e vídeo). Essa abordagem mais centrada no usuário acaba por incentivar a formação de redes sociais e comunidades virtuais nesses tipos de sistema [14, 21]. Diversas aplicações Web 2.0 podem ser encontradas na *Internet*, incluindo sistemas de relacionamento tais como Facebook [4] e Orkut [7], sistemas de compartilhamento de mídia que podem ser ilustrados pelo Flickr [5] e YouTube [9], sistemas de *bookmarking* social como del.icio.us [2], entre outros.

Em especial, uma categoria de aplicação da Web 2.0 é composta pelos sistemas de gerenciamento e compartilhamento de fotos e imagens, tais como o Flickr[1] e deviantART [3], que têm sido cada vez mais difundidos e utilizados devido à atual facilidade e relativo baixo custo da produção de imagens e fotos digitais. Segundo [1], esses sistemas são responsáveis por gerar grande volume de tráfego na *Internet* (em 25/04/08 o Flickr constava entre os top-50

[1]Durante a escrita da dissertação, o Flickr passou a aceitar também pequenos vídeos (maiores informações em http://blog.flickr.net/en/2008/04/09/video-on-flickr-2). No entanto, serão consideradas apenas as características relacionadas ao gerenciamento e compartilhamento de fotos.

no Brasil e no mundo). Ao permitir que usuários se comuniquem, troquem experiências e criem redes sociais de afinidades, sistemas de compartilhamento de fotos incentivam maior contribuição da comunidade em relação aos objetos (ex: trabalhos artísticos, imagens de domínio científico, técnicas de fotografia, etc), além de possibilitar melhor divulgação de imagens tanto de interesse geral quanto de interesse de algum grupo de pessoas específico (ex: pessoas interessadas em imagens em preto e branco ou em alguma técnica específica de fotografia).

Tais sistemas de compartilhamento de fotos devem armazenar imagens de um número gigantesco de usuários, provendo serviços como busca e compartilhamento de fotos de forma eficiente e inteligente. No entanto, conforme apontado por [25], a magnitude das coleções gerenciadas por tais sistemas acaba por trazer novos desafios, demandando maiores investigações. Além disso, um estudo mais aprofundado das características de sistemas de redes sociais é de significativa importância para se avaliar ou aprimorar os atuais sistemas, projetar novos sistemas baseados em redes sociais, entender o impacto das redes sociais sobre a *Internet* ou até mesmo entender melhor as relações existentes nas redes sociais humanas [34, 11]. De fato, o estudo de redes sociais tem sido muito relevante em diversas áreas da Ciência da Computação. Alguns exemplos de aplicações em outras áreas incluem busca por conteúdo na Web [32], minimização do impacto de ação maliciosa em redes Par-a-Par [43] e propostas de melhorias para sistemas móveis [31].

Neste trabalho será descrita uma análise baseada em redes sociais extraídas do Flickr, um sistema Web popular direcionado ao armazenamento, organização, busca e compartilhamento de fotos. O sistema, que iniciou as atividades em fevereiro de 2004 [26], é muito utilizado por fotógrafos profissionais e amadores, mas também largamente utilizado profissionalmente ou para uso pessoal pelos mais diferentes tipos de usuários. Mais especificamente, acredita-se que o Flickr é um bom candidato à caracterização de redes sociais em sistemas de compartilhamento de fotos devido às diferentes redes sociais que se consegue extrair do uso do sistema e pela sua magnitude e tráfego – em 2007 apresentando mais de 8,5 milhões de usuários registrados, gerando um tráfego de até 12.000 fotos sendo transferidas por segundo e o maior número de *uploads* de fotos por dia já registrado sendo superior a 2 milhões [42]. Exemplos de tipos de interação social relevantes encontrados no Flickr são possibilitar que usuários registrados adicionem outros à suas listas de contatos (ex: amigos, familiares, contatos profissionais, etc), permitir que usuários escrevam testemunhos (ex: texto elogiando tipo de fotografia ou conteúdo mantido pelo usuário) à outros e criem grupos de interesse para discussão e compartilhamento de fotos.

1.2 Objetivos

O objetivo geral deste trabalho é realizar uma caracterização de redes de relacionamentos existentes no Flickr, utilizando dados obtidos a partir de coletas de informações publicamente disponibilizadas no sistema.

Os objetivos específicos são:

- Caracterizar aspectos sociais de duas redes de relacionamentos específicas extraídas do

Flickr: a rede de testemunhos e a de contatos. Objetiva-se identificar tanto aspectos comuns quanto aspectos que distinguem estes dois tipos de interações entre usuários de um mesmo sistema (Flickr). Objetiva-se também contrastar os resultados obtidos com caracterizações anteriores de redes de testemunhos e redes de contatos de diferentes aplicações da Web 2.0.

- Analisar características individuais dos usuários existentes em cada uma das redes de relacionamentos, visando identificar padrões de comportamento diferentes.

- Em última instância, esta dissertação visa prover informações sobre como os usuários de um sistema de grande popularidade, o Flickr, utiliza suas funcionalidades. Acreditamos que tais informações possam ser úteis para projetistas de aplicações deste tipo, bem como para a elaboração de modelos de redes sociais.

1.3 Contribuições

Os principais resultados da caracterização realizada, que constituem as principais contribuições desta dissertação são:

- Há uma maior concentração dos números de fotos, contatos e grupos em alguns poucos usuários, representada por distribuições Zipf de cauda pesada. Isso indica que alguns poucos usuários fazem muito mais uso das funcionalidades do sistema.

- Os usuários que utilizam testemunhos tendem a ser mais ativos no sistema, lidando com um número maior de contatos, contribuindo mais com o conteúdo público do sistema e participando de um maior número de grupos de interesse.

- A popularidade das fotos dos usuários não está necessariamente relacionada ao número de fotos que ele mantém. Uma possível explicação para isso é que a popularidade das fotos de um usuário está mais relacionada à qualidade do conteúdo mantido por ele.

- A popularidade do conteúdo compartilhado por um usuário não depende significativamente do número de testemunhos por ele enviados ou do tamanho de sua lista de contatos. Essa popularidade depende mais da reação dos outros usuários, expressa por meio do número de vezes que um usuário é adicionado em listas de contatos e do número de testemunhos recebidos.

- A rede de contatos apresenta fortes características sociais, tais como alto coeficiente médio de agrupamento e assortatividade positiva. Já a rede de testemunhos apresenta características sociais mais fracas, sendo provavelmente mais orientada por interesse de usuários pelo conteúdo compartilhado pelos outros. Essa natureza menos social da rede de testemunhos do Flickr era de certa forma inesperada, uma vez que se poderia inicialmente imaginar o testemunho como refletindo uma relação mais forte de amizade entre duas pessoas. Observou-se ainda que o maior componente fortemente conectado

das duas redes formam comunidades mais recíprocas e ligeiramente mais coesas do que as respectivas redes completas.

- Apesar do crescimento do Flickr, refletido em termos de número de usuários, interações entre os usuários, números médios de fotos, contatos e grupos, as características sociais das duas redes se mantêm razoavelmente estáveis ao longo de seis meses.

1.4 Organização do Texto

A dissertação está organizada da seguinte forma. No Capítulo 2 será apresentada uma visão geral sobre o Flickr, além de discutir trabalhos relacionados a teoria de redes sociais e caracterização desse tipo de rede em aplicações da *Internet*. O Capítulo 3 apresenta a metodologia utilizada para a caracterização das redes de relacionamento de sistemas de gerenciamento e compartilhamento de fotos, descrevendo o processo de coleta e limpeza dos dados, além das métricas escolhidas para caracterização das redes. O Capítulo 4 apresenta os resultados da caracterização realizada. Finalmente o Capítulo 5 oferece conclusões e indicações de linhas de pesquisa para futuros trabalhos.

Capítulo 2

Contextualização

Neste capítulo serão apresentadas informações para contextualizar o leitor e assim proporcionar um melhor entendimento do presente trabalho. O capítulo está estruturado da seguinte forma: na seção 2.1 será apresentada uma visão geral do Flickr e suas funcionalidades mais relevantes. Na seção 2.2 serão apresentados alguns fundamentos relacionados ao estudo e à análise de redes sociais. A seção 2.3 aborda trabalhos relacionados à caracterizações de redes sociais na *Internet*, enquanto a seção 2.4 aborda trabalhos relacionados a outras análises sociais desenvolvidas em relação ao Flickr.

2.1 O Sistema Flickr

O Flickr é um sistema de gerenciamento e compartilhamento de fotos *online*, que permite aos usuários armazenarem e organizarem grandes coleções de fotos e imagens, além de fornecer funcionalidades para, por exemplo, realizar buscas e disponibilizar conteúdo para toda a rede ou somente para um grupo de pessoas. O Flickr permite também diferentes tipos de interação social entre os usuários. Nesta seção serão descritas as funcionalidades mais relevantes do sistema afim de proporcionar um melhor entendimento do presente trabalho.

Um aspecto primário nesse tipo de sistema é a gerência de fotos. Usuários do Flickr administram fotos que podem ser públicas ou privadas, podendo definir o tipo de licença e adicionar metadados tais como título, descrição, rótulos (*tags*), data de criação e localização geográfica do local onde foi tirada a foto. Caso o usuário deseje, poderá configurar o sistema para permitir que outros usuários adicionem comentários, além de outros rótulos ou notas às suas fotos. O sistema armazena ainda, para cada foto, estatísticas de acesso como, por exemplo, o número de visualizações que uma foto recebeu até o momento corrente. Usuários podem adicionar fotos de outros usuários em uma lista de favoritos, facilitando encontrá-las futuramente.

A figura 2.1 ilustra uma página contendo as informações sobre uma foto no sistema.

O Flickr permite também que cada usuário mantenha uma lista de contatos[1] (ex: amigos,

[1] O Flickr impõe um limite de 3.000 contatos não-recíprocos (i.e., caso em que o usuário adicionou um outro em sua lista de contatos mas não foi adicionado por ele) para a lista de contatos de um usuário, não afetando a quantidade de usuários que adicionam o mesmo em suas listas de contatos [6]. Esse limite foi imposto

Figura 2.1: Exemplo de uma foto no Flickr.

familiares e contatos profissionais) que pode ser usada, por exemplo, para restringir o compartilhamento de fotos a um subgrupo das pessoas em sua lista de contatos. Após adicionar um contato à lista de contatos, o usuário poderá visualizar as fotos mais recentemente postadas pelo contato, o que faz com que essa funcionalidade possa ser usada também como uma forma de criar um atalho para um usuário com conteúdo de interesse.

Uma outra forma de interação encontrada em alguns sistemas baseados em redes sociais é permitir que usuários criem, mantenham e se afiliem a grupos de interesse. Particularmente no Flickr, esses grupos permitem que usuários possam compartilhar fotos, assim como realizar discussões acerca da temática do grupo. Os grupos no Flickr podem ser privados ou públicos, onde grupos públicos podem ser completamente abertos ou baseados em convites. Exemplos de temas de grupos incluem “fotos ou imagens abstratas”, “imagens noturnas”, “imagens em preto e branco”, imagens relativas a algum monumento específico ou sobre algum tipo de técnica de fotografia, entre outros.

A figura 2.2 apresenta um exemplo de grupo público no Flickr.

A cada usuário do Flickr é atribuído um página de perfil, que pode ser editada para que o usuário possa adicionar informações sobre si, tais como contatos, estado civil, preferências,

visando melhor desempenho do sistema, além de tentar minimizar problemas de *spam* (maiores informações em: http://blog.flickr.net/en/2007/01).

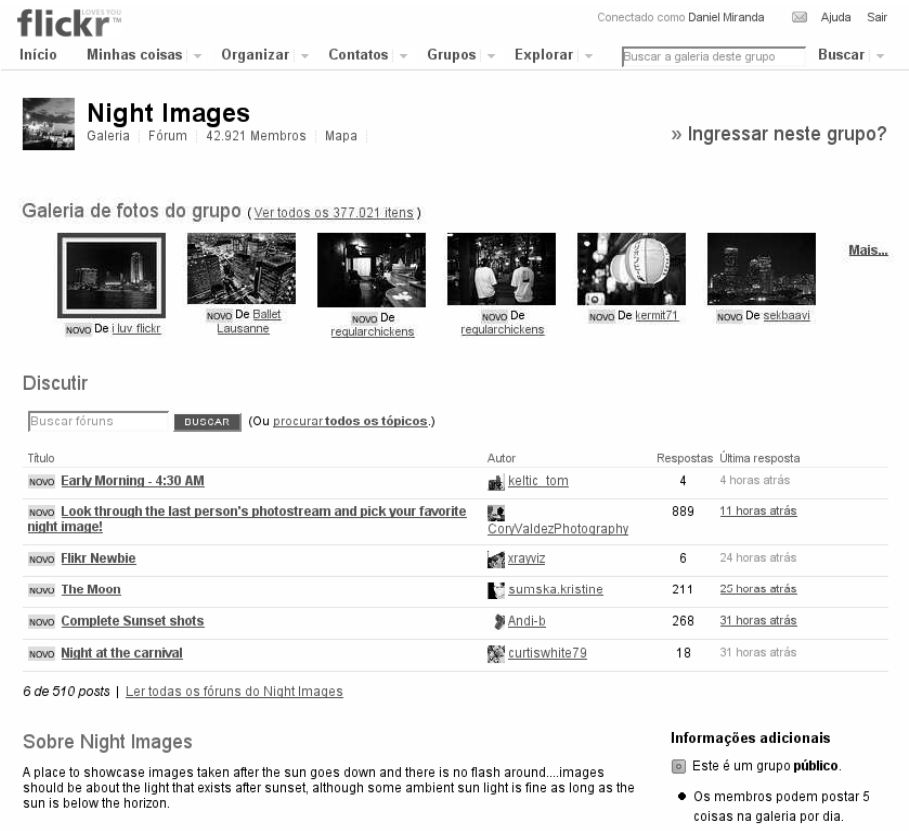


Figura 2.2: Exemplo de um grupo no Flickr.

localização geográfica, entre outras. A página de perfil resume também algumas informações sobre a utilização do sistema pelo usuário, mostrando, por exemplo, os grupos públicos em que o usuário está afiliado e o número de contatos existentes em sua lista de contatos.

Uma outra forma interessante de interação social presente no Flickr é o envio de testemunhos a outros usuários. Um usuário pode escrever testemunhos sobre outros, expressando por exemplo, desde elogios ou recomendações ao conteúdo (i.e., fotografias) mantido pelo usuário até um simples comentário acerca da personalidade daquele usuário. Para um dado usuário receber o testemunho de outro, o usuário deve ter o remetente em sua lista de contatos, embora o remetente não precise ter o recipiente em sua lista de contatos. Um usuário não pode enviar testemunhos para si, nem mesmo enviar mais de um testemunho para outro usuário. Caso um testemunho seja aceito pelo destinatário, o testemunho aparecerá na página de perfil do mesmo, podendo ser lido por qualquer outro usuário do sistema. Um testemunho pode também ser reeditado pelo remetente[2] ou apagado tanto pelo remetente quanto pelo usuário que o recebeu. Diferentemente da interação por comentários no Flickr, que está relacionada a fotos específicas de um usuário, o testemunho é geralmente relacionado a informações gerais sobre o conteúdo ou personalidade do usuário.

