Capítulo 2

Redes Sociais Online: Técnicas de Coleta e Abordagens de Medição

Fabrício Benevenuto
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Belo Horizonte - Brasil
fabricio@dcc.ufmg.br

Resumo

Redes sociais online se tornaram extremamente populares e vêm causando uma nova onda de aplicações na Web. Associado a esse crescimento, redes sociais estão se tornando um tema central em pesquisas de diversas áreas. Este trabalho oferece uma introdução ao pesquisador que pretende explorar esse tema. Inicialmente, apresentamos as principais características das redes sociais mais populares atualmente. Em seguida, discutimos as principais métricas e tipos de análises utilizadas no estudo dos grafos que formam a topologia das redes sociais. Finalmente, sumarizamos as principais abordagens utilizadas para se obter dados de redes sociais online e discutimos trabalhos recentes que utilizaram essas técnicas.

Abstract

Online social networks became extremely popular and are increasingly causing a new wave of applications on the Web. Associated to this popularity growth, online social networks are becoming a key theme in several research areas. This work offers an introduction to the researcher that aims at exploring this theme. Initially, we present the main characteristics of current online social network sites. Then, we discuss the main metrics and types of analysis used to study the graphs that represent the social network topologies. Finally, we summarize the main approaches used to collect online social networks and discuss recent work that used these approaches.

2.1. Introdução

Desde seu início, a Internet tem sido palco de uma série de novas aplicações, incluindo a Web, aplicações par-a-par e email. Atualmente, a Web vem experimentando

uma nova onda de aplicações associada à proliferação das redes sociais e ao crescimento de mídia digital. Várias redes sociais online (OSNs - *Online Social Networks*) surgiram, incluindo redes de profissionais (ex., LinkedIn), redes de amigos (ex., MySpace, Facebook, Orkut), e redes para o compartilhamento de conteúdos específicos tais como mensagens curtas (ex., Twitter), diários e blogs (ex., LiveJournal), fotos (ex., Flickr), e vídeos (ex., YouTube).

Redes sociais online têm atraído milhões de usuários. De acordo com a *Nielsen Online* [80], mídia social passou na frente de email como a atividade online mais popular. Mais de dois terços da população online global visita ou participa de redes sociais e blogs. Como comparação, se o Facebook fosse um país, seus 500 milhões de usuários registrados colocariam o Facebook como terceiro país mais populoso do mundo [5]. Tanta popularidade está associada a uma funcionalidade comum de todas as redes sociais online que é permitir que usuários criem e compartilhem conteúdo nesses ambientes. Este conteúdo pode variar de simples mensagens de texto comunicando eventos do dia-a-dia até mesmo a conteúdo multimídia, como fotos e vídeos. Como conseqüência, as estatísticas sobre conteúdo gerado pelos usuários nesses sítios Web são impressionantes. O Facebook compartilha mais de 60 bilhões de fotos, que ocupam mais de 1.5 PB de espaço [10]. A quantidade de conteúdo que o YouTube armazena em 60 dias seria equivalente ao conteúdo televisionado em 60 anos, sem interrupção, pelas emissoras NBC, CBS e ABC juntas [12]. De fato, o YouTube foi acessado por mais de 100 milhões de usuários apenas em Janeiro de 2009 [1], com uma taxa de upload de 10 horas de vídeo por minuto [14].

Apesar de tanta popularidade e da enorme quantidade de conteúdo disponível, o estudo de redes sociais ainda está em sua infância, já que esses ambientes estão experimentando novas tendências e enfrentando diversos novos problemas e desafios. A seguir sumarizamos alguns elementos motivadores para o estudo de redes sociais online.

- **Comercial**: Com usuários passando muito tempo navegando em redes sociais online, esses sítios Web tem se tornado um grande alvo para propagandas. De fato, em 2007, 1,2 bilhões de dólares foram gastos em propagandas em redes sociais online no mundo todo, e é esperado que este número triplique até 2011 [21]. Além disso, redes sociais online são lugares onde usuários compartilham e recebem uma grande quantidade de informação, influenciando e sendo influenciado por amigos [38]. Conseqüentemente, redes sociais online estão se tornando cada vez mais um alvo de campanhas políticas [51] e de diversas outras formas de marketing viral, onde usuários são encorajados a compartilhar anúncios sobre marcas e produtos com seus amigos [64].

- **Sociológica**: No passado o estudo de redes sociais era um domínio de sociólogos e antropólogos, quando ferramentas típicas para se obter dados eram entrevistas [88]. Como conseqüência, muitos desses esforços foram realizados baseados em amostras de dados pequenas e pouco representativas. Com o surgimento de redes sociais online, surgiu a oportunidade de estudos nesse tema com o uso de grandes bases de dados. Sistemas como Facebook, Twitter, Orkut, MySpace e YouTube possuem milhões de usuários registrados e bilhões de elos que os conectam. Redes sociais permitem o registro em larga escala de diversos aspectos da sociologia e da natureza humana relacionados à comunicação e ao comportamento humano. Além

disso, redes sociais online vem funcionando como um novo meio de comunicação e modificando aspectos de nossas vidas. Redes sociais online permitem que as pessoas interajam mais, permite que pessoas mantenham contato com amigos e conhecidos e permitem indivíduos se expressar e serem ouvidas por uma audiência local ou até mesmo global.

- **Melhorias dos sistemas atuais**: Assim como qualquer sistema Web, redes sociais online são vulneráveis a novas tendências e estão sujeitas a verem seus usuários rapidamente se mudar para outros sistema sem aviso prévio. Por exemplo, o MySpace experimentou um crescimento exponencial no número de usuários seguido de uma forte queda depois de abril de 2008 devido a um aumento no número de usuários do Facebook [85]. No início, o Orkut cresceu rapidamente em diversos lugares, mas sua popularidade foi concretizada somente em alguns países, dos quais o Brasil é o país com maior número de usuários registrados [11]. Várias razões podem explicar este tipo de fenômeno, incluindo a interface e novas utilidades do sistema, problemas de desempenho e características dos usuários, etc. Finalmente, o grande volume de dados disponível em redes sociais online abre um novo leque para a pesquisa relacionada à recuperação de conteúdo, onde estratégias de busca e recomendação de usuários e conteúdo são cada vez mais importante.

  Outro aspecto importante está relacionado ao tráfego gerado pelas redes sociais online. Intuitivamente, existe uma diferença crucial entre publicar conteúdo na Web tradicional e compartilhar conteúdo através de redes sociais online. Quando as pessoas publicam algum conteúdo na Web, elas tipicamente fazem isso para que todos os usuários da Internet, em qualquer lugar, possam acessar. Por outro lado, quando usuários publicam conteúdo em redes sociais online, eles geralmente possuem uma audiência em mente, geralmente, seus amigos. Algumas vezes, a audiência é explicitamente definida por um usuário ou pela política do sistema. Conseqüentemente, redes sociais online constituem uma classe única de aplicações com potencial para remodelar o tráfego da Internet com sua popularidade crescente. Estudar aspectos de sistemas relacionados a redes sociais pode ser de grande importância para a próxima geração da infra-estrutura da Internet e sistemas de distribuição de conteúdo [59, 81].

- **Segurança e conteúdo indesejável**: Redes sociais estão cada vez mais se tornando alvo de usuários maliciosos ou oportunistas que enviam propagandas não solicitadas, spam, e até mesmo *phishing*. O problema se manifesta de diversas maneiras, como postagens em listas de vídeos mais populares contendo spam [29, 26], spam no Twitter [24], conteúdo com metadados que não descrevem o conteúdo [27], etc. Conteúdo não solicitado consome a atenção humana, talvez o recurso mais importante na era da informação. O ruído e o distúrbio causados por alguns usuários reduzem a efetividade da comunicação online e é um problema cada vez maior.

Redes sociais compõem ambientes perfeitos para o estudo de vários temas da computação, incluindo sistemas multimídia e interação humano-computador. Além disso, por permitir que usuários criem conteúdo, redes sociais vêm se tornando um tema chave em

pesquisas relacionadas à organização e tratamento de grandes quantidades de dados, além de constituírem um ambiente ideal para extração de conhecimento e aplicação de técnicas de mineração de dados. Neste mini-curso apresentamos uma visão geral sobre redes sociais, oferecendo uma base necessária ao pesquisador que pretende explorar o tema. Inicialmente, apresentamos as principais características das redes sociais mais populares atualmente. Em seguida, discutimos as principais métricas e tipos de análises utilizadas no estudo dos grafos que formam a topologia das redes sociais. Finalmente, sumarizamos as principais abordagens utilizadas para se obter dados de redes sociais online e discutimos trabalhos recentes que utilizaram essas técnicas.

2.2. Definições e Características de Redes Sociais Online

Esta seção apresenta uma visão geral sobre as redes sociais online, suas características em comum e os principais mecanismos de interação entre os usuários.

2.2.1. Definição

O termo rede social online é geralmente utilizado para descrever um grupo de pessoas que interagem primariamente através de qualquer mídia de comunicação. Conseqüentemente, baseado nessa definição, redes sociais online existem desde a criação da Internet. Entretanto, neste trabalho, nós utilizaremos uma definição um pouco mais restrita, adotada em trabalhos anteriores [33, 69]. Nós definimos uma rede social online como um serviço Web que permite indivíduos (1) construir perfis públicos ou semi-públicos dentro de um sistema, (2) articular uma lista de outros usuários com os quais compartilham conexões e (3) visualizar e percorrer suas listas de conexões e outras listas feitas por outros no sistema.

Com base nessa definição existem várias redes sociais online disponíveis na Web, que variam de acordo com seus propósitos. A tabela 2.1 sumariza os propósitos de várias redes sociais online populares. Uma lista atualizada e exaustiva de redes sociais online, com mais de 150 sítios Web, pode ser encontrada em [8].

