Coleta e Análise de Grandes Bases de Dados de Redes Sociais Online

Fabrício Benevenuto[†] Jussara M. Almeida[‡] Altigran S. da Silva[‡⋆]

[†]Departamento de Ciência da Computação
Universidade Federal de Ouro Preto
Ouro Preto, MG, Brasil
fabricio@iceb.ufop.br

[‡]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brasil
jussara@dcc.ufmg.br

[⋆]Departamento de Ciência da Computação
Universidade Federal do Amazonas
Manaus, AM, Brasil
alti@dcc.ufam.edu.br

Resumo

Redes sociais online têm se tornado extremamente populares, levando ao surgimento e à crescente popularização de uma nova onda de aplicações na Web. Associado a esse crescimento, redes sociais estão se tornando um tema central de pesquisas em diversas áreas da Ciência da Computação. Este mini-curso 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 análises utilizadas no estudo dos grafos que formam a topologia das redes sociais. Finalmente, sumarizamos as principais abordagens para coleta e tratamento de dados de redes sociais online, e discutimos trabalhos recentes que ilustram o uso de tais técnicas.

Abstract

Online social networks have become extremely popular, causing the deployment and increasing popularity of a new wave of applications on the Web. Associated with this popularity growth, online social networks are becoming a key theme in several research areas. This short course offers an introduction to the researcher who aims at exploring this theme. Initially, we present the main characteristics of

current online social network applications. Then, we discuss the main metrics and analyses employed to study the graphs that represent the social network topologies. Finally, we summarize the main approaches used to collect and process online social network datasets, and discuss recent work that illustrates the use of such approaches.

1.1. Introdução

Desde seu início, a Internet tem sido palco de uma série de novas aplicações, incluindo email, aplicações par-a-par, aplicações de comércio eletrônico, assim como vários serviços Web. Atualmente, a Web vem experimentando uma nova onda de aplicações associada à proliferação das redes sociais online e ao crescimento da popularidade da 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* [Nielsen Online 2010], *mídia social*, termo usado em referência a conteúdo criado e disseminado via interações sociais, 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, este seria o terceiro país mais populoso do mundo, graças aos seus 500 milhões de usuários registrados [Facebook 2010c]. 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. Por exemplo, o Facebook compartilha mais de 60 bilhões de fotos, que ocupam mais de 1.5 PB de espaço [Facebook 2010d]. A quantidade de conteúdo que o YouTube armazena em 60 dias seria equivalente ao conteúdo televisionado em 60 anos, sem interrupção, pelas emissoras norte-americanas NBC, CBS e ABC juntas [New York Times 2010a]. De fato, o YouTube foi acessado por mais de 100 milhões de usuários apenas em Janeiro de 2009 [comScore 2010], com uma taxa de *upload* de 10 horas de vídeo por minuto [YouTube 2010].

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 têm 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 [eMarketer 2007]. Além disso, usuários de redes sociais online compartilham e recebem uma grande quantidade de informação, influenciando e sendo influenciados por seus amigos [Cha et al. 2010]. Conseqüentemente, redes sociais online estão se tornando cada vez mais um alvo de campanhas políticas [Gabrilovich et al. 2004] e de diversas outras formas de marketing viral, onde usuários são encorajados a compartilhar anúncios sobre marcas e produtos com seus amigos [Leskovec et al. 2007].

- **Sociológica**: No passado o estudo de redes sociais era um domínio de sociólogos e antropólogos, que utilizavam, como ferramentas típicas para se obter dados, entrevistas e pesquisas com usuários voluntários [Wasserman et al. 1994]. Como conseqüência, muitos desses estudos foram realizados com base em amostras de dados pequenas e possivelmente pouco representativas. Com a popularização de aplicações de redes sociais online, surgiu a oportunidade de estudos nesse tema com o uso de grandes bases de dados, coletadas de tais aplicações. Sistemas como Facebook, Twitter, Orkut, MySpace e YouTube possuem milhões de usuários registrados e bilhões de elos que os conectam. Redes sociais online permitem o registro em larga escala de diversos aspectos da natureza humana relacionados à comunicação, à interação entre as pessoas e ao comportamento humano, em geral: elas permitem que as pessoas interajam mais, mantenham contato com amigos e conhecidos, e se expressem e sejam ouvidas por uma audiência local ou até mesmo global. De fato, redes sociais online vêm funcionando como um novo meio de comunicação, modificando aspectos de nossas vidas.
- **Melhorias dos sistemas atuais**: Assim como qualquer sistema Web, redes sociais online são vulneráveis a novas tendências e estão sujeitas a experimentarem uma rápida transferência de seus usuários para outros sistema, sem aviso prévio. Por exemplo, inicialmente o MySpace experimentou um crescimento exponencial no número de usuários. Entretanto, este crescimento foi seguido por uma forte queda depois de abril de 2008 devido a um aumento no número de usuários do Facebook [Torkjazi et al. 2009]. Um outro exemplo é o Orkut, que inicialmente cresceu muito rapidamente em diversos lugares, mas que teve sua popularidade concretizada somente em alguns países. Dentre esses países, o Brasil é o com maior número de usuários registrados [New York Times 2010b]. 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 disponíveis em diferentes redes sociais online abre um novo leque de opções para pesquisas relacionadas à recuperação de conteúdo, onde estratégias de busca