A figura 2.3 exemplifica uma página de perfil de um usuário que recebeu dois testemunhos,

[2]O destinatário deve reaceitar o testemunho editado.

ambos acerca da qualidade do conteúdo mantido pelo usuário.

Figura 2.3: Exemplo de uma página de perfil no Flickr.

Como forma de auxiliar os usuários a encontrarem fotos interessantes no sistema, o Flickr elege diariamente, por meio de um algoritmo proprietário, 500 fotos tidas como as "mais interessantes" do dia. O sistema permite também que, ao realizar busca por fotos, o usuário escolha que o resultado seja ordenado pelas fotos mais interessantes.

Para permitir que programadores desenvolvam aplicações que possam interagir com o Flickr, por exemplo, apresentando uma interface diferente para o usuário utilizar o sistema, o Flickr fornece uma API[3] (*Application Programming Interface*). Essa interface com o sistema pode ser utilizada para coleta de dados públicos.

Finalmente, o Flickr distingue usuários entre pagantes (conta "pro") e não pagantes. Os usuários não pagantes sofrem restrições em relação a, por exemplo, limite do número de fotos, quota mensal de *upload* de fotos, assim como outras restrições de funcionalidade.

2.2 Análise de Redes Sociais

A seguir será apresentado um breve resumo da teoria da análise de redes sociais, discutindo os fundamentos principais para se entender as métricas utilizadas para a caracterização das

[3] http://www.flickr.com/services/api

redes analisadas no presente trabalho. As definições formais das métricas serão apresentadas na seção 3.3.

Recentemente, diversos trabalhos na literatura desenvolveram estudos focando em entender melhor os diferentes tipos de relacionamento entre pessoas ou objetos existentes no mundo real. Esses relacionamentos muitas vezes são analisados como redes complexas, modeladas como grandes grafos (direcionados ou não) onde vértices são entidades (ex: pessoas, computadores) e arestas representam algum tipo de interação entre as entidades (ex: amizade entre pessoas, comunicação de dados). Exemplos de redes analisadas na literatura incluem rede de co-autores de artigos científicos [39], redes representando a estrutura da *Internet* em diversos níveis [19], estrutura das páginas da Web [12], rede representando comunicação por e-mails [23], entre diversas outras. Durante este trabalho será focado principalmente nas redes sociais, que são aquelas envolvendo interações entre pessoas ou grupos de pessoas.

Segundo [40], a mais básica caracterização topológica de um grafo é dado pela distribuição de graus dos vértices. Entende-se como grau de um vértice em um grafo não direcionado o número de arestas incidentes a ele. Já no caso de um grafo direcionado, distingue-se as arestas que saem do vértice (ou arestas de saída) das arestas que chegam ao vértice (arestas de entrada). De acordo com [13], várias redes complexas apresentam distribuições dos graus dos vértices do tipo Lei de Potência (*power law*). Distribuições Lei de Potência, de acordo com [22], geralmente são representadas como $P(X > x) \approx Cx^{-\alpha}$, onde $C \neq 0$ é uma constante e α o parâmetro da distribuição. Uma distribuição que segue Lei de Potência indica a existência de alguns poucos vértices com alto grau e muitos com grau baixo. Essa desigualdade na distribuição das arestas na rede tem sido muitas vezes atribuída a um fenômeno conhecido na literatura de redes complexas como *preferential attachment*, que indica que durante a criação ou evolução da rede, novas arestas tendem a preferencialmente se conectar a vértices que já participam de muitas arestas [16, 13]. Uma distribuição que segue a Lei de Potência é a distribuição Zipf [44], utilizada em trabalhos anteriores [22] para modelar a distribuição dos graus dos vértices em redes sociais.

Outro aspecto importante de se verificar é que em muitas redes reais, além das novas arestas serem influenciadas pelo grau de entrada do vértice que é apontado pela aresta (explicado pelo *preferential attachment*), algumas vezes é influenciada também pelo grau de saída do vértice de origem, o que pode ser identificado pela análise da assortatividade da rede [36]. Uma assortatividade de graus positiva indica que vértices de alto grau preferencialmente se ligam a outros vértices de alto grau, enquanto vértices de baixo grau se ligam a outros de baixo grau. A assortatividade positiva da rede é um fenômeno comumente denominado na literatura como *assortative mixing*. Já uma assortatividade negativa indica que vértices de baixo grau tendem a se ligar a vértices de alto grau, e vice-versa, fenômeno geralmente denominado na literatura como *disassortative mixing*.

Como pode ser visto em [38], dois aspectos típicos de redes sociais são apresentar alto coeficiente de agrupamento da rede, em relação ao grafo aleatório de mesmo tamanho, além de uma assortatividade positiva. O coeficiente de agrupamento de um vértice mede a probabilidade dos vizinhos do vértice estarem também conectados entre si por meio de arestas,

enquanto o coeficiente de agrupamento da rede é a média dos coeficientes de agrupamento de todos os vértices, refletindo portanto o nível de coesão da rede. Redes sociais mais coesas podem ser vistas como comunidades mais fechadas.

Uma outra análise relevante que se pode fazer acerca de uma rede direcionada diz respeito à reciprocidade, que está relacionada à probabilidade de dois vértices estarem ligados por arestas de direções opostas [37]. Em [20], ilustra-se diversas implicações de uma alta ou baixa reciprocidade da rede tal como, por exemplo, maior ou menor velocidade na difusão de informações.

Segundo [40], é também importante ainda analisar a alcançabilidade dos vértices das redes complexas, que é a possibilidade de partindo-se de um vértice alcançar outro por meio das arestas existentes no grafo. Os componentes conectados do grafo definem muitas propriedades da estrutura física da rede e têm importantes conseqüências como, por exemplo, na acessibilidade de informações em redes tais como a WWW (*World Wide Web*) [40]. Os componentes conectados de uma rede social podem representar indiretamente comunidades fechadas.

Muitos desses fundamentos apresentados nesta seção definem algumas das métricas utilizadas para análise de redes sociais. Uma descrição formal das métricas será dada na seção 3.3. Durante o trabalho, referiremos como uma rede mais social ou com forte fator social aquela que apresenta características geralmente encontradas nas redes sociais, tais como alto coeficiente de agrupamento e assortatividade positiva.

2.3 Caracterização de Redes Sociais em Aplicações da Internet

Diversos trabalhos realizaram caracterização de diferentes tipos de redes na *Internet*. Trabalhos como [18, 23] estudaram redes sociais formadas a partir da comunicação por e-mails entre usuários. O primeiro caracterizou a estrutura da rede quanto a características tais como distribuição de graus, coeficiente de agrupamento e menor caminho entre vértices. A partir dessa caracterização os autores discutem, por exemplo, a difusão de vírus considerando o tipo de topologia encontrada. O segundo teve como objetivo identificar propriedades que distinguissem tráfego legítimo de tráfego acarretado por *spam*. Verificou-se, por exemplo, que existem diversas diferenças estruturais entre as redes que representam os e-mails legítimo e aquelas que representam *spam*. O trabalho de [12] analisa a rede onde vértices são páginas da Web e as URLs são as arestas. Os autores verificaram que a Web é caracterizada por uma distribuição de graus seguindo Lei de Potência. Além disso, verificou-se no gráfico um diâmetro significativamente baixo, indicando que grande parte das informações pode ser encontrada a poucas páginas de distância quando seguindo o fluxo das URLs.

Em especial, o advento da Web 2.0 acarretou na criação de diversos sistemas baseados em redes sociais, trazendo grandes oportunidades para se entender como as pessoas interagem entre si. Encontrou-se na literatura sobre Web 2.0 caracterização de diversos tipos de redes sociais como, por exemplo, rede de contatos ou amizade de diferentes sistemas [34], rede de testemunhos do Cyworld [11] e rede de subscrição para vídeos de outros usuários do You-

Tube [29]. A seguir serão apresentados os trabalhos que abordam as redes mais relevantes para a pesquisa, incluindo trabalhos que analisaram o Flickr e outros grandes sistemas Web 2.0.

A denominada rede de contatos ou rede de amizades é particularmente bem conhecida e vastamente estudada em diferentes sistemas baseados em redes sociais da Web 2.0. A seguir serão apresentados alguns desses trabalhos.

No trabalho de [26], os autores estudaram segundo uma abordagem evolutiva a rede social de contatos dos sistemas Flickr e Yahoo! 360 por meio de dados proprietários dos sistemas onde constava informação temporal. O foco principal desse trabalho consistiu em entender a estrutura e evolução dessas redes analisando os diferentes componentes dos grafos, tais como componentes de tamanho 1, componente gigante e componentes de tamanhos intermediários. Os autores estudaram por exemplo o tipo de estrutura que era encontrada em cada tipo de componente, tamanho dos componentes ao longo do tempo e como componentes se uniam para gerar componentes maiores. Em grande parte das análises os autores utilizaram as versões não-direcionadas das redes. Com as observações realizadas acerca da estrutura da rede de contatos para esses dois sistemas, os autores propuseram um modelo para explicar os principais aspectos da evolução desse tipo de rede, validando por meio de simulação.

Um outro trabalho relevante é o de [34], que analisou as propriedades estruturais de redes sociais de grandes sistemas, tais como Orkut, YouTube, Flickr e LiveJournal, obtidas por meio de coletas de dados públicos. Os autores verificaram que mesmo analisando sistemas de diferentes tipos e tamanhos, algumas características estruturais são comuns a todos eles. Entre estas, citamos a distribuição de graus seguindo uma Lei de Potência e com alta correlação entre grau de entrada e saída. Além disso, a estrutura das redes é aparentemente composta por conjuntos de vértices de relativamente baixo grau mas muito conectados entre si. Esses conjuntos bem conectados são geralmente interligados entre si por meio de alguns poucos vértices que possuem grande número de conexões. Isso implica em vértices com coeficiente de agrupamento inversamente proporcional ao seu grau.

Em [33], abordou-se uma análise evolutiva da rede de contatos do Flickr utilizando dados obtidos por coletas diárias durante um determinado período de tempo, tentando entender como novas relações de contato acontecem. Os autores verificaram, por exemplo, que relações recíprocas tendem a acontecer pouco tempo após a criação da relação inicial entre os dois usuários. Outro resultado relevante foi constatar que apesar da evolução da rede seguir o fenômeno conhecido na literatura de redes sociais como *preferential attachment*, modelos teóricos como o de Barabási e Albert [13], conhecidos por apresentar esse tipo de fenômeno, podem não ser suficientes para explicar a estrutura observada na rede.

O presente trabalho também analisa a rede de contatos do Flickr mas, diferentemente dos demais trabalhos que estudaram a rede de contatos do Flickr, realizou-se uma caracterização de diferentes características individuais dos usuários dessa rede. De forma complementar, analisou-se também aspectos sociais com o objetivo de contrastar com outra rede de relacionamento, criada a partir de outro tipo de interação social no Flickr.

Uma segunda forma de interação social interessante e presente em diversos sistemas de

relacionamento social é o envio de testemunhos. Apesar disso, a caracterização da rede de testemunhos em sistemas Web 2.0 é um tanto escassa. Até onde sabemos, apenas o trabalho de [11] aborda esse tipo de análise e será descrito a seguir.

Em [11], os autores analisam a coleta da rede de testemunhos do Cyworld, acreditando que o testemunho no sistema pode ser visto como uma relação mais forte do que a de amizade (expressa pela rede de contatos) e por isso se aproximando mais das redes sociais encontradas na vida real. Apesar da rede de testemunhos ser direcionada, para a maior parte das métricas os autores analisaram uma versão não-direcionada da mesma. Segundo os autores, a rede de testemunhos do Cyworld é um subconjunto da rede de contatos, possuindo entretanto características distintas daquela, tais como distribuição de graus exponencial ao invés de Lei de Potência, tanto para arestas de entrada quanto para arestas de saída, além de características como *assortative mixing* e maior coeficiente de agrupamento.

Nosso trabalho apresentará uma caracterização a rede de testemunhos do Flickr. No entanto, diferentemente do sistema Cyworld, que é um sistema mais focado no relacionamento social, o Flickr é um sistema especializado no armazenamento e compartilhamento de fotos. Por esse motivo, acredita-se que a rede de testemunhos do Flickr possa apresentar características distintas daquela analisada por [11].

2.4 Outras Análises do Flickr

Diferentemente dos trabalhos anteriormente descritos, onde focava-se de forma mais geral no estudo na estrutura de redes sociais em sistemas Web 2.0 sem abordar com maiores detalhes características específicas do tipo de sistema, alguns trabalhos na literatura estudaram aspectos sociais do Flickr com o intuito de propor melhorias ou entender melhor o comportamento de usuários ao utilizar um sistema de compartilhamento de fotos. Os mais relevantes serão descritos a seguir.

Alguns trabalhos como os de [28, 42, 15], que realizam uma análise mais relacionada ao acesso às fotos do usuário, mostram que navegação no sistema por parte dos usuários é fortemente influenciada por fatores sociais. Verificou-se que a popularidade de fotos está relacionada a fatores tais como tamanho da rede de contatos do dono da foto, número de pessoas que adicionam a foto à lista de favoritos ou o número de grupos em que a foto foi adicionada. O trabalho de [42] analisa além dessa dimensão social, as dimensões temporal e geográfica dos acessos às fotos, entendendo aspectos como descoberta de novas fotos, distribuição de acessos ao longo do tempo e relação entre a popularidade das fotos e a abrangência geográfica dos usuários que as acessaram. Já em [15], estuda-se a difusão de informação por meio de relações sociais no Flickr. Os autores utilizaram um modelo epidemiológico para caracterizar o processo de difusão de algumas fotos populares. Mais especificamente, modelou-se as fotos como sendo uma doença a ser disseminada no ambiente e verificou-se, por exemplo, que a taxa com que usuários encontram fotos de seus contatos e adicionam em suas listas de fotos favoritas, ficando assim infectados, é mais alta do se conhece para doenças muito contagiosas.

O trabalho de [41] estuda os grupos do Flickr e propõe uma metodologia para classificá-los

quanto aos aspectos social e temático. Os autores analisaram um pequeno conjunto de grupos do Flickr abrangendo 450 grupos com cerca de 500 usuários cada e verificaram que os grupos são muito distintos quanto a esses aspectos.

Em [25], os autores mostram como a utilização de rótulos (*tags*) inseridos pelos usuários, informações geográficas (por meio de fotos *geo-referenciadas*) e análise do conteúdo das imagens podem ser usados para, por exemplo, gerar conhecimento adicional acerca dos dados das coleções ou mesmo aprimorar o acesso a esse tipo de conteúdo multimídia. Em especial, é apresentado um mecanismo para recuperação de um tipo específico de imagens que apresentou bons resultados em relação à métrica precisão (métrica muito utilizada para medir a qualidade do resultado de uma busca).

O trabalho de [30], que propõe um modelo e taxonomia para sistemas que utilizam rótulos (*tags*), apresenta uma caracterização do uso de rótulos no Flickr, considerando também aspectos como tamanho da coleção de fotos do usuário e o número de contatos que o usuário possui. Por esse trabalho pode-se verificar, por exemplo, que a atividade de rotular fotos é influenciada pela rede social do usuário.