Nome	**Propósito**	**URL**
Orkut	Amizades	http://www.orkut.com
Facebook	Amizades	http://www.facebook.com
MySpace	Amizades	http://www.myspace.com
Hi5	Amizades	http://www.hi5.com
LinkedIn	Professionais	http://www.linkedin.com
YouTube	Compartilhamento de vídeos	http://www.youtube.com
Flickr	Compartilhamento de fotos	http://www.flickr.com
LiveJournal	Blogs e diários	http://www.livejournal.com
Digg	Compartilhamento de (*bookmarks*)	http://digg.com
Twitter	Troca de mensagens curtas	http://twitter.com
Last FM	Compartilhamento de rádio/músicas	http://www.last.fm

Tabela 2.1. Algumas redes sociais online populares

2.2.2. Elementos das redes sociais online

A seguir discutimos várias funcionalidades oferecidas pelas redes sociais atuais. O objetivo desta seção não é prover uma lista completa e exaustiva de funcionalidades, mas apenas descrever as mais relevantes.

- **Perfis dos usuários**: Redes sociais online possuem muitas funcionalidades organizadas ao redor do perfil do usuário, na forma de uma página individual, que oferece a descrição de um membro. Perfis podem ser utilizados não só para identificar o indivíduo no sistema, mas também para identificar pessoas com interesses em comum e articular novas relações. Tipicamente, perfis contêm detalhes demográficos (idade, sexo, localização, etc.), interesses (passatempos, bandas favoritas, etc.), e uma foto. Além da adição de texto, imagens e outros objetos criados pelo usuário, o perfil na rede social também contém mensagens de outros membros e listas de pessoas identificadas como amigos na rede. Perfis são geralmente acessíveis por qualquer um que tenha uma conta na rede social online ou podem ser privados, de acordo com as políticas de privacidades definidas pelo usuário.

  Recentemente, Boyd e colaboradores [32] mostraram que para a maior parte dos usuários de redes sociais online, existe uma forte relação entre a identidade do indivíduo real e seu perfil na rede social.

- **Atualizações**: Atualizações são formas efetivas de ajudar usuários a descobrir conteúdo. Para encorajar usuários a compartilhar conteúdo e navegar por conteúdo compartilhado por amigos, redes sociais online geralmente fazem as atualizações imediatamente visíveis aos amigos na rede social. Burke e colaboradores [37] mostram que atualizações motivam contribuições de novos usuários no sistema. Eles conduziram um estudo utilizando dados de 140,000 novos usuários do Facebook para determinar que atividades como atualizações são vitais para que novos usuários contribuam para o sistema. Como atualizações podem receber comentários de outros usuários, atualizações também são formas especiais de comunicação em redes sociais online.

- **Comentários**: A maior parte dos sítios de redes sociais permite que usuários comentem em conteúdo compartilhado. Alguns sistemas também permitem que usuários comentem em outros perfis de usuários. Comentários são um meio primordial de comunicação em redes sociais online, que também expressam relações sociais [19, 44]. Como exemplo, vídeos podem receber comentários no YouTube, fotos podem receber comentários no Facebook, Flickr e Orkut, usuários do LiveJournal podem postar comentários em blogs, etc.

- **Avaliações**: Em muitas redes sociais online, o conteúdo compartilhado por um usuário pode ser avaliado por outros usuários. Avaliações podem aparecer em diferentes níveis de granularidade e formas. No Facebook, usuários podem apenas gostar de uma postagem, clicando no botão "*I like this*". Com mais de 500 milhões de usuários registrados, cada usuário do Facebook avalia 9 objetos a cada mês em média [5]. No YouTube, vídeos podem ser avaliados com até 5 estrelas, de forma similar à avaliação empregada na categorização de hotéis. O YouTube ainda provê

uma avaliação binária (positiva ou negativa) para os comentários recebidos por vídeos, na tentativa de filtrar comentários ofensivos ou com alguma forma de spam.

Avaliações de conteúdo são úteis de várias formas. Como exemplo, elas são importantes para sistemas como o YouTube para ajudar usuários a encontrar e identificar conteúdo relevante. Avaliações podem ainda ajudar administradores a identificar conteúdo de baixa qualidade ou mesmo conteúdo inapropriado. Além disso, avaliações podem ser utilizadas para outras finalidades no sistema, como conteúdo em destaque, sistemas de recomendação, etc. Uma rede social online que coloca as avaliações dos usuários no centro do sistema é o Digg. O Digg permite que usuários avaliem URLs, notícias ou estórias e utiliza aquelas mais votadas para expor o conteúdo mais popular [63].

- **Listas de Favoritos**: Várias aplicações sociais utilizam listas de favoritos para permitir usuários selecionar e organizar conteúdo. Listas de favoritos ajudam usuários a gerenciar seu próprio conteúdo e podem ser úteis para recomendações sociais. Como exemplo, usuários podem manter listas de vídeos favoritos no YouTube e fotos favoritas no Flickr. Nesses sistemas, usuários podem navegar na lista de favoritos de outros usuários para buscar novos conteúdos [40]. Conseqüentemente, listas de favoritos também funcionam como uma forma de descoberta de conteúdo e propagação de informação. Sistemas como o Orkut e o Twitter também provêem listas de favoritos (fãs).

- **Listas de Top**: Tipicamente, redes sociais que colocam algum tipo de conteúdo de mídia como elemento central do sistema, como o YouTube, provêem listas de conteúdo mais popular ou usuários mais populares. Geralmente, essas listas são baseadas em avaliações ou outras estatísticas do sistema relativas ao conteúdo (ex. número de visualizações, avaliações, número de comentários) ou relativas aos usuários (ex. número de assinantes).

- **Metadados**: Uma das novas tendências da Web 2.0 é permitir aos usuários criar conteúdo livremente e associar metadados ao conteúdo [45]. Em redes sociais online como o YouTube e o Flickr, usuários tipicamente associam metadados como título, descrição e tags ao conteúdo compartilhado. Metadados são essenciais para recuperação de conteúdo em redes sociais online.

2.3. Teoria de redes complexas

Redes sociais online são inerentemente redes complexas e vários estudos vêm estudando características de redes sociais online utilizando como base teorias existentes [15, 70, 46, 60, 65, 23, 28]. De fato, o estudo de redes complexas cobre um grande número de áreas e sua teoria tem sido utilizada como ferramenta para entender vários fenômenos, incluindo o espalhamento de epidemias [71], propagação de informação [90], busca na Web [36], e consequências de ataques a redes de computadores [17]. A seguir, várias propriedades estatísticas e métricas comumente utilizadas para analizar e classificar rede complexas são apresentadas na seção 2.3.1. As seções 2.3.2 e 2.3.3 discutem propriedades de redes small-world e redes power-law. Uma revisão detalhada sobre métricas e teoria de redes complexas pode ser encontrada na referência [76].

2.3.1. Métricas para o estudo de redes

Uma rede é um conjunto de ítens, que chamamos de vértices ou nodos, com conexões entre si, chamadas de arestas. Em outras palavras, uma rede nada mais é do que um grafo. Existem diversas propriedades estatísticas e métricas que caracterizam a estrutura dessas redes, discutidas a seguir. Assume-se que o leitor tenha um conhecimento sobre a terminologia utilizada em teoria de grafos.

2.3.1.1. Grau dos vértices

Uma característica importante sobre a estrutura de uma rede é a distribuição dos graus de seus vértices. Consequentemente, uma métrica comum utilizada para comparar redes é o expoente α obtido através da regressão linear de uma distribuição de lei de potência. Valores típicos para o expoente α ficam entre 1.0 e 3.5 [49]. Para redes direcionadas, é comum analisar o grau dos nodos em ambas as direções, o grau de entrada e o grau de saída.

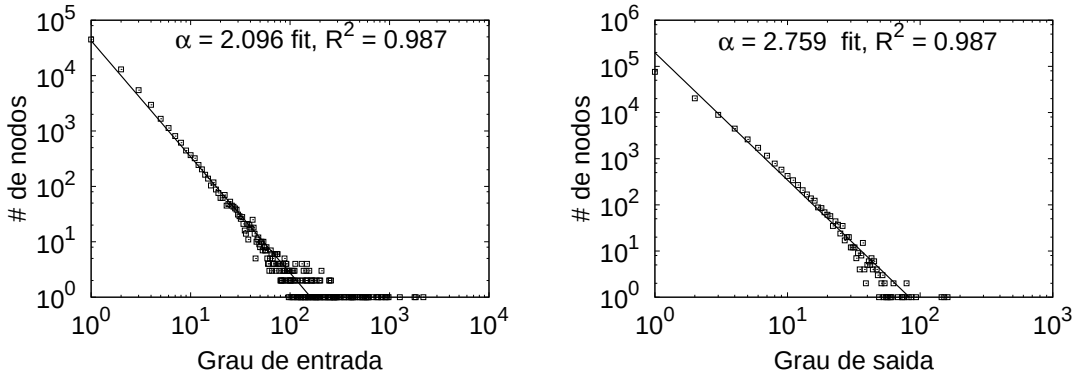


Figura 2.1. Grau de entrada e de saída de um grafo de interações entre usuários através de vídeos do YouTube

Como exemplo, a figura 2.1 mostra a distribuição dos graus de entrada e saída para um grafo formado de interações entre usuários de vídeos do YouTube, utilizado nas referências [23, 28]. Note que a curva da regressão linear, utilizada para se calcular o expoente α também é exibida nesses gráficos. Ferramentas como o Gnuplot [6] e o Matlab [9] podem ser utilizados para se realizar a regressão e calcular o valor de α. Para verificar a acurácia da regressão, é comum medir o fator R^2 [86], sendo que se o valor de R^2 for 1, significa que não há diferenças entre o modelo e os dados reais.

2.3.1.2. Coeficiente de clusterização

O coeficiente de clusterização de um nodo i, $cc(i)$, é a razão entre do número de arestas existentes entre os vizinhos de um nodo e o total de arestas possíveis entre os vizinhos de i. Como exemplo a Figura 2.2 mostra o valor do coeficiente de clusterização para o nodo escuro em três cenários diferentes. No primeiro, todos os vizinhos do nodo estão conectados entre si e, conseqüentemente, o cc do nodo é 1. No segundo cenário,

existe apenas 1 aresta entre os vizinhos do nodo dentre as 3 possíveis, deixando o nodo com cc=1/3. No último cenário, não há nenhuma aresta entre os vizinhos do nodo escuro e, portanto, o cc do nodo é 0.