e recomendação de usuários e conteúdo são cada vez mais importantes.

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 os padrões de tráfego na Internet. Estudar aspectos de sistemas relacionados a redes sociais pode ser de grande importância para a próxima geração da infra-estrutura da Internet e para o projeto de sistemas de distribuição de conteúdo mais eficientes, eficazes e robustos [Krishnamurthy 2009, Rodriguez 2009].

- **Segurança e conteúdo indesejável**: Redes sociais online 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, tais como a inclusão de vídeos contendo *spams* em listas de vídeos mais populares [Benevenuto et al. 2008b, Benevenuto et al. 2009a], o envio de *spam* no Twitter [Benevenuto et al. 2010a], a inclusão de metadados (particularmente *tags*) que não descrevem adequadamente o conteúdo associado (p.ex: *tag spamming*) [Benevenuto et al. 2010c], etc. Conteúdo não solicitado consome a atenção humana, talvez o recurso mais importante na era da informação. Em última instância, o ruído e o distúrbio causados por alguns usuários reduzem a efetividade da comunicação online.

Redes sociais online são ambientes perfeitos para o estudo de vários temas da computação, incluindo sistemas distribuídos, padrões de tráfego na Internet, mineração de dados, 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 análises utilizadas no estudo dos grafos que formam a topologia das redes sociais. Finalmente, sumarizamos as principais abordagens utilizadas para co-

letar e tratar dados de redes sociais online, e discutimos trabalhos recentes que utilizaram essas técnicas.

1.2. Definições e Características de Redes Sociais Online

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

1.2.1. 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 [Boyd e Ellison 2007, Mislove 2009]. Definimos uma rede social online como um serviço Web que permite a um indivíduo (1) construir perfis públicos ou semi-públicos dentro de um sistema, (2) articular uma lista de outros usuários com os quais ele(a) compartilha conexões e (3) visualizar e percorrer suas listas de conexões assim como outras listas criadas por outros usuários do sistema.

Com base nessa definição, existem várias redes sociais online disponíveis na Web, que variam de acordo com seus objetivos primários. A Tabela 1.1 apresenta uma lista com várias redes sociais online populares atualmente, juntamente com seus principais propósitos. Uma lista atualizada e exaustiva de redes sociais online, com mais de 150 sítios Web, pode ser encontrada em [Wikipedia 2010].

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
LastFM	Compartilhamento de rádio/músicas	http://www.last.fm

Tabela 1.1. Algumas Redes Sociais Online Populares

1.2.2. Elementos das Redes Sociais Online

Nesta seção, discutimos várias funcionalidades oferecidas pelas principais aplicações de redes sociais online 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 de um usuário na rede social também contém mensagens de outros membros e listas de pessoas identificadas como seus 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 [Boyd 2007] 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 [Burke et al. 2009] conduziram um estudo utilizando dados de 140,000 novos usuários do Facebook e concluíram que atividades como atualizações são vitais para que novos usuários contribuam para o sistema. Como atualizações podem receber comentários de outros usuários, elas também são formas especiais de comunicação em redes sociais online.

- **Comentários**: A maior parte das aplicações de redes sociais online permite que usuários comentem o conteúdo compartilhado por outros. Alguns sistemas também permitem que usuários adicionem comentários nos perfis de outros usuários. Comentários são um meio primordial de comunicação em redes sociais online, e também podem ser interpretados como expressão de relações sociais [Ali-Hasan e Adamic 2007, Chun et al. 2008]. Como exemplo, usuários do YouTube podem comentar os vídeos armazenados por outros no sistema, enquanto que tanto no Facebook quanto no Orkut, usuários podem adicionar comentários às fotos de outros. Da mesma forma, 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, por exemplo, usuários podem apenas gostar de uma postagem, clicando no botão "*I like this*". De fato, estatísticas recentes indicam que cada usuário do Facebook avalia, em média, 9 objetos por mês [Facebook 2010c]. Já 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 contendo alguma forma de *spam*.