Finalmente, em [17] desenvolveu-se uma análise do uso do Flickr abrangendo um subconjunto de usuários de cinco diferentes culturas verificando, por exemplo, parâmetros que diferem na utilização do sistema devido a fatores culturais.

Capítulo 3

Identificação das Redes, Processo de Coleta e Métricas de Caracterização

Este trabalho foca na caracterização de redes de relacionamento em sistemas de gerenciamento e compartilhamento de fotos, utilizando como estudo de caso o Flickr.

A estapa inicial do estudo aqui abordado consiste em identificar e selecionar algumas das redes de relacionamento relevantes para serem caracterizadas. Uma vez feito isso, idealmente, para o estudo dessas redes, deveria-se ter à disposição dados proprietários do sistema. No entanto, na falta de tais dados, optou-se por coletar os dados públicos disponibilizados no Flickr. As redes selecionadas para caracterização e os dados selecionados para coleta serão apresentados na seção 3.1.

Para realizar a coleta dos dados, desenvolveu-se um coletor distribuído que utiliza um mecanismo bem conhecido na literatura, denominado *Snowball* [27, 11], para determinar o fluxo de coleta. Maiores detalhes sobre o processo de coleta de dados, assim como sobre a limpeza de dados realizada objetivando minimizar ruídos podem ser encontrados na seção 3.2. As métricas selecionadas para a caracterização das redes serão descritas na seção 3.3.

3.1 Dados Coletados

Coletou-se do Flickr tanto dados relacionados ao relacionamento entre usuários (características sociais) quanto dados individuais acerca dos usuários (características individuais). A descrição dos dados relacionados às características sociais dos usuários, assim como a descrição das redes que serão geradas a partir desses dados serão apresentados na seção 3.1.1. Na seção 3.1.2 serão apresentados os dados que refletem as características individuais dos usuários.

3.1.1 Redes de Relacionamentos

Um tipo de rede de relacionamento muito estudado na literatura de redes sociais e que se pode encontrar em diversos sistemas Web 2.0 é a rede de contatos, também denominada rede de amizades, criada a partir da relação de contatos entre usuários. O estudo dessa rede é importante por permitir entender como usuários se organizam e relacionam ao utilizarem o

sistema. A rede de contatos pode ser modelada por um grafo direcionado, onde vértices são os usuários e arestas interligam um usuário a cada um de sua lista de contatos.

As informações acerca das listas de contatos dos usuários, necessárias para se analisar a rede de contatos, podem ser coletadas utilizando-se a API disponibilizada pelo Flickr. Entretanto, é possível que, por privacidade, um usuário configure seu perfil de forma que outros não o visualizem quando realizarem buscas por pessoas ou recuperarem alguma lista de contato no qual o usuário estaria incluído. Logo, esse usuário não aparecerá nas listas de contatos coletadas.

Outro tipo de interação social considerado em nosso trabalho diz respeito ao envio e recebimento de testemunhos pelos usuários. Decidiu-se analisar a rede gerada a partir dos testemunhos recebidos pelos usuários como forma de tentar entender esse tipo de relacionamento ainda pouco estudado na literatura mas presente em diversos sistemas da Web 2.0. De forma análoga à rede de contatos, a rede de testemunhos pode ser representada por um grafo direcionado, onde os vértices são os usuários da rede e arestas interligam um usuário que enviou testemunhos àqueles que receberam.

Como a API do Flickr não fornece informação acerca dos testemunhos recebidos por cada usuário, esses dados devem ser obtidos de forma indireta, coletando-se o arquivo HTML que representa a página de perfil do usuário, para então extrair deste os testemunhos recebidos.

Note que o Flickr não disponibiliza informações acerca dos depoimentos enviados por um usuário, assim como informações sobre quem adicionou um determinado usuário em sua lista de contatos. No entanto, para desenvolver uma caracterização mais completa das redes sociais, algumas vezes serão obtidas essas informações de forma indireta, a partir das interações existentes em cada rede. Logo, essas informações não obtidas diretamente do Flickr são parciais devido à visão limitada das redes.

3.1.2 Características Individuais dos Usuários

De forma a realizar uma caracterização mais completa do sistema, outras informações além das listas de contatos e testemunhos de cada usuário foram selecionadas para coleta e podem ser obtidas por meio da API disponibilizada pelo Flickr. Essas informações são:

- Número de fotos públicas do usuário: permite conhecer, por exemplo, o quanto o usuário contribui com conteúdo no sistema.
- Listas de grupos públicos do usuário: permite entender como usuários do sistema se agrupam segundo interesses.
- Tipo de conta do usuário: identifica se um usuário é pagante (conta “pro”) ou utiliza conta gratuita.
- Número de visualizações e data de inserção de cada foto pública do usuário: objetivando realizar estudo de tráfego acarretado pelas fotos dos usuários, de forma complementar a trabalhos como [28, 42], que desenvolveram análises relacionando popularidade de fotos com a relação de contatos.

3.2 Processo de Coleta e Limpeza de Dados

3.2.1 Coleta dos Dados

Para realizar a coleta de dados do Flickr de forma eficiente, desenvolveu-se um coletor distribuído do tipo cliente-servidor. Nessa aplicação, clientes realizam coleta em paralelo enquanto o servidor coordena o processo. Tanto o servidor quanto o cliente foram implementados usando a linguagem de programação Ruby [8].

O mecanismo de coleta utilizado é conhecido na literatura como *Snowball* [27, 11]. Este mecanismo parte de um conjunto de usuários (sementes), obtendo a lista de usuários a serem coletados no próximo nível. Realiza-se então a coleta em todos os usuários do segundo nível, obtendo-se a lista dos usuários não descobertos anteriormente a serem coletados no terceiro nível. Assim ocorrendo sucessivamente até que nenhum novo usuário seja encontrado ou a coleta seja interrompida. Em outras palavras, a coleta ocorre de forma similar a uma busca em largura. É conhecido que o *Snowball* tende a superestimar vértices de alto grau, uma vez que podem ser mais facilmente encontrados da rede por meio das arestas [27, 11]. Outros mecanismos existentes para coleta de dados incluem amostragem por vértices e amostragem por arestas que, segundo [11] são inferiores ao *Snowball* por serem menos eficientes (algumas vezes nem aplicáveis) e, quando a coleta não cobre um tamanho suficientemente grande dos dados, resultam em uma rede muito fragmentada e muito diferente da rede original em muitos aspectos.

Implementou-se no coletor duas estratégias (ou fluxos) diferentes para realizar o *Snowball*: (1) coleta da rede de contatos, na qual a lista de candidatos ao próximo nível consistia dos usuários que estavam na lista de contatos dos usuários coletados no nível corrente e (2) coleta da rede de testemunhos, cuja lista de candidatos ao próximo nível consistia dos usuários que escreveram testemunhos para os usuários sendo coletados no nível corrente. É importante mencionar que em ambos casos a rede é direcionada e não é possível coletar a rede na direção oposta, uma vez que o Flickr não disponibiliza informação relacionada a quem um usuário enviou testemunho ou quem adicionou um determinado usuário em sua lista de contatos. Note que essa restrição implica em não ser possível coletar por completo um componente fracamente conectado (WCC) da rede.

Dessa forma, o coletor distribuído permite dois tipos de coleta diferentes e independentes, uma para cada tipo de rede caracterizada, onde as informações coletadas para cada usuário são aquelas apresentadas na seção 3.1. O algoritmo em pseudocódigo 3.2.1 ilustra em alto nível o processo de coleta para ambos os tipos de rede. Outras requisições à API são realizadas em etapas intermediárias como, por exemplo, conversão de URLs para identificadores de usuários. No entanto, omitiu-se esses detalhes por questões de clareza.

De forma a tentar coletar um maior número de usuários e assim tentar cobrir uma maior fração do sistema, utilizou-se um grande conjunto de sementes iniciais. Essas sementes foram escolhidas como sendo os donos das fotos tidas pelo Flickr como sendo as "mais interessantes" (ver seção 2.1) dos últimos 180 dias que antecederam a coleta. A escolha de utilizar o resultado desse algoritmo proprietário do Flickr se deu devido ao fato de que possivelmente essas fotos

```
   Entrada: Lista L de usuários (sementes)
 1 para cada Usuário U na lista L faça
       /* Coleta das características individuais do usuário U (ver seção 3.1)
          */
 2     Coleta grupos públicos do usuário U usando a API do Flickr;
 3     Coleta número de fotos e tipo de conta do usuário U usando a API do Flickr;
 4     se Coletando rede de testemunhos então
 5         Coleta informações sobre cada fotos do usuário U usando a API do Flickr;
 6     fim
       /* Coleta das características sociais do usuário U (ver seção 3.1)   */
 7     Coleta HTML da página de perfil P do usuário U;
 8     Extrai da página de perfil de U, a url R dos usuários que enviaram testemunhos;
 9     para cada Url R extraída de P faça
           // Obtém usuários que enviaram testemunhos
10         Utiliza a API do Flickr para traduzir R em identificador do usuário;
11         Armazena identificador traduzido;
12     fim
13     Coleta a lista de contatos de U usando a API do Flickr;
       /* Agenda usuários para serem coletados no próximo nível             */
14     se Coletando rede de testemunhos então
15         Adiciona ao fim da lista L os novos usuários encontrados por meio de contatos;
16     senão
17         Adiciona ao fim da lista L os novos usuários encontrados por meio de
           testemunhos;
18     fim
19 fim
```

Algoritmo 3.2.1: Processo de coleta de dados para ambas redes.

são algumas das mais visitadas no Flickr, provindo de usuários provavelmente mais ativos no sistema.

Nossa carga de trabalho consiste portanto de pares de coletas independentes, onde cada par utiliza o mesmo conjunto de sementes iniciais (totalizando mais de 39.000 sementes únicas), uma seguindo a rede de testemunhos e outra seguindo a rede de contatos. Nenhuma coleta foi interrompida antes do fim do último nível, prosseguindo portanto enquanto algum novo usuário for descoberto. Note que mesmo assim os dados não necessariamente refletem um componente completo do sistema, uma vez que o Flickr não fornece informações bidirecionais acerca dos relacionamentos que formam as redes.

Abaixo resumimos as principais características do coletor implementado:

- Servidor

  O processo servidor tem como função gerenciar a coleta, mantendo listas de identificadores dos usuários coletados, dos usuários em coleta (pelos nós-cliente) e dos usuários a serem coletados no próximo nível. A única informação recebida pelo servidor para início do processo é o conjunto de sementes que iniciarão o processo.

  O servidor foi desenvolvido de forma flexível o suficiente para impedir que a queda ou

finalização de processos cliente interfiram na integridade da coleta. O servidor permite também interromper a coleta a qualquer momento ou após o finalização do atual nível de coleta. Outras tarefas administrativas foram implementadas para acompanhar ou interferir remotamente no andamento da coleta como, por exemplo, parar ou adicionar clientes.

- Clientes

  Um processo cliente recebe do servidor um conjunto de usuários e realiza a coleta destes, retornando ao servidor os identificadores dos usuários candidatos à coleta no próximo nível, considerando o tipo de coleta seguido. No caso da coleta da rede de contatos, ele retornará a lista de contatos dos usuários coletados. No caso da coleta da rede de testemunhos, ele retornará a lista de usuários que enviaram testemunhos. Para permitir que o coletor distribuído seja o mais genérico possível, possibilitando futura reutilização do núcleo do coletor distribuído para o estudo de outros sistemas que não sejam o Flickr, o módulo de comunicação com o servidor e manipulação da lista de sementes foi separado do módulo de coleta de dados propriamente dito. Somente o módulo de coleta é dependente de código desenvolvido para obter dados do sistema alvo, o Flickr.

  Além do registro e obtenção da chave obrigatória para se coletar quaisquer dados do Flickr utilizando a API disponível, verificou-se a necessidade de realizar a coleta utilizando um usuário autenticado, tanto para a obtenção dos dados recuperados pela API, quanto para os dados recuperados encapsulando arquivos HTML. Isso se deve ao fato de que a coleta de alguns dos dados selecionados era impossibilitada se realizada por usuários não autenticados. Verificou-se também que, para alguns métodos da API, requisições autenticadas obtêm um número maior de informações. Dessa forma, criou-se um único usuário que é utilizado por todos os clientes. Todos os dados foram coletados de forma autenticada.

3.2.2 Limpeza dos Dados

Após a coleta da redes de testemunhos identificou-se usuários que não haviam enviado ou recebido testemunhos. Esses usuários, contidos no conjunto das sementes utilizadas, foram removidos das análises. De forma análoga, para a coleta por contatos removeu-se os usuários que constituíam sementes que não haviam sido adicionados em alguma lista de contatos e que também não haviam adicionado usuários em sua lista de contatos. A fração encontrada desse tipo de usuário foi inferior a 0,01% para a coleta da rede de contatos e aproximadamente 28,7% para a coleta da rede de testemunhos. Essa diferença significativa se deve pelo número reduzido de usuários que enviam ou recebem testemunhos no Flickr. Um grande número dos usuários semente não haviam enviado ou recebido testemunhos.

Verificou-se também, por meio de uma mensagem encontrada na página de perfil do usuário, que alguns usuários não estavam mais ativos no sistema. Como o sistema não mais armazenava algumas informações relevantes acerca desses usuários (ex: informações acerca das fotos e grupos), optou-se por excluí-los das análises. Esses usuários correspondem a menos

de 0,1% para a rede de contatos e aproximadamente 4,5% para a rede de testemunhos. O fato de se ter uma maior fração de usuários inativos na rede de testemunhos é devido ao fato que, enquanto um usuário inativo tende a ser removido rapidamente das listas de contato no qual estaria presente, seus testemunhos enviados permanecem nas páginas de perfil dos usuários, permitindo que ele seja encontrado pela rede de testemunhos.

Finalmente, após a filtragem de usuários, verificou-se a presença de algumas arestas paralelas (menos de 0,01% para ambas redes), isto é, um usuário constando mais de uma vez em uma só lista de contatos ou um usuário recebendo mais de um testemunho de algum usuário. Como esses casos consistiam exceções e, portanto, caracterizando respostas mal formuladas dos servidores do Flickr ao responder requisições, essas arestas foram transformadas em arestas simples.

3.3 Principais Métricas Usadas na Caracterização

A caracterização das redes coletadas foi feita utilizando um conjunto de métricas que serão descritas em detalhes a seguir.

3.3.1 Métricas para Análise das Redes Sociais

Para a análise de redes sociais utilizou-se a modelagem por grafos, cuja representação foi discutida na seção 3.1.1. A teoria básica e motivação por traz das métricas de grafos que serão utilizadas foi apresentada na seção 2.2.

Métricas Relacionadas à Distribuição de Graus no Grafo

Uma métrica comumente utilizada diz respeito à distribuição dos graus do grafo. Durante o trabalho, a distribuição dos graus será apresentada em forma de função de distribuição acumulada complementar (que denominaremos como CCDF), tanto para o grau de entrada do vértice (i.e., número de testemunhos recebidos ou número de pessoas que adicionaram o usuário em suas listas de contato) quanto para o grau de saída (i.e., número de testemunhos enviados ou número de usuários adicionados na lista de contato do vértice em questão).