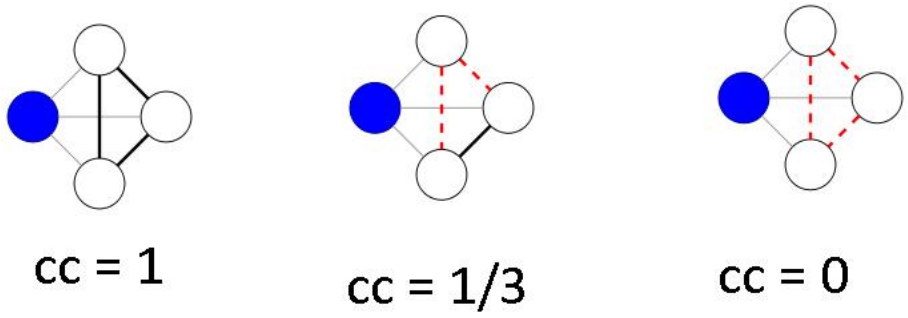


Figura 2.2. Cálculo do coeficiente de clusterização de um nodo em três cenários diferentes

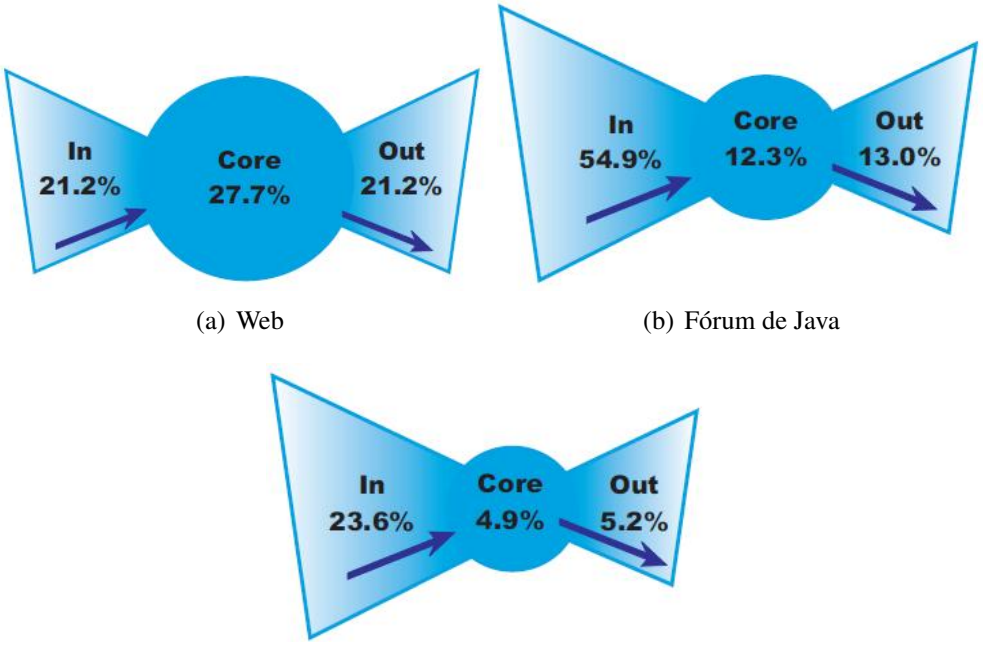
Podemos notar que o coeficiente de clusterização funciona como uma medida da densidade de arestas estabelecidas entre os vizinhos de um nodo. O coeficiente de clusterização de uma rede, *CC*, é a média do coeficiente de clusterização de todos os nodos.

2.3.1.3. Componentes

Um componente em um grafo é um conjunto de nodos, onde cada nodo possui um caminho para todos os outros nodos do conjunto. Um componente é chamado de fortemente conectado (SCC - *Strongly Connected Component*) quando o caminho entre todos os nodos do conjunto é direcionado. Por outro lado, dizemos que um componente é fracamente conectado (WCC - *Weakly Connected Component*) se o caminho é não direcionado.

Um trabalho que se tornou referência no estudo de componentes em redes complexas aborda a estrutura da Web (nodos são páginas Web e arestas são elos existentes entre as páginas) [36]. Os autores propõem um modelo que representa como os componentes no grafo da Web se relacionam. Este modelo, aplicado somente em grafos direcionados, possui um componente central que é o SCC, chamado também de *core*, e outros grupos de componentes que podem alcançar o SCC ou serem alcançados por ele. O modelo ficou conhecido como *bow tie* [36], pois a figura que ilustrada o modelo lembra uma gravata borboleta.

Este modelo tem sido utilizado por outros estudos como forma de comparar a organização dos componentes de um grafo direcionado [93, 28]. A figura 2.3 compara a estrutura dos componentes de três diferentes grafos utilizando-se o modelo *bow tie*. O componente central, *core*, das figuras corresponde à fração dos nodos do grafo que fazem parte do SCC. O componente *in* contém os nodos que apontam para algum nodo do *core*, mas não são apontados por nodos desse componente. Finalmente, o componente *out* corresponde aos nodos que são apontados por nodos do *core*.



(a) Web

(b) Fórum de Java

(c) Interações com vídeo

Figura 2.3. Estrutura dos componentes da Web [36], do Fórum de Java [93] e do grafo de interações através de vídeos [28]

2.3.1.4. Distância média e diâmetro

A distância média de um grafo é o número médio de passos entre todos os caminhos mínimos existentes para todos os nodos do grafo. Normalmente, a distância média é computada apenas no SCC para grafos direcionados ou no WCC para grafos não direcionados, já que não existe caminho entre nodos localizados em componentes diferentes. Outra métrica relacionada é o diâmetro do grafo. O diâmetro é definido como a distância do maior caminho mínimo existente no grafo (em geral, também computado somente no WCC ou no SCC).

2.3.1.5. Assortatividade

De acordo com Newman [75], assortatividade é uma medida típica de redes sociais. Uma rede exibe propriedades assortativas quando nodos com muitas conexões tendem a se conectar a outros nodos com muitas conexões. Sendo assim, definimos como $knn(k)$ como o grau médio de todos os vizinhos dos nodos com grau k. A assortatividade ou disassortatividade de uma rede é geralmente avaliando plotando-se o $knn(k)$ em função de k.

2.3.1.6. Betweenness

O Betweeness é uma medida relacionada à centralidade dos nodos ou de arestas na rede. O betweeness B de uma aresta é definido como o número de caminhos mínimos entre todos os pares de nodos em um grafo que passam pela aresta [78]. Se entre um par de nodos possui múltiplos caminhos mínimos entre eles, cada caminho recebe um peso de forma que a soma dos pesos de todos os caminhos seja 1. Conseqüentemente, o betweenness de uma aresta e pode ser expressado como

$$B(e) = \sum_{u \in V, v \in V} \frac{\sigma_e(u,v)}{\sigma(u,v)} \tag{1}$$

onde $\sigma(u,v)$ representa o número de caminhos mínimos entre u e v, e $\sigma_e(u,v)$ representa o número de caminhos mínimos entre u e v que incluem e. O betweenness de uma aresta indica a importância dessa aresta no grafo em termos de sua localização. Arestas com maior betweenness estão em mais caminhos mínimos e, portanto, são mais importantes para a estrutura do grafo.

De forma similar, o betweenness pode ser computado para um nodo ao invés de uma aresta. Neste caso, a medida do betweenness mede o número de caminhos mínimos que passam por nodo. Nodos que possuem muitos caminhos mínimos que passam por eles possuem maior betweenness do que os outros que não possuem.

2.3.1.7. Reciprocidade

Uma forma interessante de se observar a reciprocidade em um grafo direcionado é medindo a probabilidade de um nodo ter uma aresta apontando para ele para cada nodo que ele aponta. Em outras palavras, a reciprocidade ($R(x)$) é dada por:

$$R(x) = \frac{|O(x) \cap I(x)|}{|O(x)|} \tag{2}$$

onde $O(x)$ é o conjunto de nodos que recebem uma aresta de um usuário x e $I(x)$ é o conjunto de nodos que apontam x através de arestas direcionadas.

Outra métrica interessante de ser observar é o coeficiente de reciprocidade ρ, uma métrica que captura a reciprocidade da interações em toda a rede [52]. O coeficiente de reciprocidade ρ é definido pelo coeficiente de correlação entre entidades de uma matriz de adjacência de um grafo direcionado ($a_{ij} = 1$ if há uma aresta de i para j, senão $a_{ij} = 0$):

$$\rho = \frac{\sum_{i \neq j}(a_{ij} - \bar{a})(a_{ji} - \bar{a})}{\sum_{i \neq j}(a_{ij} - \bar{a})^2}, \tag{3}$$

onde o valor médio $\bar{a} = \sum_{i \neq j}(a_{ij}/N(N-1)$ e N é o número de usuários no grafo.

O coeficiente de reciprocidade indica se o número de arestas recíprocas na rede é maior ou menor do que o de uma rede aleatória. Se o valor ρ é maior do que 0, a rede é recíproca; caso contrário, anti-recíproca.

2.3.1.8. PageRank

O PageRank é um algoritmo interativo que assinala um peso numérico para cada nodo, com o propósito de medir sua importância relativa dentro dos nodos do grafo. O algoritmo foi inicialmente proposto por Brin and Page [35] para ordenar resultados de busca do protótipo do Google. A intuição por trás do PageRank é que uma página Web é importante se existem muitas páginas apontando para ela ou se existem páginas importantes apontando para ela. A equação que calcula o PageRank (PR) de um nodo é definida da seguinte forma:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{4}$$

onde u representa um nodo. $B(u)$ é é o conjunto de páginas que apontam para u. $PR(u)$ e $PR(v)$ são os valores do PageRank para os nodos u e v, respectivamente. N_v denomina o número de arestas que saem do nodo v, e o parâmetro d é um fator que pode ter valor entre 0 e 1.