  Avaliações de conteúdo são úteis de várias formas. Como exemplo, elas ajudam usuários de sistemas como o YouTube 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 identificar conteúdo em destaque, para suportar 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 [Lerman 2007].

- **Listas de Favoritos**: Várias aplicações sociais utilizam listas de favoritos para permitir que usuários selecionem e organizem seu 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 de fotos favoritas no Flickr. Nesses sistemas, usuários podem navegar na lista de favoritos de outros usuários para buscar novos conteúdos [Cha et al. 2009]. Conseqüentemente, listas de favoritos também facilitam a descoberta de conteúdo e a propagação de informação. Sistemas como o Orkut e o Twitter também provêem listas de favoritos (fãs).

- **Listas de Mais Populares (***Top Lists***)**: Tipicamente, redes sociais online que têm o compartilhamento de conteúdo como elemento central do sistema, como o YouTube que é centrado no compartilhamento de vídeos, 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, ou comentários) ou relativas aos usuários (ex: número de assinantes).

- **Metadados**: Em várias aplicações de redes sociais online, tais 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, uma vez

que grande parte dos serviços de informação (p.ex: busca, organização de conteúdo, recomendação, propaganda) ainda utilizam os metadados (particularmente as *tags*) como principal fonte de dados [Boll 2007].

1.3. Teoria de Redes Complexas

Redes sociais online são inerentemente *redes complexas*. Consequentemente, vários estudos recentes analisaram as características de diferentes redes sociais online utilizando como base teorias existentes da área de redes complexas [Adamic et al. 2003, Mislove et al. 2007, Dale e Liu 2008, Kumar et al. 2006, Leskovec e Horvitz 2008, Benevenuto et al. 2008a, Benevenuto et al. 2009b]. 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 [Moore e Newman 2000], propagação de informação [Watts 2002], busca na Web [Broder et al. 2000], e consequências de ataques a redes de computadores [Albert et al. 2000]. A seguir, várias propriedades estatísticas e métricas comumente utilizadas para analizar e classificar redes complexas são apresentadas na seção 1.3.1. As seções 1.3.2 e 1.3.3 discutem propriedades específicas comumente associadas a várias redes complexas, a saber, redes *small-world* e redes *power-law*, respectivamente. Uma revisão detalhada sobre métricas e teoria de redes complexas pode ser encontrada na referência [Newman 2003].

1.3.1. Métricas para o Estudo de Redes Complexas

Uma rede é um conjunto de elementos, que chamamos de vértices ou nodos, com conexões entre eles, chamadas de arestas. A estrutura topológica de uma rede pode ser então modelada por um grafo, que, por sua vez, pode ser caracterizado a partir de diversas métricas, discutidas a seguir. Assume-se que o leitor tenha um conhecimento sobre a terminologia utilizada em teoria de grafos.

1.3.1.1. Grau dos Vértices

Uma característica importante da estrutura de uma rede é a distribuição dos graus de seus vértices. Tal distribuição foi caracterizada em várias redes (ex: redes de emails [Gomes et al. 2007], a topologia da Internet [Faloutsos et al. 1999], a Web [Barabasi e Albert 1999], e redes neurais [Braitenberg e Schüz 1998]) como seguindo uma lei de potência. Em outras palavras, a probabilidade de um nodo ter grau k é proporcional a $k^{-\alpha}$. Consequentemente, uma métrica comumente utilizada para comparar diferentes redes é o expoente α, obtido através de uma regressão linear. Valores típicos para o expoente α ficam entre 1.0 e 3.5 [Ebel et al. 2002]. Para redes direcionadas, é comum analisar as distribuições dos graus dos nodos em ambas as direções, isto é, a distribuição do grau de entrada e a distribuição do grau de saída.