Para complementar a caracterização, será determinada a distribuição estatística que melhor representa os dados por meio de regressão linear utilizando o método dos mínimos quadrados [24].

Distribuição dos Tamanhos dos Componentes Fortemente Conectados

Em um grafo direcionado, um componente é tido como fortemente conectado se cada vértice é alcançável a partir dos outros vértices do componente.

Como feito para o caso da distribuição de graus, a distribuição do tamanho dos componentes fortemente conectados será apresentado em forma de função acumulada complementar.

Coeficiente de Agrupamento

O coeficiente de agrupamento (*clustering coefficient*) de um vértice indica a probabilidade de existir arestas entre os vértices adjacentes ao vértice em questão[1]. Em outras palavras, essa métrica captura a densidade local do grafo. O coeficiente de agrupamento da rede é definida como a média dos coeficientes de agrupamento de todos os vértices da rede.

Formalmente, o coeficiente de agrupamento de um vértice i é definido como:

$$CC(i) = \frac{|\{e_{jl}\}|}{k_i \times (k_i - 1)} : j, l \in Adj_i$$

onde j e l são vértices adjacentes ao vértice i, e_{jl} representa arestas existentes entre os vértices representados por j e l. O grau de i é representado por k_i. Logo, o coeficiente de agrupamento é a divisão entre o número real de arestas existentes entre os vértices adjacentes a i e o número total de arestas que poderiam existir entre os vértices adjacentes.

Note que o coeficiente de agrupamento é um número entre 0 e 1, onde valores mais altos indicam maior coesão. O coeficiente de agrupamento é definido apenas para usuários com pelo menos dois vértices adjacentes.

Assortatividade

A assortatividade diz respeito ao estudo da relação entre o grau de um vértice e os graus de seus vértices vizinhos. Essa métrica permite entender se na rede os vértices em geral se ligam a outros com grau próximo ao do vértice em questão ou tendem a se ligar a vértices com grau muito diferente (muito maiores ou muito menores).

Visualmente pode-se representar a assortatividade a partir de um gráfico onde, para cada grau k encontrado em pelo menos um vértice da rede, representa-se o grau médio k_{nn} dos vizinhos dos vértices de grau k. Este gráfico pode ser feito separadamente considerando graus de entrada e de saída. A assortatividade também pode ser representada numericamente por meio do coeficiente de Pearson [11]:

$$r = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sqrt{(\langle k_i^2 \rangle - \langle k_i \rangle^2)(\langle k_j^2 \rangle - \langle k_j \rangle^2)}}$$

onde k_i e k_j são os graus dos vértices que participam de uma aresta enquanto a notação $\langle\ \rangle$ representa a média sobre todas as arestas da rede. O coeficiente r é um valor entre -1 e 1, onde um valor positivo mais alto implica maior probabilidade de haver arestas entre nós de graus semelhantes (usuários com grande número de arestas tendendo a se ligar a outros também com grande número de arestas) enquanto um valor negativo implica maior probabilidade de encontrar arestas entre usuários de graus diferentes.

[1] O coeficiente de agrupamento é normalmente definido para grafos não direcionados, não havendo ainda um consenso quanto a definição para o caso de grafos direcionados. Logo, a definição utilizada aqui pode variar em relação a outros trabalhos.

Reciprocidade

Uma outra métrica geralmente utilizada para caracterizar redes direcionadas é a reciprocidade, que está relacionada à probabilidade de dois vértices estarem ligados por arestas de direções opostas [37].

Utilizou-se para o cálculo da reciprocidade da rede a definição apresentada em [20]:

Sendo r a fração entre o número de arestas recíprocas $L^{\leftrightarrow}$ (apontando em direções opostas) sobre o total L de arestas, ou seja:

$$r \equiv \frac{L^{\leftrightarrow}}{L}$$

Sendo $\overline{a}$ a densidade da rede, ou seja:

$$\overline{a} = \frac{L}{N \times (N-1)}$$

onde N é o número total de vértice do grafo. Define-se então a reciprocidade da rede como sendo:

$$recip = \frac{r - \overline{a}}{1 - \overline{a}}$$

A reciprocidade da rede é um valor entre -1 e 1, onde valores positivos indicam alta reciprocidade e valores negativos indicam baixa reciprocidade. O valor 0 indica um número de arestas recíprocas próximo ao de um grafo aleatório.

Além da métrica acima apresentada, que calcula a reciprocidade da rede, utilizou-se uma métrica para representar a reciprocidade de cada vértice individualmente. Essas duas métricas de reciprocidade são calculadas de forma independente. Define-se então a reciprocidade de um vértice como sendo a fração de arestas de entrada para as quais existe uma aresta de saída. No caso em questão estamos medindo a fração de contatos ou testemunhos recebidos para os quais o usuário responde. Durante o trabalho, a reciprocidade dos vértices será apresentada na forma de distribuição acumulada complementar.

3.3.2 Métricas para Análise das Características Individuais dos Usuários

Análise dos Números de Contatos, Fotos e Grupos dos Usuários

Serão apresentadas as médias e coeficientes de variação dos números de contatos, fotos e grupos dos usuários de cada uma das coletas realizadas.

Complementarmente, para cada coleta serão caracterizadas as distribuições do número de contatos, fotos e grupos por usuários, por meio de funções de distribuição acumulada complementar. Os dados serão aproximados de uma distribuição encontrada por meio de regressão linear utilizando o método dos mínimos quadrados.

Correlação entre Características dos Usuários

Será utilizado o coeficiente de correlação linear [24] para analisar a correlação entre diversas características do usuário como, por exemplo, o número de contatos ou testemunhos com o número de contatos, fotos e grupos do usuário.

Análise de Tráfego das Fotos dos Usuários

Utilizando uma métrica de tráfego médio das fotos de um usuário, a ser apresentada a seguir, serão realizados cálculos da correlação linear entre diferentes características dos usuários e o tráfego acarretado pelas fotos.

Considere um usuário u, o número de fotos F do usuário, a data da coleta D das fotos do usuário, data de inserção d_i da i-ésima do usuário foto no sistema, além do número de visualizações V_i da i-ésima foto do usuário. A métrica tráfego médio por dia do usuário u será definida formalmente a seguir:

Uma vez que o Flickr não disponibiliza o número de visualizações de cada foto por dia, mas apenas o valor acumulado de visualizações de cada foto desde a data de inserção no sistema até o momento da coleta, aproximou-se o número médio de visualizações por dia. Tem-se então o tráfego médio por dia da i-ésima foto do usuário ($\overline{V_i}$) como sendo:

$$\overline{V_i} = \frac{V_i}{D - d_i}$$

De posse de $\overline{V_i}$, calcula-se então o tráfego do dia j, $\overline{V_j}$, como sendo a soma do tráfego médio por dia de cada foto no dia j. Ou seja:

$$\overline{V_j} = \sum_{i=1;d_i \leq j}^{F} \overline{V_i}$$

Seja d_{min} a data de inserção da foto mais antiga do usuário, ou seja:

$$d_{min} = min(d_i) \forall i$$

O próximo passo consiste em obter para cada um dos dias entre a inserção da foto mais antiga do usuário e a data em que foi realizada a coleta dos dados do usuário, a soma do tráfego das fotos para aquele dia. A média do valor sobre todos os dias finaliza o cálculo da métrica utilizada durante o trabalho. A métrica tráfego médio por dia do usuário u é portanto definida como:

$$t_u = \frac{\sum_{j=d_{min}}^{D} \overline{V_j}}{D - d_{min}}$$

A métrica t_u permite identificar usuários que atraem maior tráfego no sistema, podendo ser utilizada por exemplo para indicar usuários mais interessantes a serem contemplados em buscas por conteúdo, sugerir promoção ou realizar classificação de usuários.

Capítulo 4

Caracterização das Redes de Relacionamentos

Este capítulo apresenta as caracterizações realizadas, estando estruturado em três partes. A primeira parte parte (seção 4.1) apresenta uma análise das características dos usuários. Já a segunda parte (seção 4.2) aborda uma caracterização das coletas do ponto de vista de redes sociais. Finalmente, na terceira parte (seção 4.3) discute-se acerca da evolução das redes, analisando-se as características dos usuários ao longo do tempo.

4.1 Caracterização dos Usuários

Nesta seção será apresentada uma caracterização dos usuários nas coletas realizadas, visando entender melhor as características dos mesmos e contrastar o padrão de comportamento dos usuários em cada coleta. Uma análise separada dos usuários de cada coleta é importante pelo seguinte motivo: apesar da coleta por testemunhos estar quase totalmente dentro da coleta por contatos, a coleta por testemunhos possui um tamanho muito inferior. Logo, os usuários da coleta por testemunhos podem possivelmente apresentar características muito distintas daquelas encontradas nos usuários da outra coleta mas, devido ao tamanho, ser mascaradas por quaisquer análises que envolvam todos os usuários da coleta por contatos.

A tabela 4.1 apresenta informações sobre as duas coletas que serão discutidas ao longo do capítulo. Utilizou-se o termo interação para indicar uma interação social (testemunho ou contato). A sigla CV se refere ao coeficiente de variação. Durante o capítulo, as coletas por meio das informações de testemunhos e contatos serão também denominadas respectivamente como rede de testemunhos e rede de contatos. Note que o número de usuários coletados na coleta por contatos é muito superior ao número de usuários coletados em trabalhos anteriores que também analisaram a rede de contatos no Flickr [34, 15, 33]. Até onde sabemos, este é o primeiro trabalho que aborda a coleta de usuários por meio de relações de testemunhos no Flickr.

Uma observação geral que se pode fazer a partir da análise da tabela é relativo à grande diferença entre as coletas. Provavelmente o aspecto mais marcante é a diferença de tamanho,

	Coleta a partir dos testemunhos dos usuários	**Coleta a partir dos contatos dos usuários**
Data de coleta	04/04/2008	04/04/2008
# usuários	56.718	3.531.505
# interações	139.786	57.504.406
# médio de contatos por usuário	317,38 (CV = 1,98)	16,28 (CV = 7,20)
# médio de fotos por usuário	707,41 (CV = 2,74)	135,98 (CV = 17,90)
# médio de grupos por usuário	108,46 (CV = 1,27)	6.88 (CV = 5.20)
# usuários na interseção entre as coletas	53.005 (93,45% da coleta)	53.005 (0,0015% da coleta)

Tabela 4.1: Informação sobre as coletas utilizadas na análise (após limpeza dos dados).

que varia em ordens de grandeza. Esse fato pode ser explicado pela utilização das funcionalidades do Flickr pelos seus usuários: grande parte dos usuários do Flickr nunca receberam um testemunho. De fato, a fração de usuários da rede de contatos (a maior coleta) que receberam pelo menos 1 testemunho é inferior a 2,5%. Além disso, os números médios de fotos, grupos e contatos dos usuários da rede de testemunhos é bastante superior (até duas ordens de grandeza) do que os números equivalentes para os usuários na rede de contatos. Note ainda a maior variabilidade (maior CV) dessas métricas nesta segunda rede. Estes resultados, juntamente com a grande interseção dos usuários da rede de testemunhos com a rede de contatos (maior que 93%) pode indicar a existência de uma classe ou subconjunto de usuários da rede de contatos (aqueles que fazem uso de testemunhos) que utiliza o sistema de forma diferente, levantando perguntas tais como: *como cada um desses tipos de usuário utiliza o sistema? Quem são os usuários que fazem uso dos testemunhos? Há relação entre a popularidade das fotos de um usuário e número de testemunhos e contatos que ele possui?*

Para entender melhor as características de cada coleta, será apresentada na seção 4.1.1 uma análise mais aprofundada acerca de como os usuários utilizam o sistema. A seção 4.1.2 analisa a relação entre popularidade das fotos dos usuários existentes na interseção das duas coletas e aspectos sociais do usuário (em relação a testemunhos e contatos do usuário).

4.1.1 Análise das Características Individuais dos Usuários

A tabela 4.1 apresenta características como os números médios de contatos, fotos e grupos por usuário. Note os grandes valores encontrados para os coeficientes de variação. Esta grande variabilidade pode indicar que a média pode não ser uma boa medida para refletir o padrão de comportamento dos usuários. A seguir será descrita uma análise mais completa acerca do comportamento dos usuários de ambas coletas.

A figura 4.1 apresenta as distribuições dos números de contatos públicos dos usuários em cada coleta, representando o fator social do usuário na utilização do sistema. Conforme visto em outros sistemas baseados em redes sociais [34, 11, 26], a distribuição do número de

contatos segue uma distribuição de cauda pesada (Lei de Potência), indicando a existência de poucos usuários com muitos contatos e muitos usuários com poucos contatos. Por meio de aproximação, verificou-se que os dados podem ser representados por uma Zipf [44] ($P[x] = kx^{-\alpha}$), cujo parâmetro (α) e a qualidade de aproximação (fator de determinação R^2) são $\alpha_{contatos} = 1,55$ ($R^2 = 0,91$) e $\alpha_{testemunhos} = 1,28$ ($R^2 = 0,86$) para as coletas por contatos e testemunhos, respectivamente[1]. Valores de R^2 mais altos indicam uma melhor aproximação dos dados. O menor valor de α para a coleta por testemunhos indica que ela é mais uniforme em relação ao número de contatos mantido pelos usuários. Percebe-se ainda um número de contatos muito superior para os usuários da coleta por testemunhos. Em particular, a probabilidade de um usuário na rede de testemunhos ter mais que cem contatos é de 56,17%. Já considerando os usuários na rede de contatos, esta probabilidade é de apenas 2,82%. Logo, nota-se uma grande diferença no padrão geral de comportamento dos usuários das redes: os usuários contidos na rede de testemunhos tendem a interagir por meio de contatos com um número significativamente maior de pessoas no sistema. Acredita-se que essa diferença seja devido à natureza da rede de contatos que acaba por concentrar diversos tipos de usuários, entre eles usuários que não exploram as características sociais do sistema. Alguns exemplos desses tipos de usuários seriam aqueles que utilizam o sistema focando mais em ter espaço para armazenamento pessoal de fotos, usuários que compartilham suas fotos com grupos pequenos ou restritos de pessoas (ex: família), ou mesmo usuários que registraram no sistema apenas para conhecer as suas funcionalidades. Já a coleta por testemunhos inclui usuários que fazem maior uso das funcionalidades de interação social do sistema.

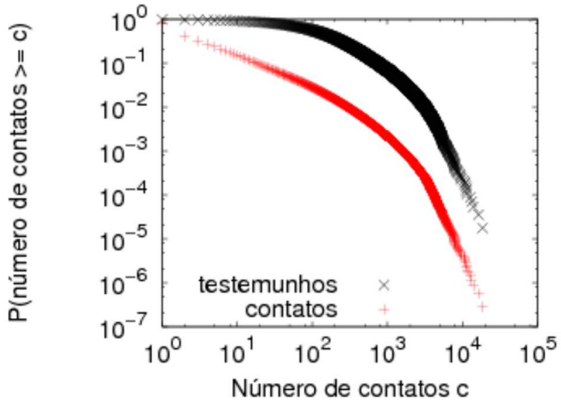


Figura 4.1: Distribuição do número de contatos por usuário nas duas coletas.