O algoritmo do PageRank tem sido aplicado em outros contextos, como por exemplo, para encontrar usuários experientes em fóruns especializados [93] e usuários influentes no Twitter [91, 61]. Além disso, existem outras modificações do PageRank com propósitos específicos, como por exemplo, a detecção de spam na Web [56].

2.3.2. Redes small-world

O conceito de redes small-world ficou bastante conhecido com o famoso experimento de Milgram [68]. Seu experimento consistiu de um grupo de voluntários tentando enviar uma carta para uma pessoa alvo através de outras pessoas que eles conheciam. Milgram enviou cartas a várias pessoas. As cartas explicavam que ele estava tentando atingir uma pessoa específica final em uma cidade dos EUA e que o destinatário deveria repassar a carta para alguém que ele achasse que poderia levar a carta o mais próximo do seu destino final (ou entregá-la diretamente, caso o destinatário final fosse uma pessoa conhecida). Antes de enviar a carta, entretanto, o remetente adicionava seu nome ao fim da carta, para que Milgram pudesse registrar o caminho percorrido pela carta. Das cartas que chegaram com sucesso ao destino final, o número médio de passos requerido para o alvo foi 6, resultado que ficou conhecido como o princípio dos *seis graus de separação*.

Em termos das propriedades das redes sociais que discutimos, uma rede pode ser considerada small-world se ela tiver duas propriedades básicas: coeficiente de clusterização alto e um pequeno diâmetro [89]. Estas propriedades foram verificadas em várias redes como a Web [18, 36], redes de colaboração científica [74, 77] (pesquisadores são nodos e arestas ligam co-autores de artigos), atores de filmes [20] (atores são nodos e arestas ligam atores que participaram do mesmo filme), e redes sociais on-

line [15, 70, 46, 65, 23, 28]. Em particular, Mislove e colaboradores [70] verificaram propriedades small-world em quatro redes sociais online: LiveJournal, Flickr, Orkut, e YouTube.

2.3.3. Redes power-law

Redes power-law consistem de grafos cuja distribuição de grau segue uma distribuição de lei de potência. Em outras palavras, nessas redes a probabilidade que um nodo tenha grau k é proporcional a $k^{-\alpha}$ para $\alpha > 1$. Várias redes reais mostram distribuições de grau que seguem distribuições de lei de potência, incluindo a topologia da Internet [50], a Web [22], e redes neurais [34].

Redes livres de escala (scale free) são classes de redes que seguem leis de potência, onde os nodos de grau alto tendem a se conectar a outros nodos de grau alto. Barabási e colaboradores [22] propuseram um modelo para gerar redes livre de escala, introduzindo o conceito de conexão preferencial (*preferential attachment*). O modelo diz que a probabilidade de um nodo se conectar a outro nodo é proporcional ao seu grau. Os autores do modelo ainda mostraram que, sob certas circunstâncias, este modelo produz redes que seguem leis de potência. Mais recentemente, Li e colaboradores [66] criaram uma métrica para medir se uma rede é livre de escala ou não, além de prover uma longa discussão sobre o tema.

2.4. Técnicas de Coleta de Dados

Em um passado recente, redes sociais eram um domínio de sociólogos e antropólogos, quando ferramentas típicas de coleta de dados de redes sociais eram pesquisas e entrevistas com pequenos grupos de usuários [88]. Com o surgimento das redes sociais online, a obtenção desse tipo de dados em larga escala se tornou possível e diversas áreas da computação começaram a realizar coletas de dados.

Diferentes áreas de pesquisa demandam diferentes tipos de dados e, por isso, existem várias formas de se obter dados de redes sociais online. A figura 2.4 apresenta possíveis pontos de coleta de dados, que variam desde entrevistas com os usuários até à instalação de coletores localizados em servidores Proxy ou aplicações. A seguir discutimos essas diferentes abordagens, bem como trabalhos que adotaram essas estratégias.

2.4.1. Dados dos usuários

Um método comum de se analisar o uso de redes sociais online consiste em conduzir entrevistas com usuários desses sistemas. Em particular, esta estratégia tem sido bastante empregada pela comunidade da área de interface homem-máquina [84, 57, 41, 31, 79], onde entrevistas estruturadas são as formas mais populares de obtenção de dados.

Como exemplo, através de entrevistas com usuários do Facebook, Joinson e seus colaboradores [57] identificaram várias razões pelas quais usuários utilizam o Facebook, incluindo conexão social, compartilhamento de interesses, compartilhamento e recuperação de conteúdo, navegação na rede social e atualização do seu estado atual. Chapman e Lahav [41] conduziram entrevistas e analisaram a navegação de 36 usuários de quatro nacionalidades diferentes para examinar diferenças etnográficas no uso de redes sociais online.

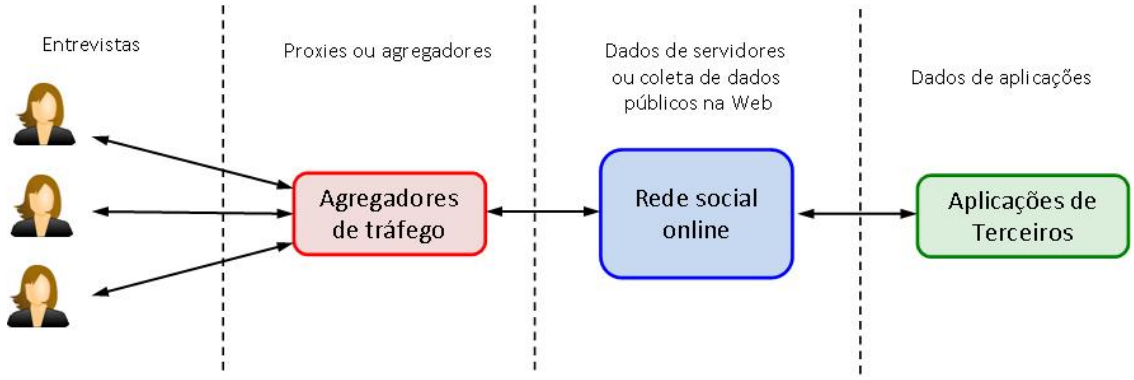


Figura 2.4. Possíveis pontos de coleta de dados

2.4.2. Dados de pontos intermediários

Existem duas técnicas comuns utilizadas para coletar dados de pontos de agregação de tráfego na rede. A primeira consiste em coletar os dados que passam por um provedor de serviços Internet (ISP) e filtrar as requisições que correspondem a acessos às rede sociais online. A segunda consiste em coletar dados diretamente de um agregador de redes sociais. A seguir, discutimos alguns trabalhos que fizeram o uso dessas estratégias.

2.4.2.1. Servidores proxy

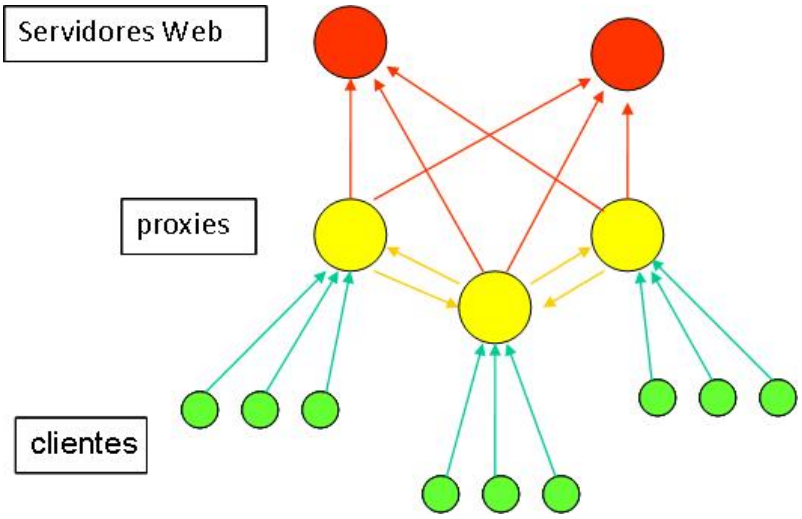


Figura 2.5. Exemplo de um servidor proxy intermediando o tráfego entre clientes e servidores

Coletar dados de um servidor proxy tem sido uma estratégia utilizada em vários estudos sobre o tráfego da Internet [47, 55, 67, 92]. A figura 2.5 ilustra como um servidor funciona como um agregador de tráfego de seus clientes. Tais servidores são utilizados para delimitar uma porção da rede, onde computadores estão localizados em uma mesma localização geográficas.

Dado o crescente interesse por vídeos na Web, alguns trabalhos recentes utilizaram

servidores proxy para obter dados do tráfego gerado por sistemas como o YouTube. Gill e colaboradores [53] caracterizaram o tráfego do YouTube coletando dados de um servidor proxy localizado na universidade de Calgary, no Canadá. Eles mostraram que requisições de HTTP GET, utilizadas para baixar conteúdo do servidor, correspondem a 99.87% do total das requisições que passam pelo servidor proxy. Eles ainda caracterizaram diversas medidas tais como a duração, a idade e a categoria dos vídeos. Mais recentemente, os mesmos autores caracterizam sessões no YouTube [54]. Eles mostraram que uma sessão típica possui cerca de 40 minutos e utilizaram esse valor para analisar medidas relativas à sessão dos usuários, tais como duração da sessão, tempo entre sessões e tipos de conteúdo transferidos em cada sessão. Finalmente, Zink e colaboradores [94] também estudaram tráfego coletado no proxy de uma universidade. Eles caracterizaram medidas típicos no tráfego do YouTube e, utilizando essas medidas, eles simularam abordagens de caches locais e globais para vídeos, bem como o uso de uma arquitetura P2P para distribuição de vídeos. De maneira geral, eles mostraram que essas abordagens poderiam reduzir tráfego significativamente e permitir acesso aos vídeos de forma mais rápida.