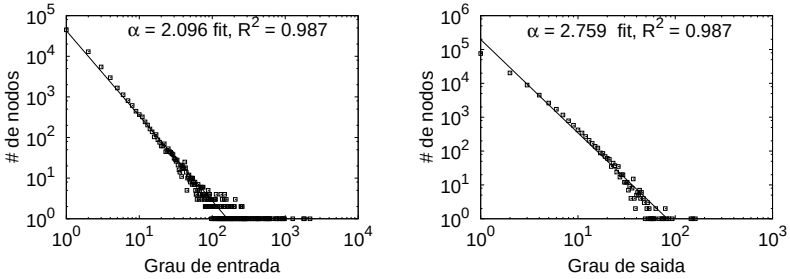


Figura 1.1. Distribuições dos Graus de Entrada e de Saída em um Grafo de Interações entre Usuários através de Vídeos do YouTube

Como exemplo, a Figura 1.1 mostra as distribuições dos graus de entrada (esquerda) e de saída (direita) para um grafo formado a partir das interações entre usuários de vídeos do YouTube [Benevenuto et al. 2008a, Benevenuto et al. 2009b]. Note que a curva da regressão linear, utilizada para se calcular o expoente α, também é exibida nesses gráficos. Ferramentas como o Gnuplot [Gnuplot 2010] e o Matlab [mathworks 2010] podem ser utilizados para realizar a regressão e calcular o valor de α. Para verificar a acurácia da regressão, é comum medir o coeficiente de determinação R^2 [Trivedi 2002]. Quanto mais próximo o valor de R^2 for de 1 (regressão perfeita), menor será a diferença entre o modelo de regressão e os dados reais.

1.3.1.2. Coeficiente de Agrupamento

O coeficiente de agrupamento (*clustering coefficient*) de um nodo i, $cc(i)$, é a razão entre o número de arestas existentes entre os vizinhos de i e o número máximo de arestas possíveis entre estes vizinhos. Como exemplo, a Figura 1.2 mostra o valor do coeficiente de agrupamento para o nodo escuro em três cenários diferentes[1]. 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.

Podemos notar que o coeficiente de agrupamento representa uma medida da densidade de arestas estabelecidas entre os vizinhos de um nodo. O coeficiente de agrupamento de uma rede, CC, é calculado como a média dos coeficientes de agrupamento de todos os seus nodos.

[1] Arestas tracejadas não existem e apenas ilustram as possíveis conexões entre os vizinhos do nodo escuro.

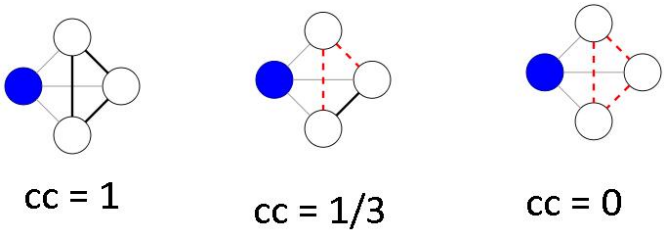


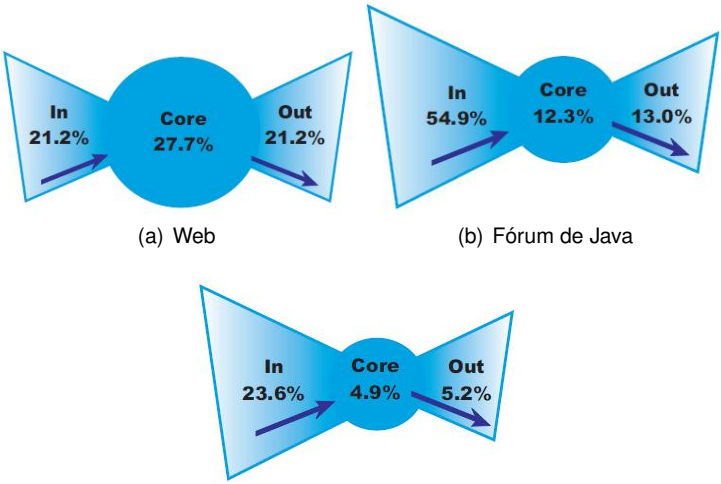
Figura 1.2. Cálculo do Coeficiente de Agrupamento de um Nodo em Três Cenários Diferentes

1.3.1.3. Componentes

Um componente em um grafo é um conjunto de nodos, onde cada nodo possui um caminho para todos os outros nodos do conjunto. Para grafos direcionados, um componente é chamado de fortemente conectado (SCC - *Strongly Connected Component*) quando existe um caminho direcionado entre cada par de nodos do conjunto. 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 representada por um grafo em que nodos são páginas Web e arestas são elos existentes entre as páginas [Broder et al. 2000]. 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* [Broder et al. 2000], pois a figura que ilustra 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 [Zhang et al. 2007, Benevenuto et al. 2009b]. A Figura 1.3, por exemplo, ilustra o uso do modelo *bow tie* para comparar a estrutura dos componentes de três redes diferentes, a saber, a rede Web, uma rede formada pelas conexões estabelecidas em um Fórum de Java e a rede estabelecida entre usuários do YouTube. 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*. Note que há diferenças estruturais significativas entre as 3 redes: a rede Web é muito mais balanceada, em termos do tamanho dos componentes, enquanto que, para as outras duas, o componente *in* contém um