Conforme mencionado anteriormente, um número muito pequeno de usuários na rede de contatos recebeu algum testemunho. A figura 4.2 ilustra esse fato apresentando a distribuição do número de testemunhos recebidos pelos usuários em ambas coletas. As curvas são bem modeladas por distribuições Zipf com parâmetro $\alpha_{contatos} = 2,23$ ($R^2 = 0,97$) e $\alpha_{testemunhos} = 2,13$ ($R^2 = 0,97$). Note que pelas distribuições serem do tipo Lei de Potência, alguns usuários

[1]Daqui por diante será utilizada a notação $\alpha_{contatos}$ e $\alpha_{testemunhos}$ para se referir respectivamente ao parâmetro α das distribuições encontradas para as coletas por contatos e por testemunhos.

recebem muito mais testemunhos (provindos de contatos diferentes) do que outros. Verificou-se ainda em ambas coletas uma forte correlação do número de testemunhos recebidos com o número de contatos dos usuários. Esta correlação é de 0,38 e 0,40 nas coletas por testemunhos e contatos respectivamente. Este resultado é intuitivo uma vez que, conforme mencionado na seção 2.1, um usuário somente recebe testemunhos de quem está em sua lista de contatos. A forte correlação encontrada para a coleta por contatos sugere ainda que os usuários que não receberam testemunhos (parte dominante da coleta) em geral apresentam um número baixo de contatos (isto é, correspondem à cauda da distribuição). Uma discussão complementar acerca da distribuição do número de contatos e testemunhos será feita ao realizar a análise de redes sociais na seção 4.2.

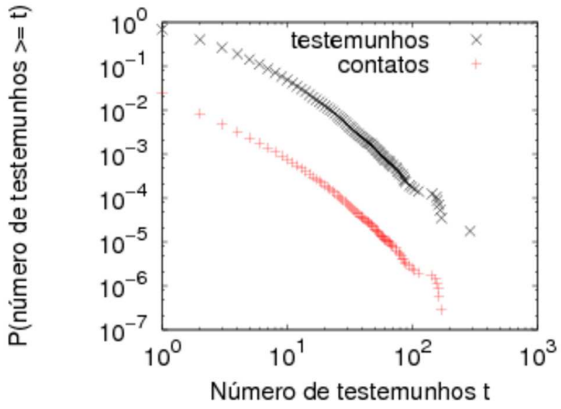


Figura 4.2: Distribuição do número de testemunhos por usuário nas duas coletas.

Outra característica analisada diz respeito ao quanto cada usuário contribui com conteúdo publicamente disponibilizado no sistema. A figura 4.3 apresenta as distribuições dos números de fotos por usuário, que podem ser aproximadas por distribuições Zipf com $\alpha_{contatos} = 1,52$ ($R^2 = 0,93$) e $\alpha_{testemunhos} = 1,14$ ($R^2 = 0,87$). No gráfico pode-se distinguir uma irregularidade nas curvas, mais especificamente no ponto onde a abscissa é igual a 200, que pode ser explicada pelo limite de 200 fotos imposto pelo Flickr para usuários não pagantes (conta gratuita). Percebe-se que os usuários da rede de testemunhos tendem a contribuir um pouco mais com o conteúdo do sistema em relação ao que é visto como padrão geral na rede de contatos, que é provavelmente o padrão geral dos usuários no sistema. De fato, aproximadamente 57% da rede de testemunhos é composto por usuários pagantes, enquanto apenas 13% da rede de contatos é composto por esse tipo de usuário, enfatizando que a rede de testemunhos é constituída, em maior parte, de usuários com maior compromisso com a utilização do sistema, possivelmente utilizando inclusive para fins profissionais. Exemplos de usuários que utilizam o sistema para fins profissionais incluem fotógrafos profissionais, modelos, entre outros.

Analisou-se ainda as distribuições dos números de grupos públicos aos quais um usuário pertence. A figura 4.4 apresenta estas distribuições. Novamente estas distribuições são bem modeladas por distribuições de cauda pesada (Zipf) com parâmetro $\alpha_{contatos} = 2,06$ ($R^2 =$

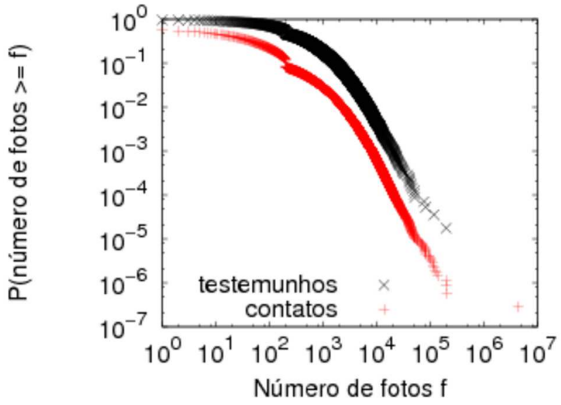


Figura 4.3: Distribuição do número de fotos por usuário nas duas coletas.

$0,91$) e $\alpha_{testemunhos} = 1,77$ ($R^2 = 0,82$). A figura indica que o conjunto de usuários da rede de testemunhos são membros, em geral, de um maior número de grupos. Por exemplo, enquanto 37,45% dos usuários da rede de testemunhos são membros de mais de 100 grupos, esse número é apenas 1,77% para a rede de contatos. Além disso, observou-se que o número de grupos distintos encontrados nas redes de testemunhos e contatos foi de respectivamente 123.396 e 193.925, onde 121.861 estavam presentes em ambas coletas, indicando que grande parte dos grupos de interesse encontrados podem ser descobertos a partir usuários da rede de testemunhos.

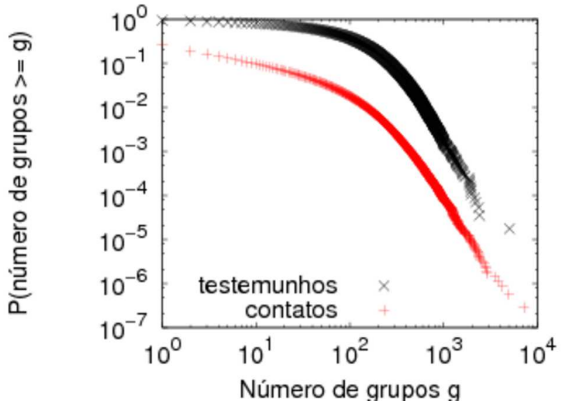


Figura 4.4: Distribuição do número de grupos por usuário nas duas coletas.

Para entender melhor a distribuição de usuários em grupos no Flickr, calculou-se a distribuição do número de membros por grupo (figura 4.5(a)) e o número acumulado de usuários únicos encontrados no sistema a partir dos $M\%$ grupos maiores (figura 4.5(b)). Para esta análise, considerou-se apenas os grupos que apareceram em cada coleta realizada. A figura 4.5(a) indica que os grupos são muito variados em relação ao tamanho, com alguns poucos grupos muito grandes (por exemplo incluindo grupos de temas mais gerais, tais como "Black and White" e "Night Images") e muitos grupos com poucos membros. Percebe-se ainda para a

coleta por contatos, devido à sua magnitude, a prevalência de grupos maiores. Por exemplo, a fração de grupos com mais 100 usuários na coleta por testemunhos é de 9,31%, enquanto para a coleta por contatos é de 17,69%. Já a figura 4.5(b) mostra que se os grupos encontrados forem ordenados por tamanho, uma pequena fração dos maiores grupos já abrangeria uma parte significativa dos usuários únicos. Em especial, para a coleta por testemunhos esse fato é muito mais evidente. Neste caso, os 10 maiores grupos abrangem um total de 57,70% dos usuários únicos, enquanto que na coleta por contatos a mesma quantidade de grupos abrange 19,71% dos usuários únicos.

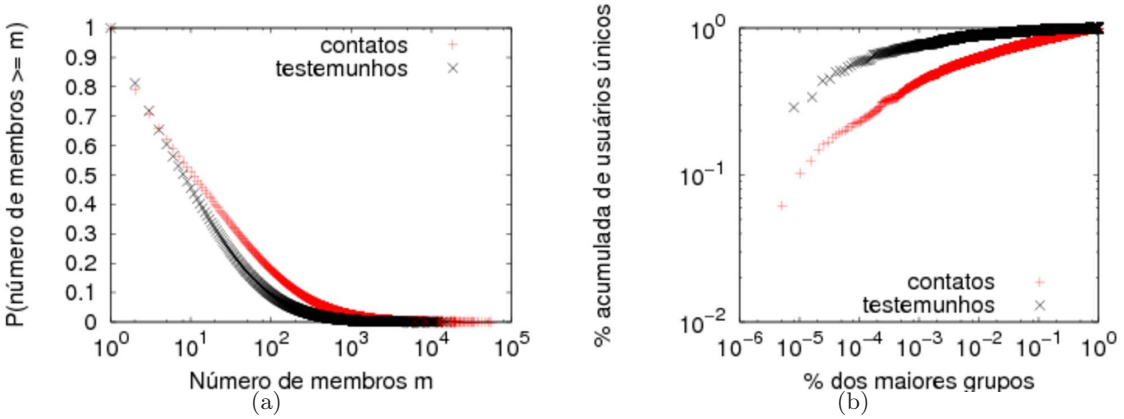


Figura 4.5: Análise da composição dos grupos considerando a visão local das coletas.

Em suma, verificou-se uma maior concentração dos números de fotos, contatos e grupos em alguns poucos usuários, representada por distribuições Zipf de cauda pesada. Alguns poucos usuários fazem muito mais uso das funcionalidades do sistema. Além disso, observou-se que os usuários da rede de testemunhos formam um conjunto de usuários que podem ser considerados muito ativos no sistema, lidando com um número maior de contatos, contribuindo mais para o conteúdo do sistema e participando de um número maior de grupos de interesse. Finalmente, a grande interseção das duas coletas sugere que a rede de testemunhos possa ser considerada um subconjunto de usuários da rede de contatos que utiliza o sistema de forma mais ativa, fortalecendo a importância de uma análise mais profunda dessa rede.

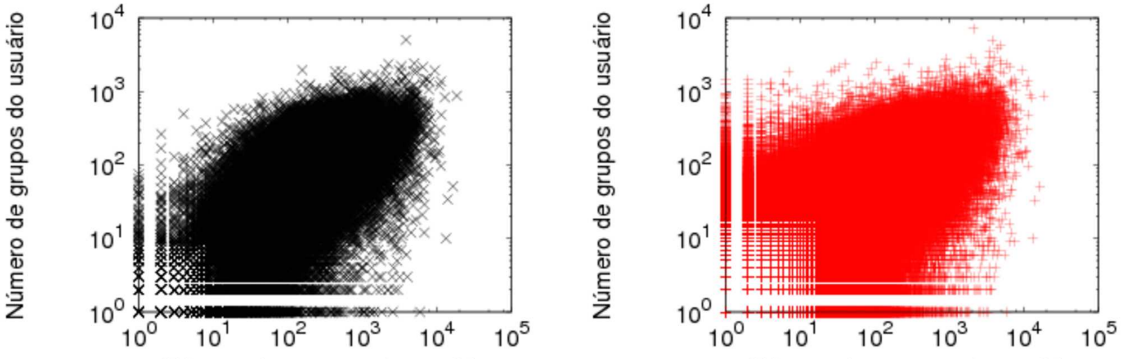
Para completar as análises anteriores, calculou-se também para cada rede as correlações do número de interações (contatos e testemunhos) com os números de grupos e de fotos dos usuários. A tabela 4.2 apresenta os resultados. De forma a analisar também o número de testemunhos enviados pelos usuários da coleta por testemunhos e número de vezes que usuários da coleta por contatos foram adicionados em listas de contatos, ambas constituindo informações não disponibilizadas pelo Flickr, obteve-se esses dados a partir das interações existentes na respectiva coleta. Note portanto que os dados acerca desses dois tipos de interação constituem informações parciais. Para facilitar a apresentação dos dados, a partir deste momento será denominado como "contatos recebidos" o número de vezes que um usuário é adicionado em listas de contatos. Da mesma forma, poderá ser denominado tanto como "contatos adicionados"

quanto simplesmente "contatos" o tamanho da lista de contatos do usuário.

	Rede de testemunhos	**Rede de contatos**
# contatos adicionados x # fotos	0,0303	0,0226
# contatos recebidos x # fotos	–	0,0255
# testemunhos recebidos x # fotos	0,0690	0,0255
# testemunhos enviados x # fotos	0,0423	–
# contatos adicionados x # grupos	0,4077	0,5037
# contatos recebidos x # grupos	–	0,4404
# testemunhos recebidos x # grupos	0,2592	0,2985
# testemunhos enviados x # grupos	0,1897	–

Tabela 4.2: Correlações entre número de contatos ou testemunhos com o número de fotos e grupos.

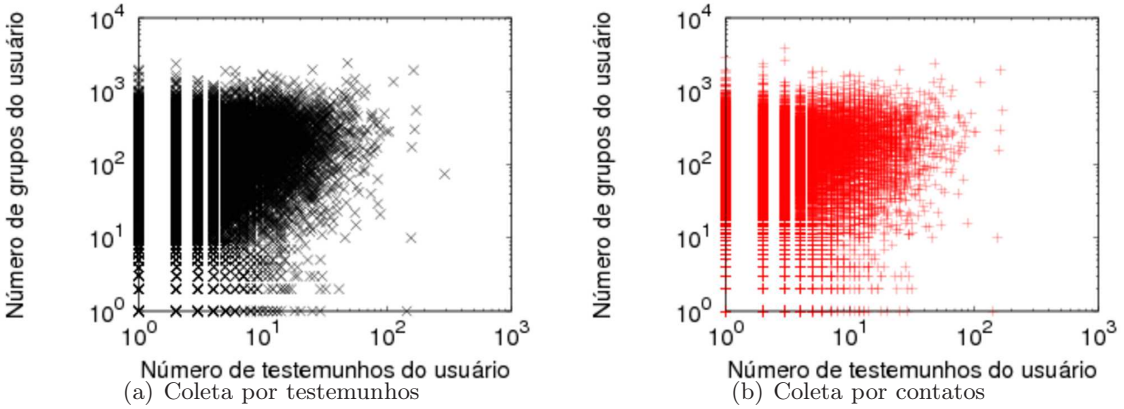
A tabela mostra correlações altas do número de contatos recebidos e adicionados do usuário com o número de grupos em que ele participa. A figura 4.6 ilustra a correlação do número de contatos adicionados de um usuário com o número de grupos para ambas coletas. Uma correlação um pouco mais baixa, porém ainda significativa, foi encontrada entre os números de testemunhos recebidos e enviados pelos usuários com o número de grupos dos mesmos. A figura 4.7 ilustra a correlação do número de testemunhos recebidos com o número de grupos. Esses resultados podem sugerir que usuários formam ou encontram grupos de interesse à medida que conhecem outros usuários no sistema ou que grupos podem ser boas formas de conhecer e interagir com novas pessoas.