Em um trabalho recente, Schneider e colaboradores [82] extraíram dados de redes sociais online de um provedor de acesso a Internet e reconstruíram ações realizadas pelos usuários em suas navegações por diferentes redes sociais online. Em outras palavras, eles criam o que chamamos de *clickstream* [42] para redes sociais online, que captura cada passo da navegação dos usuários do ISP. Eles oferecem uma ampla discussão sobre a metodologia para reconstruir os acessos dos usuários e, com base nesses dados, eles analisaram as seqüências de requisições realizadas pelos usuários em redes sociais online como o Facebook.

2.4.2.2. Agregador de redes sociais

Agregadores de redes sociais são sistemas que permitem acesso a várias redes sociais simultaneamente, através de um portal único. Esses sistemas ajudam usuários que utilizam várias redes sociais online a gerenciar vários perfis de uma forma mais simples e unificada [58, 83]. Ao logar em um agregador de redes sociais online, usuários acessam suas contas através de uma interface única, sem precisar logar em cada rede social separadamente. Isto é feito através de uma conexão HTTP em tempo real realizada em duas etapas. A primeira etapa ocorre entre o usuário e o agregador de redes sociais e a segunda etapa ocorre entre o sistema agregador e as redes sociais. Agregadores tipicamente comunicam com redes sociais online através de APIs, como o OpenSocial [7], e todo o conteúdo é exibido através da interface do sistema agregador. A figura 2.6 descreve o esquema de interação entre os usuários, um sistema agregador e algumas redes sociais online. Através dessa interface, um usuário pode utilizar várias funcionalidades de cada rede social que ele está conectado, como checar atualizações de amigos, enviar mensagens e compartilhar fotos.

Recentente, nós utilizamos esta estratégia para obter dados de *clickstream* de redes sociais online [30]. Nós colaboramos com um agregador de redes sociais online e obtivemos dados da navegação dos usuários em 4 redes sociais online: Orkut, Hi5, MySpace e LinkedIn. A tabela 2.2 mostra o número de usuários, sessões e requisições HTTP para

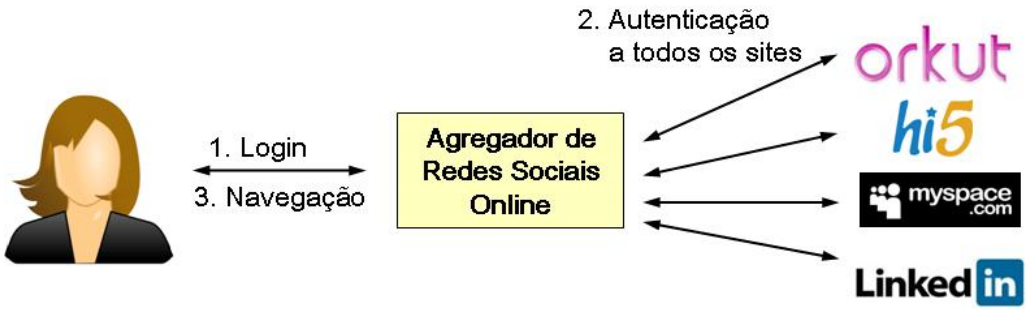


Figura 2.6. Ilustração de um usuário se conectando a múltiplas redes sociais online simultaneamente através de um portal agregador

cada uma dessas redes. Baseados nesses dados e em dados coletados do Orkut, nós examinamos o comportamento dos usuários nas redes sociais online, bem como características das interações entre os usuários através das várias atividades que eles realizam.

	# usuários	# sessões	# requisições
Orkut	36.309	57.927	787.276
Hi5	515	723	14.532
MySpace	115	119	542
LinkedIn	85	91	224
Total	37.024	58.860	802.574

Tabela 2.2. Sumário dos dados obtidos de um agregador de redes sociais

2.4.3. Dados de servidores de redes sociais online

Idealmente, servidores de redes sociais são os locais mais adequados para a coleta de dados. Entretanto, a maior parte desses sistemas evita prover dados, mesmo que anonimizados. Existem alguns poucos trabalhos que utilizaram dados obtidos diretamente de servidores de uma rede social. Chun e seus colaboradores [44] estudaram interações textuais entre os usuários do Cyworld, uma rede social bastante popular na Coréia do Sul, através de dados obtidos diretamente do servidor. Eles compararam a rede de amizades explicita com a rede criada por mensagens trocadas no livro de visitas do Cyworld, discutindo diversas similaridades e diferenças em termos da estrutura da rede. Baluja e colaboradores obtiveram dados dos servidores do YouTube. Eles utilizaram dados do histórico da navegação dos usuários do YouTube para criar um grafo onde cada vídeo é um nodo e arestas ligam vídeos freqüentemente vistos em seqüência. Baseados nesse grafo, eles criaram um mecanismo capaz de prover sugestões de vídeo personalizadas para os usuários do YouTube. Finalmente, Duarte e seus colaboradores [48] caracterizaram o tráfego em um servidor de blogs do UOL (www.uol.com.br). Recentemente, nós estudamos a navegação dos usuários em um servidor de vídeos do UOL, chamado UOL Mais [25].

Dada a dificuldade em se obter dados diretamente de servidores de redes sociais online, uma estratégia comum consiste em visitar páginas de redes sociais com o uso de uma ferramenta automática, que chamamos de *crawler* ou robô, e coletar sistematicamente informações públicas de usuários e objetos. Tipicamente, os elos entre usuários de

uma rede social online podem ser coletados automaticamente, permitindo que os grafos de conexões entre os usuários sejam reconstruídos. Essa estratégia tem sido amplamente utilizada em uma grande variedade de trabalhos, incluindo estudos sobre a topologia das redes sociais online [70, 16], padrões de acesso no YouTube [39] e interações reconstruídas através de mensagens trocadas pelos usuários [87]. A seguir discutimos vários aspectos relacionados à coleta de dados de redes sociais online.

2.4.3.1. Coleta por amostragem

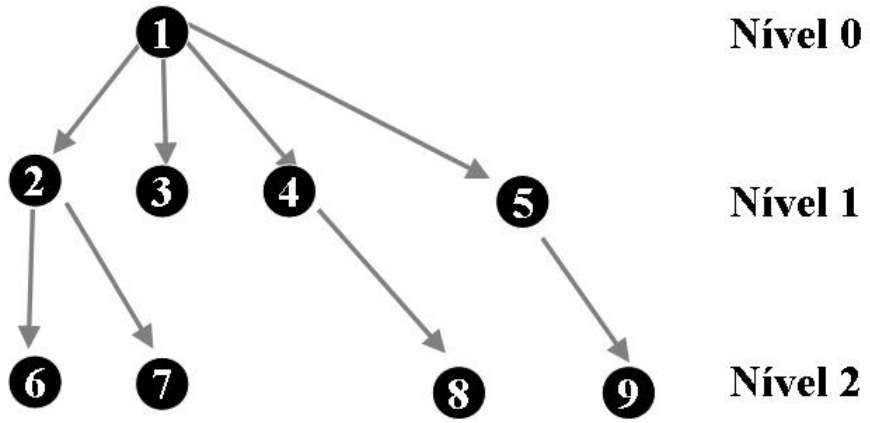


Figura 2.7. Exemplo de busca em largura em um grafo

Idealmente, é sempre mais interessante coletar o grafo inteiro de uma rede social online para evitar que a coleta seja tendenciosa a um grupo de usuários da rede. Entretanto, na maior parte das vezes, não há uma forma sistemática de se coletar todos os usuários de uma rede social online. Para esses casos é necessário coletar apenas parte do grafo. Uma abordagem comumente utilizada chama-se snowball. Snowball consiste em coletar o grafo de uma rede social online seguindo uma abordagem de busca em largura, como ilustra a figura 2.7. A coleta inicia-se a partir de nodo semente. Ao coletar a lista de vizinhos desse nodo, novos nodos são descobertos e então coletados no segundo passo, que só termina quando todos os nodos descobertos no primeiro passo são coletados. No passo seguinte todos os nodos descobertos no passo anterior são coletados, e assim sucessivamente. Recomenda-se o uso de um grande número de nodos sementes para evitar que a coleta não fique restrita a um pequeno componente do grafo.

O que caracteriza a coleta por snowball é a interrupção da coleta em um passo intermediário, antes que todos os nodos alcançáveis pela busca em largura sejam atingidos. Dependendo do objetivo da coleta, snowball pode ser uma estratégia viável. Por exemplo, se realizarmos 3 passos da coleta por snowball, podemos calcular o coeficiente de clusterização dos nodos semente. Entretanto, se quisermos computar o coeficiente de clusterização médio de toda a rede ou outras métricas como distribuição de graus, distância média, etc., a coleta por snowball pode resultar em números tendenciosos [62, 16].

Outra abordagem bastante difundida consiste em coletar o maior componente fracamente conectado (WCC) do grafo. A coleta de um componente inteiro pode ser realizada com uma estratégia baseada em um esquema de busca em largura ou busca em

profundidade. Quanto maior o número de sementes utilizadas maior a chance de se coletar o maior componente do grafo. Em trabalhos recentes, nós realizamos uma busca por palavras aleatórias no YouTube para verificar se o componente coletado era o maior componente [28, 23]. Como a maior parte dos usuários encontrados nessas buscas estavam no WCC do nosso grafo, os resultados desse teste sugeriram que o componente coletado era o maior WCC. Mislove e colaboradores [70] argumenta que o maior WCC de um grafo é estruturamente a parte mais interessante de ser analisada, pois é o componente que registra a maior parte das atividades dos usuários. Além disso, eles mostram que usuários não incluídos no maior WCC tendem fazer parte de um grande grupo de pequenos componentes isolados ou até mesmo totalmente desconectados.