(a) Web (b) Fórum de Java

(c) Interações com vídeo

Figura 1.3. Estrutura dos Componentes da Web [Broder et al. 2000], do Fórum de Java [Zhang et al. 2007] e do Grafo de Interações de Usuários do YouTube [Benevenuto et al. 2009b]

percentual muito maior dos nodos.

1.3.1.4. Distância Média e Diâmetro

A distância média de um grafo é o número médio de arestas em todos os caminhos mínimos existentes entre todos os pares de 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 e, em geral, é também computado somente para nodos do WCC ou do SCC.

1.3.1.5. Assortatividade

De acordo com Newman [Newman 2002], 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. Para caracterizar a assortatividade de uma rede, medimos o grau médio de todos os vizinhos dos nodos com grau k, dado por $knn(k)$. A assortatividade ou disassortatividade de uma rede é geralmente estimada avaliando os valores de $knn(k)$ em função de k. Valores crescentes indicam assortatividade, isto é, nodos com graus maiores tendem a se conectar a nodos com um número maior de conexões. Valores decrescentes de $knn(k)$ em função de k, por sua vez, indicam uma rede disassortativa.

1.3.1.6. *Betweenness*

Betweeness é uma medida relacionada à centralidade dos nodos ou de arestas na rede. O *betweeness* $B(e)$ de uma aresta e é definido como o número de caminhos mínimos entre todos os pares de nodos em um grafo que passam por e [Newman e Girvan 2004]. Se existem múltiplos caminhos mínimos entre um par de nodos, 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* fazem parte de um número maior de 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 pelo nodo dado. Nodos que possuem muitos caminhos mínimos que passam por eles possuem maior *betweenness*, indicando uma maior importância para a estrutura da rede.

1.3.1.7. Reciprocidade

Uma forma interessante de se observar a reciprocidade de um nodo i em um grafo direcionado é medindo a porcentagem dos nodos apontados por i que apontam para ele. Em outras palavras, a reciprocidade ($R(i)$) é dada por:

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

onde $O(i)$ é o conjunto de nodos apontados por i e $I(i)$ é o conjunto de nodos que apontam para i.

Outra métrica interessante de ser observada é o coeficiente de reciprocidade, ρ, que captura a reciprocidade das interações em toda a

rede [Garlaschelli e Loffredo 2004]. O coeficiente de reciprocidade ρ é definido pelo coeficiente de correlação entre entidades da matriz de adjacência representativa do grafo direcionado. Em outras palavras, seja a matriz a definida como a_{ij} = 1 se há uma aresta de i para j no grafo, e a_{ij} = 0, caso contrário. Definimos a reciprocidade da rede ρ como:

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

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, ela é anti-recíproca.

1.3.1.8. PageRank

O PageRank é um algoritmo interativo que assinala um peso numérico para cada nodo com o propósito de estimar sua importância relativa no grafo. O algoritmo foi inicialmente proposto por Brin and Page [Brin e Page 1998] para ordenar resultados de busca do protótipo de máquina de busca da 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 i, $PR(i)$, é definida da seguinte forma:

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

onde $S(i)$ é o conjunto de páginas que apontam para i, 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 [Zhang et al. 2007] e usuários influentes no Twitter [Weng et al. 2010, Kwak et al. 2010]. Além disso, existem outras modificações do PageRank com propósitos específicos, como, por exemplo, a detecção de spam na Web [Gyöngyi et al. 2004].

1.3.2. Redes Small-World

O conceito de redes *small-world* ficou bastante conhecido com o famoso experimento de Milgram [Milgram 1967]. Este 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 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 requeridos 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 agrupamento alto e diâmetro pequeno [Watts 1999]. Estas propriedades foram verificadas em várias redes como a Web [Albert et al. 1999, Broder et al. 2000], redes de colaboração científica [Newman 2001, Newman 2004] em que pesquisadores são nodos e arestas ligam co-autores de artigos, redes de atores de filmes [Amaral et al. 2000] em que atores são nodos e arestas ligam atores que participaram do mesmo filme, e redes sociais online [Adamic et al. 2003, Mislove et al. 2007, Dale e Liu 2008, Leskovec e Horvitz 2008, Benevenuto et al. 2008a, Benevenuto et al. 2009b]. Em particular, Mislove e colaboradores [Mislove et al. 2007] verificaram propriedades *small-world* em quatro redes sociais online: LiveJournal, Flickr, Orkut, e YouTube.