(a) Coleta por testemunhos (b) Coleta por contatos

Figura 4.6: Relação entre número de contatos e número de grupos de cada usuário.

Inicialmente poderia-se imaginar que usuários que mantêm maior número de fotos teriam maiores chances de terem suas fotos encontradas por outros usuários e assim conseguir expandir sua rede de contatos e possibilitar maiores interações por meio de testemunhos. No entanto, os resultados mostram uma fraca correlação entre todas as interações e o número de fotos públicas do usuário no sistema. Uma possível hipótese para explicar esse fato seria que o número de relacionamentos do usuário no sistema, tanto por meio de contatos quanto por



(a) Coleta por testemunhos (b) Coleta por contatos

Figura 4.7: Relação entre número de testemunhos recebidos e número de grupos de cada usuário.

meio de testemunhos, tende a ser influenciado pela qualidade do conteúdo do usuário e não necessariamente pelo número de fotos. De fato, os trabalhos de [28, 42] indicam uma forte relação entre as visualizações das fotos do usuário e rede social dele. A próxima seção apresenta uma análise relacionando a atividade do usuário (relacionada às interações de contatos e testemunhos) com a popularidade de suas fotos no sistema.

4.1.2 Popularidade do Conteúdo dos Usuários

Realizou-se um estudo acerca da correlação entre o número de interações sociais, expressas por meio de contatos ou testemunhos, e a popularidade do conteúdo compartilhado pelos usuários, capturada pela métrica t_u (vide seção 3.3), que expressa o número médio de visualizações por dia para todas as fotos do usuário u. Um estudo da popularidade das fotos dos usuários permite identificar os usuários que atraem maior tráfego no sistema.

Devido à magnitude da coleta por contatos, que inviabilizava coletar as informações sobre as fotos de cada usuário, obteve-se essas informações somente para os usuários da coleta por testemunhos. De forma a analisar o número de testemunhos enviados pelos usuários e o número de contatos recebidos (conforme descrito na seção 4.1.1), limitou-se a análise aos usuários contidos na interseção das duas coletas. O número total de fotos desses usuários e portanto analisadas aqui é 38.641.429.

A figura 4.8 apresenta a distribuição de t_u para todos os usuários analisados. Nota-se que, enquanto muitos usuários possuem um conteúdo na média pouco popular, alguns poucos usuários concentram conteúdos mais requisitados no sistema.

A tabela 4.3 apresenta correlações entre os números de contatos (adicionados e recebidos) e de testemunhos (enviados e recebidos) com o número médio de visualizações por dia t_u.

Verifica-se que a maior correlação encontrada foi de t_u com o número de vezes que o usuário u é adicionado como contato (isto é, número de contatos recebidos por u). Esse fato possivelmente pode ser explicado pela funcionalidade existente no Flickr que permite aos

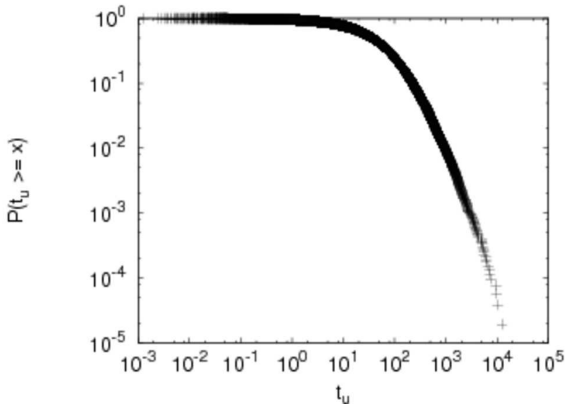


Figura 4.8: Distribuição do número médio de visualizações por dia.

# contatos adicionados por u vs. t_u	0.2104
# testemunhos recebidos por u vs. t_u	0.3948
# contatos recebidos por u vs. t_u	0.4777
# testemunhos enviados por u vs. t_u	0.1734

Tabela 4.3: Correlações de t_u com características do usuário.

usuários monitorarem novas fotos adicionadas pelos seus contatos. Dessa forma, usuários que adicionam fotos muito apreciadas por um grande número de pessoas tendem a ser referenciados como contato por diversos usuários. Verifica-se assim que a interação por contatos, além de refletir relações sociais de contato como, por exemplo, amizade entre duas pessoas no mundo real, pode também nesses casos apresentar relações de interesse de usuários pelo conteúdo de outro. Uma discussão complementar acerca dessa relação de interesse da rede de contatos do Flickr pode ser encontrada em [35]. Os nossos resultados corroboram aqueles obtidos em [28, 42], que também indicam a forte correlação entre rede social de contatos do usuário e visualização de suas fotos.

Outra correlação significativamente forte encontrada é de t_u com o número de testemunhos recebidos pelos usuários. Ao analisar o perfil de diferentes usuários no sistema pode-se verificar que predominantemente os testemunhos estão relacionados ao conteúdo do usuário (fotos ou mesmo textos escritos pelo usuário), sugerindo que essa funcionalidade é normalmente utilizada para indicar o interesse dos usuários pelo conteúdo de algum outro.

A figura 4.9 ilustra estas duas correlações, apresentando para cada usuário u que já foi adicionado às listas de contato de um determinado número de usuários (figura 4.9(a)) ou que recebeu um determinado número de testemunhos (figura 4.9(b)), o número médio de visualizações recebidos por dia desde a inserção de sua primeira foto.

Verificou-se ainda correlações mais fracas de t_u com os números de testemunhos enviados e com o tamanho da lista de contatos do usuário, sugerindo que a popularidade do conteúdo compartilhado por um usuário não depende significativamente desses dois fatores.

Finalmente, para verificar a relação entre o tamanho do conteúdo mantido pelos usuários

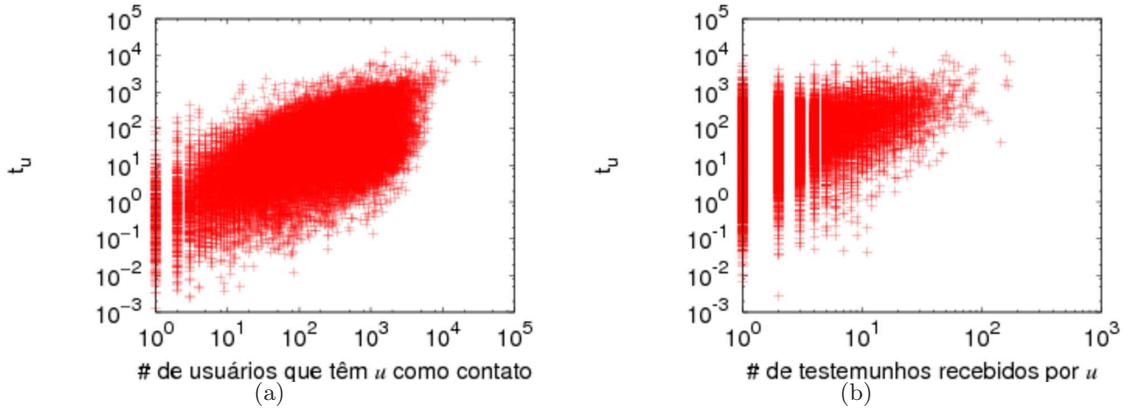


Figura 4.9: Relação entre número de testemunhos ou contatos recebidos e t_u.

e o tráfego acarretado por ele, obteve-se também a correlação do número de fotos com t_u. O valor encontrado foi 0,2760, expressando uma correlação relativamente baixa em relação às duas maiores correlações vistas na tabela 4.3. A figura 4.10 ilustra a relação entre as duas características. Logo, a popularidade das fotos é aparentemente mais influenciada pela qualidade do conteúdo do usuário e não necessariamente pelo número de fotos que o usuário mantém.

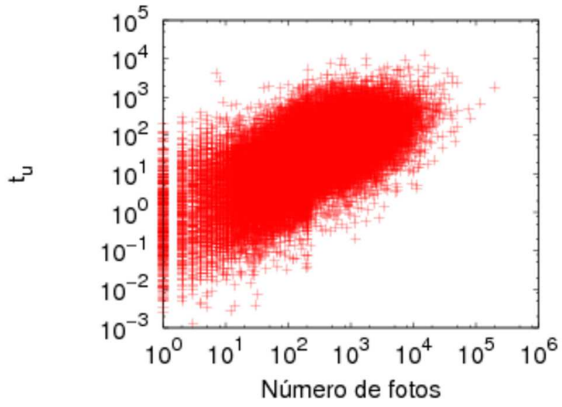


Figura 4.10: Relação entre número de fotos e t_u.

Em suma, verificou-se que o número de contatos (recebidos ou adicionados) e o número de testemunhos (recebidos ou enviados) por um usuário são mais influenciados pela popularidade do conteúdo mantido pelo usuário e não tanto pelo tamanho da coleção mantida por ele. Além disso, observou-se que a popularidade do conteúdo compartilhado por um usuário depende mais da reação dos demais usuários em relação ao conteúdo dele, expressa por meio de contatos recebidos e testemunhos recebidos, não dependendo significativamente do número de testemunhos enviados, tamanho da lista de contatos ou número de fotos mantidas em sua coleção.

4.2 Caracterização das Redes de Contatos e de Testemunhos

Nesta seção será apresentada uma caracterização das estruturas das redes sociais formadas a partir da interação entre usuários por meio de contatos ou testemunhos. Mais especificamente, serão analisados e contrastados dois tipos de grafos, um formado pelos usuários e relações de contatos extraídos da coleta por contatos (denominado rede de contatos) e outro formado pelos usuários e relações de testemunhos extraídos da coleta por testemunhos (denominado rede de testemunhos).

A tabela 4.4 apresenta o tamanho, em número de arestas e de vértices, dos grafos e de seus respectivos componentes gigantes. Um componente gigante se refere ao maior componente fortemente conectado do grafo e será discutido com maiores detalhes na próxima seção. As métricas utilizadas na caracterização das redes são aquelas apresentadas da seção 3.3.

	Rede de testemunhos	**Rede de contatos**
Grafo completo		
# vértices	56.718	3.531.505
# arestas	139.786	57.504.406
Componente gigante		
# vértices	23.813	2.686.355
# arestas	94.674	54.875.148

Tabela 4.4: Tamanho das redes.

4.2.1 Componentes Fortemente Conectados

Uma etapa importante para se entender a estrutura das redes é analisar os componentes fortemente conectados do grafo, que podem representar indiretamente comunidades de usuários. A figura 4.11 apresenta a distribuição do tamanho dos componentes fortemente conectados para ambas redes.

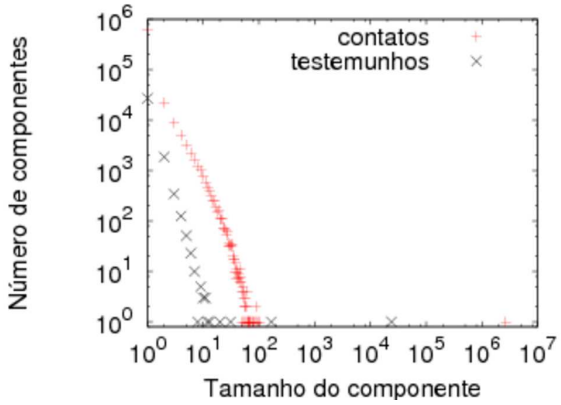


Figura 4.11: Distribuição dos componentes fortemente conectados.

A estrutura das redes é composta por um maior componente fortemente conectado ou componente gigante, diversos componentes intermediários (i.e., 50.084 e 2.436 componentes nas redes de contatos e testemunhos, respectivamente) e um número grande de componentes contendo somente um usuário (i.e., 617.211 e 26.819 componentes nas redes de contatos e testemunhos respectivamente). O componente gigante de ambas redes contém uma fração significativa dos usuários, aproximadamente 76% e 42% das redes de contatos e testemunhos respectivamente. Já o segundo maior componente fortemente conectado de cada rede contém uma fração mínima de seus usuários, menos de 0,001% da rede de contatos e menos de 0,3% da rede de testemunhos. Ou seja, o componente gigante é muito maior do que os diversos componentes intermediários. Além disso, note que existe a diferença significativa da fração da coleta abrangida pelos componentes gigantes de cada rede. Essa diferença pode ser explicada pelo menor coeficiente de agrupamento da rede de testemunhos, indicando que a rede de testemunhos é menos coesa que a de contatos. Uma discussão mais aprofundada acerca do coeficiente de agrupamento nas redes é apresentada na seção 4.2.4.

Como o componente gigante representa uma parte relevante das amostras, concentrando a maior parte dos usuários e das interações da respectiva coleta, o estudo das características sociais das redes incluirá tanto a análise das redes completas como dos respectivos componentes gigantes.

4.2.2 Distribuição dos Graus

Outra característica importante para se entender a estrutura de uma rede direcionada diz respeito às distribuições dos graus de entrada e saída dos vértices da rede, representando o número de pessoas com quem usuários interagem no sistema.

As figuras 4.12 e 4.13 apresentam as distribuições dos graus e as respectivas aproximações para distribuições do tipo Zipf. Os coeficientes de determinação R^2 para as redes de testemunhos e contatos são respectivamente $0,988$ e $0,907$ para as distribuições dos graus de saída e $0,972$ e $0,894$ para as distribuições dos graus de entrada. Note que a distribuição dos graus de saída da rede de contatos é a mesma visto na figura 4.1, que apresentava o tamanho das listas de contatos. Da mesma forma, a distribuição dos graus de entrada da rede de testemunhos é a mesma da figura 4.2, que apresentava o número de depoimentos recebidos pelos usuários. Complementando a análise realizada na seção 4.1.1, a análise das redes que será realizada a seguir engloba também informações acerca do grau de entrada da rede de contatos e grau de saída da rede de testemunhos. É válido mencionar ainda que as distribuições encontradas foram próximas às de [34]. Note ainda que a distribuição de graus da rede de testemunhos do Flickr se diferencia daquela vista para o caso da rede de testemunhos do Cyworld [11] na medida que a última segue uma distribuição exponencial.

Uma observação inicial que se pode fazer sobre os gráficos é relativo ao coeficiente da aproximação da distribuição Zipf. A rede de testemunhos apresenta um parâmetro α maior, indicando haver um número menor de usuários que possuem um grau muito alto em relação ao resto da rede, ou seja, possuindo um pequeno número de usuários muito mais populares ou ativos do que o resto da rede. Note ainda para o caso da rede de contatos um maior

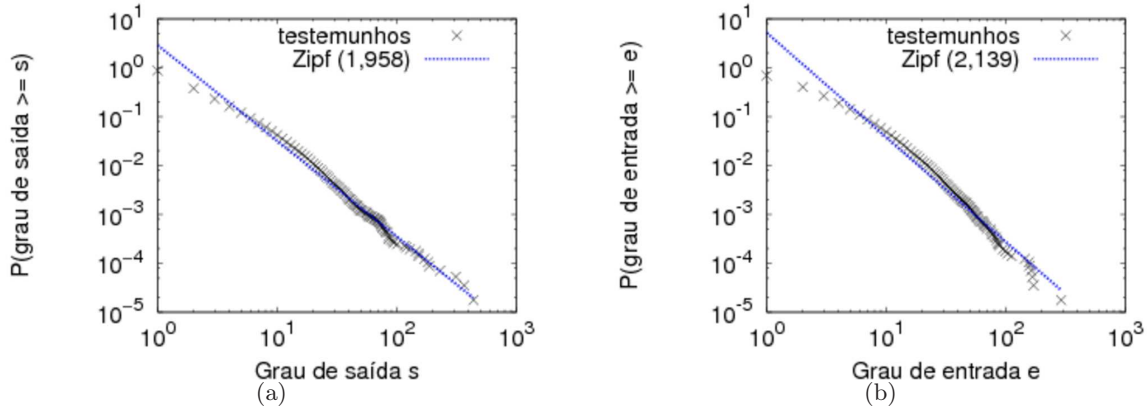


Figura 4.12: Distribuição de graus da rede de testemunhos.