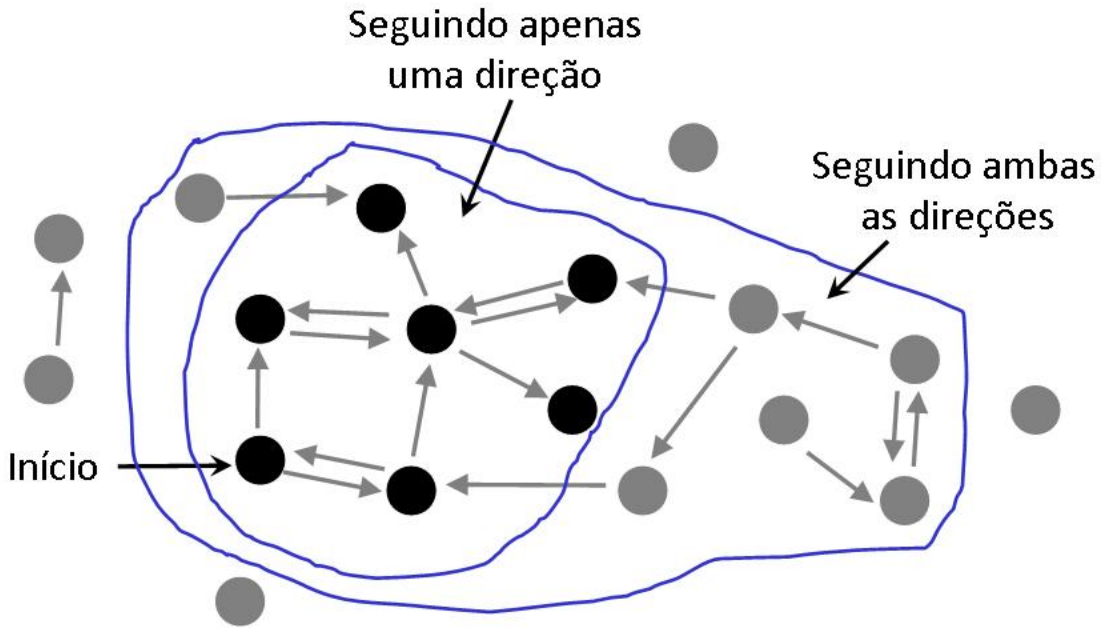


Figura 2.8. Exemplo de coleta do WCC em um grafo direcionado

Finalmente, é importante observar que para coletar o WCC em um grafo direcionado é necessário percorrer as arestas do grafo em ambas as direções. Algumas redes sociais online como o Twitter ou o Flickr, o conceito de amizade é direcionado. Ou seja, um usuário pode seguir outro, mas não ser seguido pelo mesmo. Se coletarmos o grafo seguindo as arestas em apenas uma direção, não necessariamente vamos coletar todo o WCC. A figura 2.8 mostra que o conjunto de nodos coletados quando seguimos as arestas em ambas as direções é maior do que quando seguimos apenas uma das direções. Em algumas redes não é possível percorrer o grafo em ambas as direções e, portanto, não é possível coletar o maior WCC. Essa limitação é típica de estudos que envolvem a coleta da Web [36]. Tipicamente, a Web é freqüentemente coletada seguindo apenas uma direção os elos entre as páginas, já que não é possível determinar o conjunto de páginas que apontam para uma página.

2.4.3.2. Coleta em larga escala

A coleta de grandes bases de dados de redes sociais online geralmente envolve o uso de coletores distribuídos em diversas máquinas. Isso acontece não só devido ao processamento necessário para tratar e salvar os dados coletados, mas também para evitar que servidores de redes sociais interpretem a coleta de dados públicos como um ataque a seus servidores.

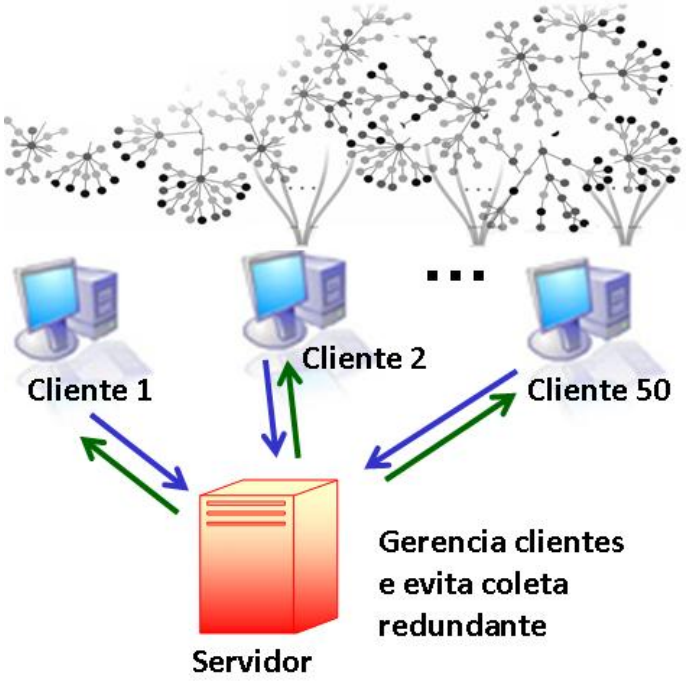


Figura 2.9. Exemplo de coleta feita de forma distribuída

Uma forma de se realizar tal coleta, conforme descrito em [43], está representada na figura 2.9. A estratégia consiste em utilizar (1) uma máquina mestre que mantém uma lista centralizada de usuários a serem visitados e (2) máquinas escravas, que coletam os dados, salvam esses dados, fazem um parser dos dados coletados e identificam novos usuários. Finalmente, as máquinas escravas retornam listas de novos usuários identificados à máquina mestre, que por sua vez, distribui novos usuários para as máquinas escravas.

2.4.3.3. Coleta por inspeção de IDs

Como discutido anteriormente, ao coletar uma rede social online o ideal é coletar toda a rede e não só uma porção dela. Em alguns sistemas como o MySpace e o Twitter é possível realizar uma coleta completa. Esses sistemas atribuem um identificador (ID) numérico e seqüencial para cada usuário cadastrado. Como novos usuários recebem um identificador seqüencial, podemos simplesmente percorrer todos os IDs, sem ter que verificar a lista de amigos desses usuários em busca de novos IDs para coletar.

Recentemente, nós realizamos uma coleta do Twitter seguindo essa estratégia. Nós pedimos aos administradores do Twitter para realizar uma coleta em larga escalae eles adicionaram 58 IPs de nossas máquinas em uma lista banca, com permissão para coletar. Cada uma das 58 máquinas, localizadas no *Max Planck Institute for Software Systems (MPI-SWS)*, na Alemanha[1], teve permissão para realizar uma taxa máxima de 20 mil requisições por hora ao Twitter. Utilizando a API do Twitter, nosso coletor investigou todos 80 milhões de IDs de forma seqüência, coletando todas as informações públicas sobre os usuários, bem como seus elos de seguidores e seguidos e todos os seus tweets. Dos 80 milhões de contas inspecionadas, nós encontramos cerca 55 milhões em uso. Isso acontece porque o Twitter apaga contas inativas por um período maior do que 6 meses. No total, coletamos cerca de 55 milhões de usuários, quase 2 bilhões de elos sociais e cerca de 1.8 bilhões de tweets. Ao inspecionar as listas de seguidores e seguidos coletadas, nós não encontramos nenhum ID acima do 80 milhões inspecionados, sugerindo que nós coletamos todos os usuários. Esses dados foram utilizados recentemente em dois trabalhos, um sobre detecção de spammers no Twitter [24] e o outro sobre medição de influência no Twitter [38].

Torkjazi e colaboradores [85] também usufruíram dos IDs sequênciais do MySpace para inspecionar o surgimento de novos usuários no sistema.

2.4.3.4. Utilizando APIs

Figura 2.10. Exemplo da API do Twitter: http://twitter.com/users/show/fbenevenuto.xml

No contexto de desenvolvimento Web, uma API é tipicamente um conjunto de tipos de requisições HTTP juntamente com suas respectivas definições de resposta. Em redes sociais é comum encontrarmos APIs que listam os amigos de um usuário, seus objetos, suas comunidades, etc. APIs são uma nova tendência na Web 2.0 onde sistemas não só oferecem seus serviços, mas permitem que outros acessem dados através de APIs.

APIs são perfeitas para a coleta de dados de redes sociais, pois oferecem os dados em formatos estruturados como XML e JSON. Vários sistemas como YouTube e Twitter

[1]Esta coleta foi realizada durante uma visita de 5 meses ao MPI-SWS

oferecem APIs. Como exemplo, a figura 2.10 mostra o resultado de uma busca por informações de um usuário no Twitter. Além dessa função, o Twitter ainda oferece várias outras funções em sua API. Com a API do Twitter é possível coletar 5000 seguidores de um usuário através de uma única requisição. A coleta dessa informação através do sítio Web convencional necessitaria centenas de requisições, visto que o Twitter só mostra alguns seguidores por página. Além disso, cada página viria com uma quantidade muito maior de informação desnecessária.

Vários sistemas possuem APIs, incluindo Twitter, Flickr, YouTube, Google Mapas, Yahoo Mapas, etc. Com tantas APIs existentes, é comum ver aplicações que utilizam duas ou mais APIs para criar um novo serviço, que é o que chamamos de Mashup. Uma interessante aplicação chamada Yahoo! Pipes [13], permite a combinação de diferentes APIs de vários sistemas para a criação automatizada de Mashups.