1.3.3. Redes Power-Law e Livres de Escala

Redes cujas distribuições dos graus dos nodos seguem uma lei de potência (Seção 1.3.1.1) são ditas redes *power-law*. Redes livres de escala (*scale free*) são uma classe de redes que seguem leis de potência caracterizadas pela seguinte propriedade: nodos de grau alto tendem a se conectar a outros nodos de grau alto. Barabási e colaboradores [Barabasi e Albert 1999] propuseram um modelo para gerar redes livres 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 [Li et al. 2005] criaram uma métrica para medir se uma rede é livre de escala ou não, além de prover uma longa discussão sobre o tema.

1.4. Técnicas de Coleta de Dados

Em um passado recente, redes sociais eram um domínio de sociólogos e antropólogos, que utilizavam pesquisas e entrevistas com pequenos grupos de usuários como ferramentas de coleta de dados [Wasserman et al. 1994]. Com o surgimento das redes sociais online, a obtenção de dados reais em larga escala se tornou possível, e pesquisadores de 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 Fi-

gura 1.4 apresenta possíveis pontos de coleta de dados, que variam desde entrevistas com os usuários até à instalação de coletores localizados em servidores proxy ou em aplicações. A seguir discutimos essas diferentes abordagens, bem como trabalhos que ilustram o uso dessas estratégias.

1.4.1. Dados dos Usuários

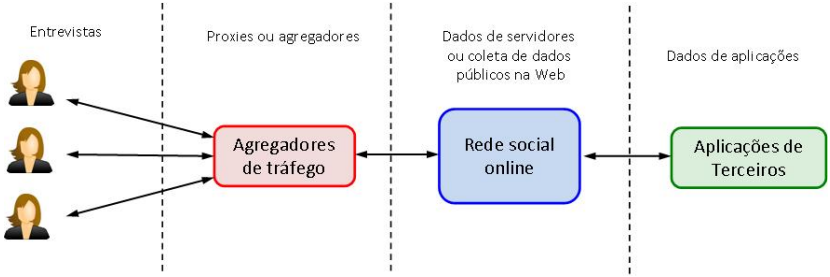


Figura 1.4. Possíveis Pontos de Coleta de Dados

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 [Thom-Santelli et al. 2008, Joinson 2008, Chapman e Lahav 2008, Binder et al. 2009, Otterbacher 2009], para a qual entrevistas estruturadas são as formas mais populares de obtenção de dados.

Como exemplo, através de entrevistas com usuários do Facebook, Joinson e seus colaboradores [Joinson 2008] identificaram várias razões pelas quais usuários utilizam o sistema, tais como conexão social, compartilhamento de interesses, compartilhamento e recuperação de conteúdo, navegação na rede social e atualização do seu estado atual. Chapman e Lahav [Chapman e Lahav 2008] conduziram entrevistas e analisaram os padrões de navegação de 36 usuários de quatro nacionalidades diferentes para examinar diferenças etnográficas no uso de redes sociais online.

1.4.2. 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 redes 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.

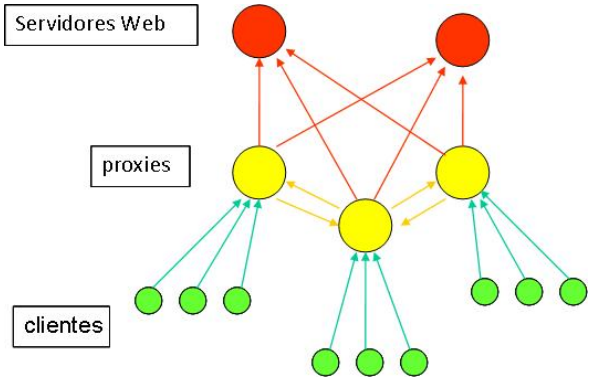


Figura 1.5. Exemplo de um Servidor Proxy Intermediando o Tráfego entre Clientes e Servidores

1.4.2.1. Servidores Proxy

Coletar dados de um servidor proxy tem sido uma estratégia utilizada em vários estudos sobre o tráfego da Internet [Duarte et al. 2006, Gummadi et al. 2003, Mahanti et al. 2000, Williamson 2002]. A Figura 1.5 ilustra como um servidor proxy funciona como um agregador de tráfego de seus clientes, podendo ser utilizado para delimitar uma porção da rede, composta por computadores em uma mesma localização geográfica.