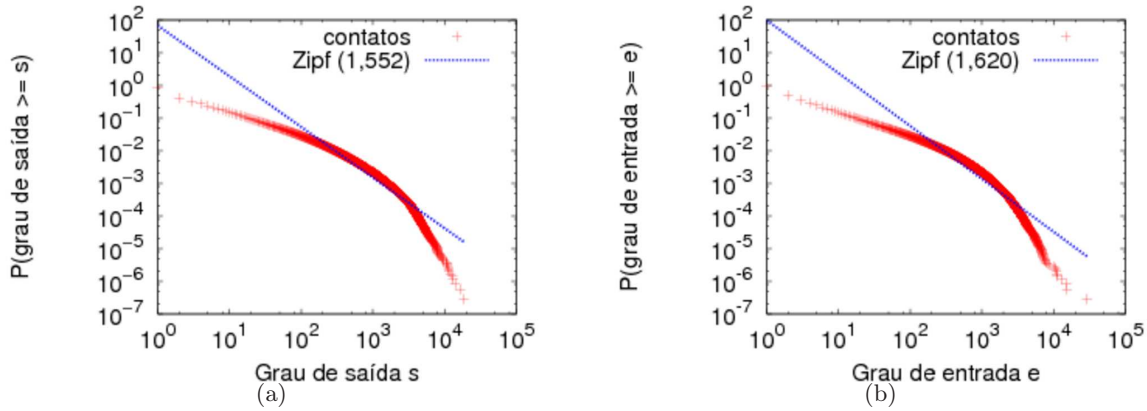


Figura 4.13: Distribuição de graus da rede de contatos.

distanciamento entre a aproximação e o comportamento real dos usuários no que diz respeito aos pontos da curva que recaem sobre os valores iniciais da abcissa. Nessa região há um achatamento da curva, onde muitos usuários apresentam números de contatos adicionados ou recebidos bem mais próximos. Essa constatação pode sugerir a existência de usuários com padrão de comportamento diferentes do previsto.

Ainda, o estudo de [34] apresenta que a rede social de contato de vários sistemas, incluindo o Flickr, tende a apresentar uma alta correlação entre grau de entrada e grau de saída, indicando que usuários que possuem muitos contatos ou enviam muitos testemunhos tendem também a ser muito referenciados como contatos ou receber muitos testemunhos. De fato, a rede de contatos apresentou uma forte correlação entre grau de entrada e grau de saída tanto para a rede completa quanto para o componente gigante (respectivamente 0,7549 e 0,7674). No entanto, a rede de testemunhos apresentou uma correlação um pouco mais fraca (mas ainda bastante significativa) para a rede completa do que para o componente gigante (respectivamente 0.6274 e 0.7256). Uma possível explicação é apresentada na próxima seção,

que discute a reciprocidade nas duas redes.

4.2.3 Reciprocidade

A reciprocidade, que diz respeito às arestas duplas (em direções opostas) existentes entre os pares de vértices, é uma métrica geralmente utilizada no estudo de redes sociais direcionadas [37]. A tabela 4.5 apresenta a reciprocidade média para as redes analisadas, calculada conforme [20].

	Rede de contatos	**Rede de testemunhos**
Reciprocidade da rede completa	0,6247	0,4700
Reciprocidade do componente gigante	0,6479	0,6140

Tabela 4.5: Reciprocidade das redes.

Pela tabela percebe-se que a reciprocidade para a rede de contatos é superior àquela encontrada para a rede de testemunhos, tanto para o caso do grafo completo quanto para o componente gigante. Uma funcionalidade presente no Flickr, assim como em outros sistemas de redes sociais, que pode contribuir para a grande reciprocidade da rede de contatos é o fato de o usuário ser notificado quando outro o adiciona na lista de contatos. O valor da reciprocidade encontrado para a rede de contatos é próximo ao encontrado por [34].

Um aspecto interessante é a menor reciprocidade da rede de testemunhos em relação à reciprocidade do seu componente gigante, que foi inicialmente discutido na seção 4.2.2. Uma possível explicação para essa diferença entre rede de testemunhos completa e seu componente gigante é a menor reciprocidade média na rede completa de testemunhos (0,4700) do que a encontrada no seu componente gigante (0,6140), sugerindo uma diferença entre o padrão encontrado no componente gigante e aquele visto na rede de testemunhos como um todo. Note que uma maior reciprocidade indica que uma maior fração dos testemunhos recebidos são respondidos, acarretando maior proximidade entre os graus de saída e entrada. Uma análise do segundo maior componente fortemente conectado da rede de testemunhos indica uma correlação entre graus (0,8516) e alto valor de reciprocidade (0.7422), reforçando que os componentes fortemente conectados do grafo refletem comunidades mais recíprocas de usuários do que a rede como um todo. A figura 4.14 apresenta a distribuição da reciprocidade dos usuários (conforme definido em 3.3) para as redes e seus componentes gigantes. A grande quantidade de usuários com reciprocidade baixa na rede completa de testemunhos reforça a menor reciprocidade média.

Uma possível explicação para a reciprocidade mais baixa da rede de testemunhos é que testemunhos são em grande parte orientados por interesse pelo conteúdo de um usuário, sendo assim menos relacionado ao fator social. Nesse caso, intuitivamente muitas das relações não seriam mútuas. Para exemplificar, note que nem todo usuário apreciador de fotos necessariamente é um produtor de fotos ou mesmo um produtor de fotos consideradas de alta qualidade. Já a rede de contatos apresenta tanto características sociais quanto características de interesse por conteúdo.

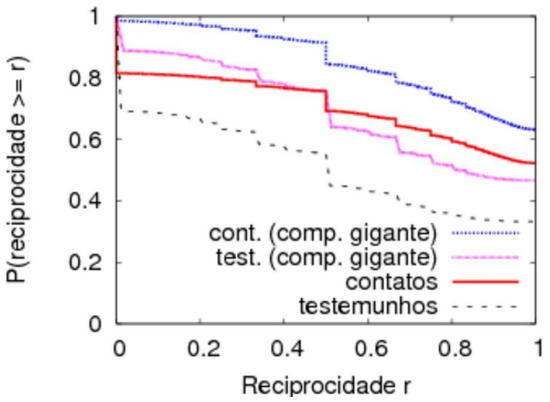


Figura 4.14: Distribuição da reciprocidade.

4.2.4 Coeficiente de Agrupamento

Uma medida muito utilizada no estudo de redes sociais diz respeito ao coeficiente de agrupamento, que mede a coesão da rede. O coeficiente de agrupamento médio da rede reflete o grau de coesão da rede. Redes sociais tendem a apresentar coeficientes de agrupamento não triviais [38].

O coeficiente de agrupamento médio (CC) encontrado para ambas redes e os seus respectivos componentes gigantes são apresentados na tabela 4.6. É válido mencionar ainda que o CC encontrado aqui é bem próximo, mas de valor um pouco menor ao encontrado por [34]. Comparando-se com a rede de testemunhos do Cyworld analisada em [11], verificou-se que no Flickr a rede de testemunhos apresenta um CC menor. Note, no entanto, que o CC apresentado naquele trabalho foi calculado utilizando uma versão não-direcionada do grafo. A rede de contatos do Cyworld não é direcionada e apresenta um CC inferior àquele encontrado aqui para a rede de contatos do Flickr.

	Rede de contatos	**Rede de testemunhos**
CC da rede completa	0,2676	0,1003
CC do componente gigante	0,2857	0,1164

Tabela 4.6: Coeficiente de agrupamento das redes.

Nota-se pela tabela que, tanto para o grafo completo quanto para o componente gigante, a rede de testemunhos é menos coesa que a rede de contatos. Um alto valor do coeficiente de agrupamento para a rede de contatos, refletindo o forte aspecto social da mesma, foi também encontrado em redes de contatos de outros sistemas da Web 2.0 [34] e indica que usuários presentes em uma mesma lista de contatos apresentam uma alta probabilidade de serem também contatos entre si. Já a rede de testemunhos do Flickr apresenta um coeficiente de agrupamento significativo, mas inferior ao da rede de contatos. Essa constatação, juntamente com a menor reciprocidade da rede de testemunhos, sugere que diferentemente do que foi visto para as redes sociais do sistema Cyworld analisadas por [11], a rede de testemunhos do

Flickr provavelmente não é constituída pela relação mais forte de amizade entre pessoas, como concluído naquele trabalho, mas sim direcionada mais pelo conteúdo por elas compartilhado.

Verifica-se também pela análise da tabela que, em ambas redes, o coeficiente de agrupamento é um pouco superior nos componentes gigantes, mas ainda assim apresentando valores muito próximos aos encontrados para as redes completas. Isto é, os componentes gigantes refletem comunidades um pouco mais coesas.

Para complementar a análise, a figura 4.15 apresenta a distribuição dos coeficientes de agrupamento dos usuários para ambas redes. Pela figura percebe-se uma diferença significativa principalmente na fração de usuários que possuem um coeficiente igual a zero (22,39% da rede de contatos e 61,8% da rede de testemunhos), reforçando a menor coesão da rede de testemunhos.

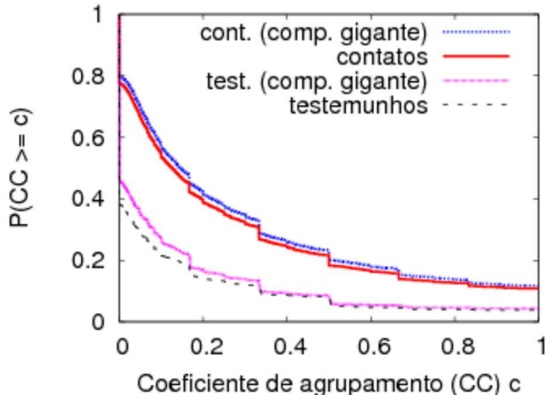


Figura 4.15: Distribuição dos coeficientes de agrupamento dos usuários.

4.2.5 Assortatividade

Esta seção apresenta os resultados relacionados à assortatividade das duas redes analisadas. A assortatividade representa a correlação entre o grau de um vértice e o grau dos vértices conectados a ele. Uma assortatividade negativa indica que vértices de maior grau (usuários mais populares) tendem a se ligar a vértices de grau menor (usuários menos populares), e vice-versa. Já uma assortatividade positiva indica que vértices de alto grau tendem a se ligar a outros de também alto grau, enquanto que vértices de baixo grau tendem a se ligar a outros de baixo grau. Como as redes analisadas no presente trabalho tratam de grafos direcionados, onde separa-se grau de entrada do grau de saída, os resultados serão apresentados para cada combinação possível de graus. Os resultados para os componentes gigantes são muito semelhantes e foram, por isso, omitidos.

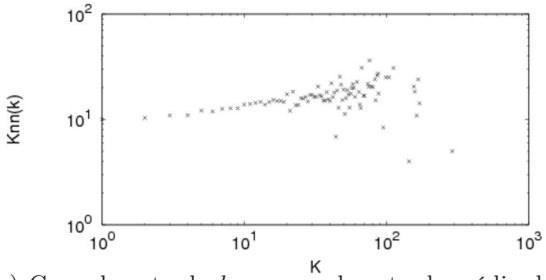
A tabela 4.7 apresenta os coeficientes de Pearson (r) para ambas redes. Utilizou-se a notação $r^{x \to y}$ para representar o coeficiente de Pearson para as diferentes combinações entre tipos de grau, onde x e y podem possuir um dos dois seguintes valores: e para grau de entrada ou s para grau de saída. A variável x equivale ao tipo de grau dos vértices origem das arestas,

enquanto y equivale ao tipo de grau dos vértices destino das arestas. Como exemplo, tem-se $r^{e \to s}$ sendo o coeficiente de Pearson que correlaciona o grau de entrada dos vértices origem das arestas ao grau de saída dos vértices onde as arestas incidem.

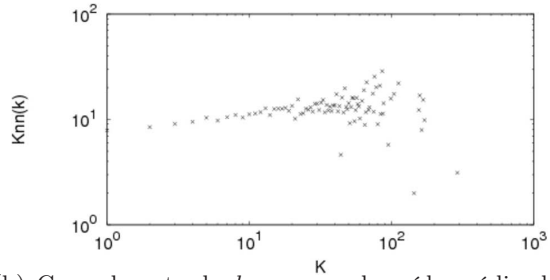
	Rede de contatos	**Rede de testemunhos**
$r^{e \to e}$	0,0539	0,0479
$r^{e \to s}$	0,0954	0,0547
$r^{s \to e}$	0,0324	-0,0004
$r^{s \to s}$	0,0545	0,0071

Tabela 4.7: Coeficientes de Pearson para as duas redes.

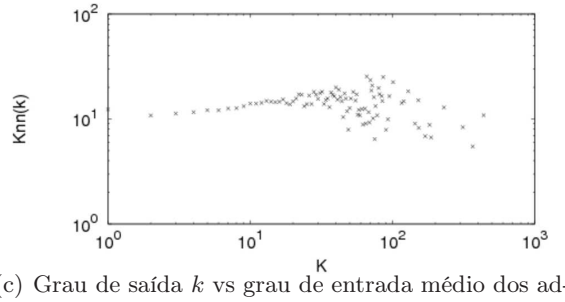
Por uma análise geral da tabela, verifica-se que a rede de testemunhos apresenta coeficientes menores aos encontrados para a rede de contatos. Note ainda que o valor de r muito pequeno para o grau de saída para a rede de testemunhos pode ser decorrência da visão limitada da rede de testemunhos em relação a testemunhos enviados.



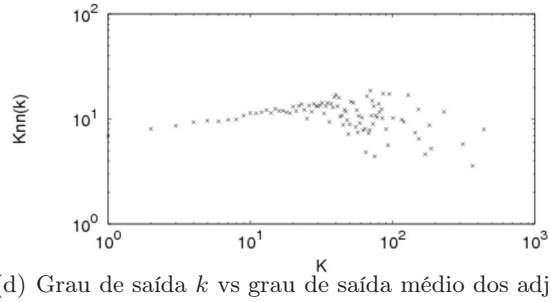
(a) Grau de entrada k vs grau de entrada médio dos adjacentes.



(b) Grau de entrada k vs grau de saída médio dos adjacentes.