2.4.3.5. Ferramentas e bibliotecas

```
#!/usr/bin/perl

use LWP;

$ua = LWP::UserAgent->new();
$req = new HTTP::Request(GET =>
   "http://twitter.com/followers/ids/44446416.xml?page=1");
$content = $ua->request($req)->content;

print "$content";
```

Figura 2.11. Exemplo do uso da biblioteca LWP em Perl

Desenvolver um crawler pode ser uma tarefa bastante complicada devido à diversidade de formatos de páginas. Entretanto, coletar redes sociais online é, em geral, diferente de coletar páginas da Web tradicional. As páginas de uma rede social online são, em geral, bem estruturadas e possuem o mesmo formato, pois são geradas automaticamente, enquanto na Web tradicional as páginas podem ser criadas por qualquer pessoa em qualquer formato. Além disso, como cada indivíduo ou objeto em uma rede social, em geral, possui um identificador único, temos certeza sobre quais as informações obtivemos quando coletamos uma página.

```
#!/usr/bin/python

import urllib

req = urllib.urlopen("http://twitter.com/followers/ids/44446416.xml?page=1")
content = req.read()

print content
```

Figura 2.12. Exemplo do uso da biblioteca URLLIB em Python

Existem várias ferramentas que podem ser utilizadas para se coletar dados de redes sociais online. Como cada pesquisa requer um tipo de coleta e cada coleta de dados possui sua particularidade, desenvolver o próprio coletor pode ser necessário. A figura 2.11 mostra o uso da biblioteca LWP na linguagem Perl. Este código realiza a coleta dos seguidores de um usuário no Twitter através de sua API. De maneira similar, o código em Python da figura 2.12 utiliza a biblioteca URLLIB para realizar a mesma tarefa. O resultado da execução dos crawlers é a lista de seguidores de um usuário do Twitter em formato XML, como ilustra a figura 2.13.

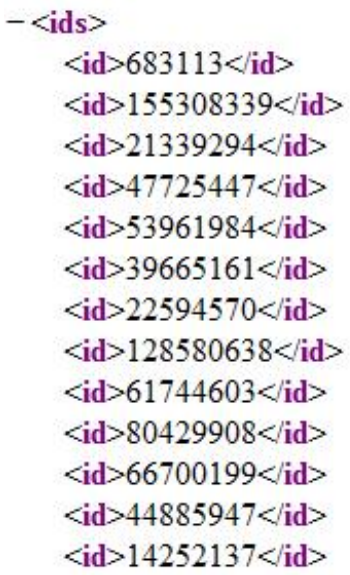


Figura 2.13. Resultado da execução dos crawlers em Perl e Python

2.4.3.6. Ética dos crawlers

Coletar dados da Web não é apenas útil para a pesquisa sobre redes sociais, mas constitui um importante passo para qualquer mecanismo de busca, como o Google [35]. Entretanto, crawlers ou robôs são ferramentas com habilidade para causar vários problemas para empresas relativos à coleta e disponibilização de conteúdo indevido. Sendo assim, foi convencionado através de um protocolo conhecido como robots.txt que sítios Web podem regulamentar o que pode e o que não pode ser coletado no sistema. Este método é bastante utilizado pelos administradores de sistemas para informar aos robôs visitantes quais diretórios de um sítio Web não devem ser coletados. O robots.txt nada mais é do que um arquivo que fica no diretório raiz dos sítios e contém regras para robôs específicos ou de uso geral para qualquer robô. Ao visitar um site, os robôs devem buscar primeiro pelo arquivo robots.txt e verificam suas permissões. Exemplos desses arquivos seriam:

http://portal.acm.org/robots.txt
http://www.google.com/robots.txt

http://www.globo.com/robots.txt
http://www.orkut.com.br/robots.txt
http://www.youtube.com/robots.txt
http://www.robotstxt.org/robots.txt

A seguir mostramos um exemplo simples de regra em um arquivo robots.txt. Essa regra restringe todos os crawlers de acessarem qualquer conteúdo no sistema.

```
User-agent: *
Disallow: /
```

É possível ainda especificar restrições a alguns robôs específicos ou restringir o acesso a alguns diretórios específicos. Como exemplo, o sítio Web www.globo.com oferece restrições para todos os robôs nos seguintes diretórios.

```
User-agent: *

Disallow: /PPZ/
Disallow: /Portal/
Disallow: /Java/
Disallow: /Servlets/
Disallow: /GMC/foto/
Disallow: /FotoShow/
Disallow: /Esportes/foto/
Disallow: /Gente/foto/
Disallow: /Entretenimento/Ego/foto/
```

No caso de coleta de redes sociais é importante verificar não só o arquivo robots.txt dos sistemas, mas também os termos de uso do sistema.

2.4.4. Dados de aplicações

Em uma tentativa bem sucedida de enriquecer a experiência dos usuários de redes sociais online, o Facebook realizou uma de suas maiores inovações: abriram sua plataforma para desenvolvedores de aplicações [4]. Com esta inovação, desenvolvedores são capazes de criar diferentes tipos de aplicações. Com o sucesso no Facebook, outros sistemas como Orkut e MySpace também adotoram essa estratégia. O tipo e número de aplicações nesses sistemas se tornou ilimitada com o novo modelo de API aberta implementado. O Facebook sozinho possui mais de 81,000 aplicações [2]. Empresas como a Zynga, especializadas em desenvolver essas aplicações, contam com mais de 80 milhões de usuários registrados em suas aplicações [2]. Uma grande lista de aplicações do Facebook pode ser encontrada na seguinte referência [3].

Uma estratégia para se obter dados de redes sociais online é através do desenvolvimento de aplicações sociais. A figura 2.14 mostra o funcionamento de uma aplicação terceirizada em execução em redes sociais online como o Facebook ou o Orkut. Aplicações são caracterizadas pela presença de um servidor da rede social intermediando toda

a comunicação entre o cliente e o servidor da aplicação. Tipicamente, o cliente envia a requisição ao servidor da rede social online, que repassa ao servidor da aplicação. Então, o servidor da aplicação manda de volta a resposta ao servidor da rede social, que repassa ao cliente [73].

Aplicações podem ser utilizadas para o estudo de interações entre os usuários que utilizam as aplicações e também podem ser úteis para coletar outras informações dos usuários. Como exemplo, aplicações podem pedir permissão aos usuários para acessar informações como lista de amigos e atividades executadas dentro de uma sessão.

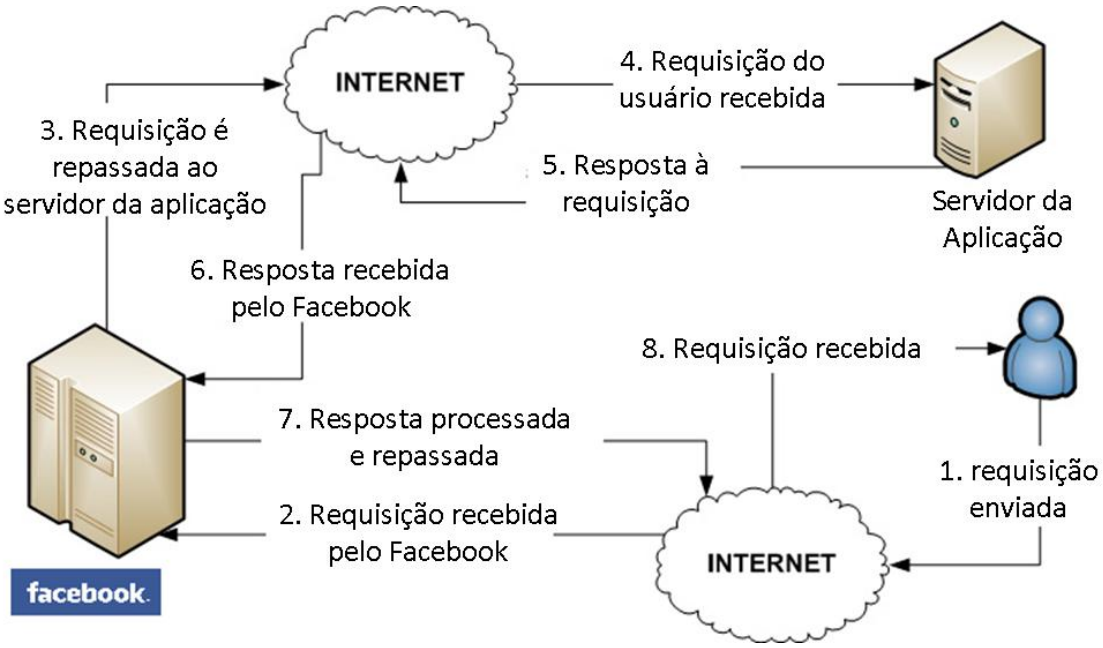


Figura 2.14. Funcionamento de aplicações no Facebook e no Orkut

Alguns trabalhos fizeram uso dessa estratégia de medição para estudar os usuários de redes sociais online. Nazir e colaboradores [72] analisaram características de aplicações no Facebook, desenvolvendo e lançando suas próprias aplicações. Em particular, eles estudaram a formação de comunidades online a partir de grafos de interação entre os usuários de suas aplicações. Mais recentemente, Nazir e colaboradores [73] estudaram várias características relacionadas ao desempenho de suas aplicações no Facebook.

2.5. Conclusões

Redes sociais se tornaram extremamente populares e parte do nosso dia a dia, causando o surgimento de uma nova onda de aplicações disponíveis na Web. A cada dia, grandes quantidades de conteúdo são compartilhadas e milhões de usuários interagem através de elos sociais. Apesar de tanta popularidade, o estudo de redes sociais ainda está em sua infância, já que estes ambientes estão ainda experimentando novas tendências e enfrentando diversos novos problemas e desafios.

Redes sociais compõem ambientes perfeitos para o estudo de vários temas da computação, incluindo sistemas multimídia e interação humano-computador. Além disso, por permitir que usuários criem conteúdo, redes sociais vêm se tornando um tema chave

em pesquisas relacionadas à organização e tratamento de grandes quantidades de dados, além de constituírem um ambiente ideal para extração de conhecimento e aplicação de técnicas de mineração de dados.

Este trabalho oferece uma introdução ao pesquisador que pretende explorar o tema. Inicialmente, foram apresentadas as principais características das redes sociais mais populares atualmente. Em seguida, discutimos as principais métricas e tipos de análises utilizadas no estudo dos grafos que formam a topologia das redes sociais. Finalmente, sumarizamos as principais abordagens utilizadas para se obter dados de redes sociais online e discutimos trabalhos recentes que utilizaram essas técnicas.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Instituto Nacional de Ciência e Tecnologia para a Web.