Dado o crescente interesse por vídeos na Web, alguns trabalhos recentes utilizaram servidores proxy para obter dados do tráfego gerado por sistemas de compartilhamento de vídeos, como o YouTube. Gill e colaboradores [Gill et al. 2007] 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 fazer o *download* de conteúdo do YouTube, correspondem a 99.87% do total das requisições que passam pelo servidor proxy. Eles ainda caracterizaram diversas métricas relacionadas a estas requisições, tais como a duração, a idade e a categoria dos vídeos. Mais recentemente, os mesmos autores caracterizam sessões de usuários no YouTube [Gill et al. 2008]. Eles mostraram que uma sessão típica dura cerca de 40 minutos, caracterizando também o tempo entre sessões e os tipos de conteúdo transferidos em cada sessão. Finalmente, Zink e colaboradores [Zink et al. 2008] também estudaram o tráfego de vídeos do YouTube coletado no servidor proxy de uma universidade. Eles também analisaram, via simulação, os benefícios da adoção de abordagens de caches locais e globais para vídeos, bem como do uso de arquiteturas P2P para distribuição de vídeos. De maneira geral, eles mostraram que essas abordagens poderiam reduzir significativamente o volume de tráfego, permitindo acesso aos vídeos

de forma mais rápida.

Em um trabalho recente, Schneider e colaboradores [Schneider et al. 2009] 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 criaram o que foi chamado de *clickstream* [Chatterjee et al. 2003] para redes sociais online, capturando cada passo da navegação dos usuários do ISP. Eles discutiram amplamente a metodologia de reconstrução dos acessos dos usuários e, com base nesses dados, analisaram as seqüências de requisições realizadas pelos usuários de vários sistemas, incluindo o Facebook.

1.4.2.2. Agregadores 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 seus perfis de uma forma mais simples e unificada [King 2007, Schroeder 2007]. Ao se conectarem a um agregador de redes sociais online, usuários podem acessar suas contas através de uma interface única, sem precisar se conectar 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, enquanto 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 [OpenSocial 2010], e todo o conteúdo é exibido através da interface do sistema agregador. A Figura 1.6 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, tais como checar atualizações de amigos, enviar mensagens e compartilhar fotos.

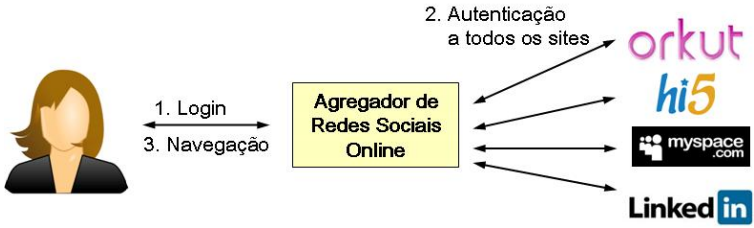


Figura 1.6. Ilustração de um Usuário se Conectando a Múltiplas Redes Sociais Online Simultaneamente Através de um Portal Agregador

Recentemente, a partir de uma colaboração com um agregador de

redes sociais online, coletamos dados da navegação dos usuários (i.e., *clickstream*) em 4 redes sociais online: Orkut, Hi5, MySpace e LinkedIn [Benevenuto et al. 2009c]. A Tabela 1.2 mostra o número de usuários, sessões e requisições HTTP para cada uma dessas redes. Baseados nesses dados e em dados coletados do Orkut, nós examinamos o comportamento dos usuários nas redes sociais online, caracterizando as interações estabelecidas entre eles através das várias atividades realizadas.