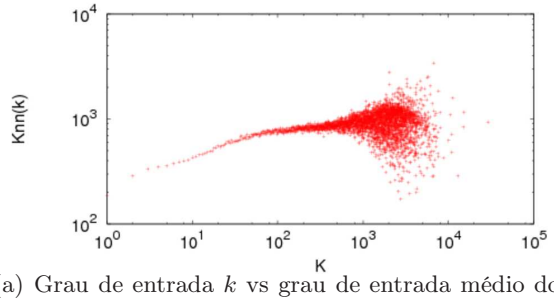
(c) Grau de saída k vs grau de entrada médio dos adjacentes.



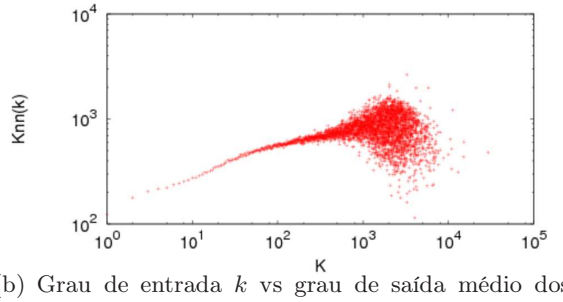
(d) Grau de saída k vs grau de saída médio dos adjacentes.

Figura 4.16: Relação entre grau e o grau médio dos adjacentes na rede de testemunhos.

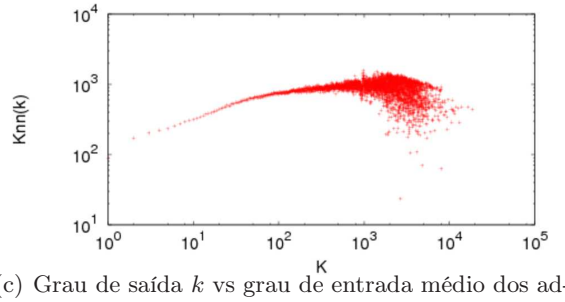
As figuras 4.16 e 4.17 apresentam a relação entre grau dos vértices e o grau médio dos vértices adjacentes. Verifica-se que a rede de contatos apresenta claramente uma assortatividade positiva, uma característica considerada única das redes sociais e encontrada na maioria dessas redes [11]. No entanto, graficamente a rede de testemunhos não aparenta apresentar uma tendência forte em nenhuma direção, mesmo sendo os coeficientes $r^{e \to e}$ e $r^{e \to s}$ da rede de testemunhos próximos ou de mesma ordem daqueles equivalentes na rede de contatos. Isto possivelmente é explicado pela dispersão nos usuários de maior grau (como encontrado



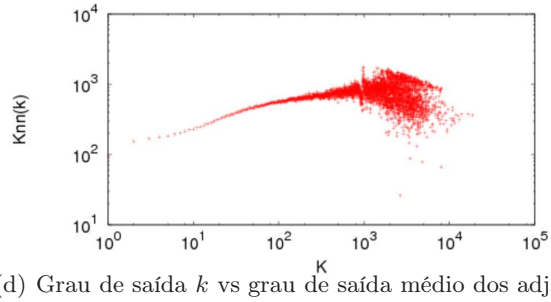
(a) Grau de entrada k vs grau de entrada médio dos adjacentes.



(b) Grau de entrada k vs grau de saída médio dos adjacentes.



(c) Grau de saída k vs grau de entrada médio dos adjacentes.



(d) Grau de saída k vs grau de saída médio dos adjacentes.

Figura 4.17: Relação entre grau e o grau médio dos adjacentes na rede de contatos.

em [11]), onde os valores dos coeficientes de Pearson para a rede de contatos podem ser mais influenciados pelos vértices de alto grau, que aparentemente não seguem uma tendência ou sugerem usuários com assortatividade diferente daquela encontrada nos demais usuários da rede. Neste caso, esses usuários de grau mais alto são responsáveis por diminuir o valor da assortatividade. Esses resultados sugerem que a rede de testemunhos do Flickr se difere daquela analisada para o sistema Cyworld [11], onde se encontrou alto valor para assortatividade.

Em suma, verificou-se que a rede de contatos apresenta características sociais mais altas do que a rede de testemunhos, tais como maior coeficiente de agrupamento e assortatividade positiva. Essa menor característica social da rede de testemunhos, juntamente com os resultados obtidos na seção 4.1.2, sugere que a relação de interesse por conteúdo pode ser o principal motivador de usuários enviarem testemunhos a outros. Além disso, verificou-se que os componentes gigantes das duas redes podem ser vistos como comunidades mais recíprocas e ligeiramente mais coesas do que as respectivas redes completas.

4.3 Evolução das Características dos Usuários e das Redes

De forma a verificar a variação das características dos usuários e das redes durante o tempo, realizou-se a coleta das redes em diferentes datas. As tabelas 4.8 e 4.9 apresentam as principais características para três coletas. Os dados presentes na terceira coluna das tabelas são relativos à coleta analisada e discutida em detalhes nas últimas duas seções anteriores, enquanto as demais colunas correspondem a coletas realizadas anteriormente. O período de

tempo abrangido pelas três coletas é de pouco menos que seis meses.

	Coleta 1 (28/10/07)	**Coleta 2 (31/01/08)**	**Coleta 3 (04/04/08)**
# vértices	43.757	53.424	56.718
# arestas	104.241	131.087	139.786
# médio de contatos por usuário	299,09 (CV = 2,03)	315,70 (CV = 2,00)	317,38 (CV = 1,98)
# médio de fotos por usuário	672,22 (CV = 2,54)	694,54 (CV = 2,69)	707,41 (CV = 2,74)
# médio de grupos por usuário	100,51 (CV = 1,24)	106,16 (CV = 1,27)	108,46 (CV = 1,27)
Distribuição dos graus de entrada (α)	2,107	2,098	2,139
Distribuição dos graus de saída (α)	1,974	1,999	1,958
Reciprocidade	0,472	0,472	0,470
CC	0,100	0,100	0,100
$r^{e \to e}$	0,030	0,026	0,048
$r^{e \to s}$	0,041	0,038	0,055
$r^{s \to e}$	0,000	-0,001	0,000
$r^{s \to s}$	0,007	0,008	0,007

Tabela 4.8: Comparação entre diferentes coletas por testemunhos.

	Coleta 1 (29/10/07)	**Coleta 2 (31/01/08)**	**Coleta 3 (04/04/08)**
# vértices	2.758.472	3.177.725	3.531.505
# arestas	41.556.467	50.941.359	57.504.406
# médio de contatos por usuário	15,07 (CV = 7,35)	16,03 (CV = 7,23)	16,28 (CV = 7,20)
# médio de fotos por usuário	129,60 (CV = 8,18)	133,99 (CV = 17,09)	135,98 (CV = 17,90)
# médio de grupos por usuário	5,67 (CV = 5,18)	6,64 (CV = 5,16)	6,88 (CV = 5.20)
Distribuição dos graus de entrada (α)	1,607	1,617	1,620
Distribuição dos graus de saída (α)	1,521	1,539	1,552
Reciprocidade	0,642	0,632	0,625
CC	0,280	0,273	0,268
$r^{e \to e}$	0,065	0,062	0,054
$r^{e \to s}$	0,094	0,100	0,095
$r^{s \to e}$	0,032	0,035	0,032
$r^{s \to s}$	0,046	0,055	0,054

Tabela 4.9: Comparação entre diferentes coletas por contatos.

Para ambas redes (contatos e testemunhos), observa-se um crescimento em relação ao número de vértices, arestas e características individuais dos usuários. Esses resultados sugerem que, enquanto novos usuários estão entrando no sistema, os usuários antigos estão fazendo maior uso das funcionalidades, indicando um aumento na popularidade do Flickr. Nota-se ainda para a rede de contatos que o parâmetro α da aproximação da Zipf para distribuição dos graus (de entrada e de saída) está aumentando ao longo do tempo. Isso indica que a rede tende a um maior distanciamento entre usuários de maior grau e os de menor grau.

Observa-se ainda para as demais características sociais dos dois tipos de redes que, apesar de apresentarem pequenas variações, eles estão de certa forma estáveis durante período de tempo abrangido pelas três coletas. Logo, isso evidencia que as conclusões obtidas se mantêm pelo menos para esse período de tempo, não sendo apenas relacionadas a uma determinada coleta específica.

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho caracterizou-se e contrastou-se dois tipos de redes de relacionamentos extraídas do Flickr, criadas a partir de diferentes tipos de interações entre usuários: contato e testemunho.

Uma primeira vertente da caracterização foi relativo a características individuais dos usuários de cada uma das redes. Observou-se que, enquanto poucos usuários utilizam muito mais as diversas funcionalidades do sistema, grande parte utiliza pouco. Mais especificamente, verificou-se que as distribuições dos números de contatos, testemunhos, grupos e fotos são bem modeladas por distribuições do tipo Zipf.

Além disso, verificou-se que os usuários presentes na rede de testemunhos são quase totalmente um subconjunto pequeno dos usuários da rede de contatos, porém apresentando um padrão de comportamento diferente daquele visto para a rede de contatos como um todo. Os usuários da rede de testemunhos são mais ativos no sistema, apresentando maior número de fotos, grupos e contatos, além de grande parte ser composta por usuários de conta paga, indicando maior compromisso com a utilização do sistema.

Analisando as fotos dos usuários nas duas redes, constatou-se que o número de interações de um usuário não se deve necessariamente ao tamanho da coleção de fotos mantida por ele, mas está mais relacionado à popularidade do conteúdo mantido por ele. Ainda, verificou-se que a popularidade das fotos de um usuário é bastante correlacionada ao número de testemunhos recebidos e ao número de vezes que o usuário foi adicionado em listas de contatos.

A caracterização dos aspectos sociais das duas redes mostrou que a rede de contatos apresenta características mais sociais, tais como maior coeficiente de agrupamento e assortatividade positiva. Uma explicação para o menor fator social da rede de testemunhos é que esta pode estar mais relacionada a relações de interesse entre usuários pelo conteúdo de outros, por isso sendo menos social. Esse resultado mostra uma rede de testemunhos com características distintas daquelas encontradas para no caso do sistema de relacionamento social Cyworld [11], onde verificou-se que o testemunho representava uma interação mais social do que o contato. Essa diferença entre as redes de testemunhos nos dois sistemas sugere que as características das diversas redes de relacionamentos em um sistema podem ser muito dependentes do objetivo ou foco principal do sistema (ex: compartilhamento de fotos, relacionamento social).

Verificou-se ainda que apesar do crescimento do Flickr, as características sociais das duas redes se mantiveram razoavelmente estáveis ao longo de seis meses.

Existem várias direções possíveis para extensões do trabalho realizado. Uma possibilidade seria estender a análise realizada para outras redes existentes no Flickr, tais como a rede de fotos favoritas, onde usuários adicionam fotos de outros à sua lista de favoritas, rede formada pelos comentários que usuários fazem sobre fotos de outros, entre outras. Complementarmente, a caracterização das redes existentes em outros sistemas de compartilhamento de fotos, como o deviantART [3] e o Zooomr [10], possibilitaria identificar semelhanças ou diferenças nos padrões de comportamento dos usuários. Uma outra direção a ser explorada é uma caracterização mais detalhada do tráfego para o Flickr, abordando um estudo acerca da evolução da popularidade de fotos e a relação com os diferentes tipos de interação.

Referências Bibliográficas

[1] The alexa web site. http://www.alexa.com.

[2] Del.icio.us. http://del.icio.us.

[3] DeviantART. http://www.deviantart.com.

[4] Facebook. http://www.facebook.com.

[5] Flickr. http://www.flickr.com.

[6] Flickr faq. http://www.flickr.com/help/faq.

[7] Orkut. http://www.orkut.com.

[8] Ruby programming language. http://www.ruby-lang.org.

[9] YouTube. http://www.youtube.com.

[10] Zooomr. http://www.zooomr.com.

[11] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[12] R. Albert, H. Jeong, and A. L. Barabasi. The diameter of the world wide web. *Nature*, 401:130–131, 1999.

[13] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[14] Susanne Boll. Multitube–where web 2.0 and multimedia could meet. *IEEE MultiMedia*, 14(1):9–13, 2007.

[15] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks (WOSN'08)*, August 2008.

[16] Birgitte Freiesleben de Blasio, Åke Svensson, and Fredrik Liljeros. Preferential attachment in sexual networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(26), 2007.

[17] Amir Dotan. A cross-cultural analysis of flickr users from peru, israel, iran, taiwan and the united kingdom. Master's thesis, City University, 2008.

[18] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66:035103, 2002.

[19] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA, 1999. ACM.

[20] Diego Garlaschelli and Maria I. Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93:268701, 2004.

[21] Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. Youtube traffic characterization: a view from the edge. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 15–28, New York, NY, USA, 2007. ACM.

[22] Luiz Henrique Gomes. Análise e modelagem do comportamento dos spammers e dos usuários legítimos em redes email. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil, May 2006.

[23] Luíz Henrique Gomes, Rodrigo B. Almeida, Luis M. A. Bettencourt, Virgílio A. F. Almeida, and Jussara M. Almeida. Comparative graph theoretical characterization of networks of spam. In *CEAS 2005 - Second Conference on Email and Anti-Spam*, California, USA, 2005.

[24] Raj Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* John Wiley and Sons, 1991.

[25] Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 631–640, New York, NY, USA, 2007. ACM.

[26] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM Press.

[27] S. H. Lee, P. J. Kim, and H. Jeong. Statistical properties of sampled networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 73(1 Pt 2), January 2006.

[28] Kristina Lerman and Laurie Jones. Social browsing on flickr. In *International Conference on Weblogs and Social Media*, Mar 2007.

[29] Marcelo Maia, Virgílio Almeida, and Jussara Almeida. Identifying user behavior in online social networks. In *Proceedings of the ACM Workshop on Social Network Systems. Co-located with EuroSys Conference*, Apr 2008.

[30] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[31] Andrew Miklas, Kiran Gollu, Kelvin Chan, Stefan Saroiu, Krishna Gummadi, and Eyal de Lara. Exploiting social interactions in mobile systems. pages 409–428. 2007.

[32] Alan Mislove, Krishna P. Gummadi, and Peter Druschel. Exploiting social networks for internet search. In *Proceedings of the 5th Workshop on Hot Topics in Networks (HotNets'06)*, November 2006.

[33] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the flickr social network. In *Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks (WOSN'08)*, August 2008.

[34] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM.

[35] Radu Negoescu. An analysis of the network of flickr. Technical report, Laboratory of Nonlinear Systems - LANOS, School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Switzerland, July 2007.

[36] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701+, October 2002.

[37] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[38] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, Sep 2003.

[39] M.E.J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, 2001.

[40] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach.* Cambridge University Press, New York, NY, USA, 2004.

[41] Nicolas Pissard and Christophe Prieur. Thematic vs. social networks in web 2.0 communities: A case study on flickr groups. In *Proceedings of the Algotel Conference*, Mai 2007.

[42] Roelof van Zwol. Flickr: Who is looking? In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 184–190, Washington, DC, USA, 2007. IEEE Computer Society.

[43] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. Sybilguard: defending against sybil attacks via social networks. In *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 267–278, New York, NY, USA, 2006. ACM.

[44] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.