Referências

[1] comscore: Americans viewed 12 billion videos online in may 2008. http://www.comscore.com/press/release.asp?press=2324. Acessado em Março/2010.

[2] Developer analytics. http://www.developeranalytics.com. Acessado em Março/2010.

[3] Facebook application directory. http://www.facebook.com/apps. Acessado em Março/2010.

[4] Facebook platform. http://developers.facebook.com. Acessado em Março/2010.

[5] Facebook Press Room, Statistics. http://www.facebook.com/press/info.php?statistics. Acessado em Março/2010.

[6] Gnuplot. http://www.gnuplot.info/. Acessado em Agosto/2010.

[7] Google OpenSocial. http://code.google.com/apis/opensocial/. Acessado em Março/2010.

[8] List of social network web sites. http://en.wikipedia.org/wiki/List_of_social_networking_websites. Acessado em Março/2010.

[9] Matlab. http://www.mathworks.com/products/matlab/. Acessado em Agosto/2010.

[10] Needle in a Haystack: Efficient Storage of Billions of Photos. Facebook Engineering Notes, http://tinyurl.com/cju2og. Acessado em Março/2010.

[11] New york times. a web site born in u.s. finds fans in brazil. http://www.nytimes.com/2006/04/10/technology/10orkut.html. Acessado em Março/2010.

[12] New york times. uploading the avantgarde. http://www.nytimes.com/2009/09/06/magazine/06FOB-medium-t.htm. Acessado em Julho/2010.

[13] Yahoo! pipes. http://pipes.yahoo.com/pipes. Acessado em Agosto/2010.

[14] YouTube fact sheet. http://www.youtube.com/t/fact_sheet. Acessado em Março/2010.

[15] L. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the web. *First Monday*, 8(6), 2003.

[16] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *World Wide Web Conference (WWW)*, pages 835–844, 2007.

[17] R. Albert, H., Jeong, and A. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

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

[19] N. Ali-Hasan and L. Adamic. Expressing social relationships on the blog through links and comments. In *AAAI Conference on Weblogs and Social Media (ICWSM)*, 2007.

[20] A. Amaral, A. Scala, M. Barthelemy, and E. Stanley. Classes of small-world networks. 97(21):11149–11152, 2000.

[21] B. Williamson. Social network marketing: ad spending and usage. *EMarketer Report*, 2007. http://tinyurl.com/2449xx. Acessado em Março/2010.

[22] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439), 1999.

[23] F. Benevenuto, F. Duarte, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Understanding video interactions in YouTube. In *ACM Conference on Multimedia (MM)*, pages 761–764, 2008.

[24] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[25] F. Benevenuto, A. Pereira, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Avaliação do perfil de acesso e navegação de usuários em ambientes web de compartilhamento de vídeos. In *Brazilian Symposium on Multimedia Systems and Web (WebMedia)*, pages 149–156, 2009.

[26] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 620–627, 2009.

[27] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, M. Gonçalves, and K. Ross. Video pollution on the web. *First Monday*, 15(4), April 2010.

[28] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)*, 5(4):1–25, 2009.

[29] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying video spammers in online social networks. In *Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pages 45–52, 2008.

[30] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 49–62, 2009.

[31] J. Binder, A. Howes, and A. Sutcliffe. The problem of conflicting social spheres: effects of network structure on experienced tension in social network sites. In *ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 965–974, 2009.

[32] D. Boyd. *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*. Cambridge, MA, 2007.

[33] D. Boyd and N. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1-2), 2007.

[34] V. Braitenberg and A. Schüz. *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer-Verlag, 1998.

[35] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[36] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33:309–320, 2000.

[37] M. Burke, C. Marlow, and T. Lento. Feed me: Motivating newcomer contribution in social network sites. In *ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 945–954, 2009.

[38] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *In 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[39] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 1–14, 2007.

[40] M. Cha, A. Mislove, and K. Gummadi. A measurement-driven analysis of information propagation in the Flickr social network. In *World Wide Web Conference (WWW)*, pages 721–730, 2009.

[41] C. Chapman and M. Lahav. International ethnographic observation of social networking sites. In *ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 3123–3128, 2008.

[42] P. Chatterjee, D. L. Hoffman, and T. P. Novak. Modeling the clickstream: implications for web-based advertising efforts. *Marketing Science*, 22(4):520–541, 2003.

[43] D. Chau, Pandit, S. Wang, and C. Faloutsos. Parallel crawling for online social networks. In *World Wide Web Conference (WWW)*, pages 1283–1284, 2007.

[44] H. Chun, H. Kwak, Y. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of Cyworld. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 57–70, 2008.

[45] G. Cormode and B. Krishnamurthy. Key differences between web 1.0 and web 2.0. *First Monday*, 13(6), 2008.

[46] X. Dale and C. Liu. Statistics and social network of YouTube videos. In *Int'l Workshop on Quality of Service (IWQoS)*, 2008.

[47] F. Duarte, F. Benevenuto, V. Almeida, and J. Almeida. Locality of reference in an hierarchy of web caches. In *IFIP Networking Conference (Networking)*, pages 344–354, 2006.

[48] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic characteristics and communication patterns in blogosphere. In *Conference on Weblogs and Social Media (ICWSM)*, 2007.

[49] H. Ebel, L. Mielsch, and S. Bornholdt. Scale free topology of e-mail networks. *Physical Review E*, 66(3):35103, 2002.

[50] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, pages 251–262, 1999.

[51] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *World Wide Web Conference (WWW)*, pages 482–490, 2004.

[52] D. Garlaschelli and M. Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26):268701, 2004.

[53] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: A view from the edge. In *ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 15–28, 2007.

[54] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Characterizing user sessions on YouTube. In *IEEE Multimedia Computing and Networking (MMCN)*, 2008.

[55] K. Gummadi, R. Dunn, S. Saroiu, S. Gribble, H. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2003.

[56] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Int'l. Conference on Very Large Data Bases (VLDB)*, pages 576–587, 2004.

[57] A. Joinson. Looking at, looking up or keeping up with people?: motives and use of Facebook. In *ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 1027–1036, 2008.

[58] R. King. When your social sites need networking, *BusinessWeek*, 2007. `http://tinyurl.com/o4myvu`. Acessado em Março/2010.

[59] B. Krishnamurthy. A measure of online social networks. In *Conference on Communication Systems and Networks (COMSNETS)*, 2009.

[60] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

[61] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Int'l World Wide Web Conference (WWW)*, 2010.

[62] S. Lee, P. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(30):102–109, 2006.

[63] K. Lerman. Social information processing in news aggregation. *IEEE Internet Computing*, 11(6):16–28, 2007.

[64] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):228–237, 2007.

[65] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *World Wide Web Conference (WWW)*, 2008.

[66] L. Li, D. Alderson, J. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4), 2005.

[67] A. Mahanti, D. Eager, and C. Williamson. Temporal locality and its impact on web proxy cache performance. *Performance Evaluation Journal*, 42(2-3):187–203, 2000.

[68] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, May 1967.

[69] A. Mislove. *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, Department of Computer Science, 2009.

[70] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 29–42, 2007.

[71] C. Moore and M. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678, 2000.

[72] A. Nazir, S. Raza, and C. Chuah. Unveiling facebook: A measurement study of social network based applications. In *ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 43–56, 2008.

[73] A. Nazir, S. Raza, D. Gupta, C. Chua, and B. Krishnamurthy. Network level footprints of facebook applications. In *ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 63–75, 2009.

[74] M. Newman. The structure of scientific collaboration networks. 98(2):404–409, 2001.

[75] M. Newman. Assortative mixing in networks. *Physical Review E*, 89(20):208701, 2002.

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

[77] M. Newman. Coauthorship networks and patterns of scientific collaboration. 101(1):5200–5205, 2004.

[78] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):26113, 2004.

[79] J. Otterbacher. 'helpfulness' in online communities: a measure of message quality. In *ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 955–964, 2009.

[80] N. O. Report. Social networks & blogs now 4th most popular online activity, 2009. `http://tinyurl.com/cfzjlt`. Acessado em Março/2010.

[81] P. Rodriguez. Web infrastructure for the 21st century. *WWW'09 Keynote, 2009*. `http://tinyurl.com/mmmaa7`. Acessado em Março/2010.

[82] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 35–48, 2009.

[83] S. Schroeder. 20 ways to aggregate your social networking profiles, *Mashable*, 2007. `http://tinyurl.com/2ceus4`. Acessado em Março/2010.

[84] J. Thom-Santelli, M. Muller, and D. Millen. Social tagging roles: publishers, evangelists, leaders. In *ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 1041–1044, 2008.

[85] M. Torkjazi, R. Rejaie, and W. Willinger. Hot today, gone tomorrow: On the migration of myspace users. In *ACM SIGCOMM Workshop on Online social networks (WOSN)*, pages 43–48, 2009.

[86] K. S. Trivedi. *Probability and statistics with reliability, queuing and computer science applications*. John Wiley and Sons Ltd., Chichester, UK, 2002.

[87] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in Facebook. In *ACM SIGCOMM Workshop on Online Social Networks (WOSN)*, pages 37–42, 2009.

[88] S. Wasserman, K. Faust, and D. Iacobucci. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, 1994.

[89] D. Watts. *Small Worlds: the Dynamics of Networks Between Order and Randomness*. Princeton University Press, 1999.

[90] D. Watts. A simple model of global cascades on random networks. 99(9):5766–5771, 2002.

[91] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *ACM international conference on Web search and data mining (WSDM)*, pages 261–270, 2010.

[92] C. Williamson. On filter effects in web caching hierarchies. *ACM Transactions on Internet Technology (TOIT)*, 2(1):47–77, 2002.

[93] J. Zhang, M. Ackerman, and L. Adamic. Expertise networks in online communities: Structure and algorithms. In *World Wide Web Conference (WWW)*, pages 221–230, 2007.

[94] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: YouTube network traces at a campus network - measurements and implications. In *IEEE Multimedia Computing and Networking (MMCN)*, 2008.