	# 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 1.2. Sumário dos Dados Obtidos de um Agregador de Redes Sociais Online

1.4.3. Dados de Servidores de Redes Sociais Online

Idealmente, servidores de aplicações de redes sociais online são os locais mais adequados para a coleta de dados, uma vez que eles têm uma visão completa de todas as ações e atividades realizadas por todos os usuários do sistema em um dado período de tempo. Entretanto, o acesso a dados armazenados por estes servidores, ainda que anonimizados, é tipicamente muito difícil. Dentre os poucos trabalhos que utilizaram dados obtidos diretamente de servidores de redes sociais online, podemos citar o estudo de Chun e seus colaboradores [Chun et al. 2008] sobre as interações textuais entre os usuários do Cyworld, uma rede social bastante popular na Coréia do Sul. Eles compararam as propriedades estruturais da rede de amizades explícita criada naquela aplicação com as propriedades da rede implícita estabelecida a partir de trocas de mensagens no livro de visitas do Cyworld, discutindo diversas similaridades e diferenças. Citamos também o trabalho de Baluja e colaboradores [Baluja et al. 2008], que utilizaram dados do histórico de navegação de 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 vídeos personalizadas para os usuários do YouTube. Finalmente, Duarte e seus colaboradores [Duarte et al. 2007] caracterizaram o tráfego em um servidor de blogs do UOL (www.uol.com.br), enquanto nós estudamos a navegação dos usuários em um servidor de vídeos do UOL, chamado UOL Mais [Benevenuto et al. 2010b].

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 [Mislove et al. 2007, Ahn et al. 2007], padrões de acesso no YouTube [Cha et al. 2007] e interações reconstruídas através de mensagens trocadas pelos usuários [Viswanath et al. 2009b]. A seguir discutimos vários aspectos relacionados à coleta de dados de redes sociais online.

1.4.3.1. Coleta por Amostragem

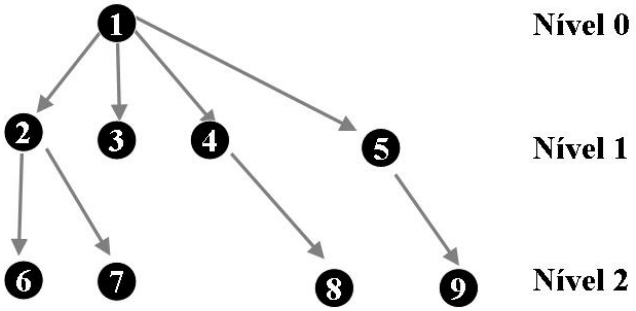


Figura 1.7. Exemplo de Busca em Largura em um Grafo

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, chamada *snowball* (ou bola de neve), consiste em coletar o grafo de uma rede social online seguindo uma busca em largura, como ilustra a Figura 1.7. A coleta inicia-se a partir de um 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.

Um ponto crucial sobre 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. Esta interrupção pode ter que ser forçada para tornar a coleta viável. Em particular, a busca completa em componentes muito grandes pode gastar um tempo excessivamente longo e proibitivo. Entretanto, é importante ressaltar que o resultado de uma coleta parcial pode gerar medidas tendenciosas, tais como distribuições de graus dos nodos com uma tendência maior para valores grandes, o que pode, em última instância, afetar as análises e conclusões observadas. Em outras palavras, os dados obtidos via coleta por *snowball* com interrupção precisam ser tratados e analisados com cautela. Em outras palavras, o tipo de análise a ser feita precisa ser considerado. Por exemplo, se realizarmos 3 passos da coleta, podemos calcular, com precisão, o coeficiente de agrupamento dos nodos semente. Entretanto, se quisermos computar o coeficiente de agrupamento médio de toda a rede ou outras métricas globais, como distribuição de graus, distância média, etc., a coleta por *snowball* pode resultar em números tendenciosos caso ela não inclua pelo menos um componente completo representativo da rede completa [Lee et al. 2006, Ahn et al. 2007].

Uma abordagem bastante difundida consiste em coletar o maior componente fracamente conectado (WCC) do grafo. Mislove e colaboradores [Mislove et al. 2007] argumentam que o maior WCC de um grafo é estruturalmente 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 a fazer parte de um grande grupo de pequenos componentes isolados ou até mesmo totalmente desconectados. 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, realizamos uma busca por palavras aleatórias no YouTube para verificar se o componente coletado era o maior componente [Benevenuto et al. 2009b, Benevenuto et al. 2008a]. 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.

Note que, em algumas redes sociais online como o Twitter ou o Flickr, o conceito de amizade é direcionado. Ou seja, um usuário u pode seguir um outro usuário v, sem necessariamente ser seguido por ele. Em casos como estes, ou seja, em grafos direcionados, é necessário percorrer as arestas do grafo em ambas as direções a fim de coletar o WCC completamente. Caso contrário, se coletarmos o grafo seguindo as arestas em apenas uma direção, não é garantido que consigamos coletar todos os nodos do WCC. A Figura 1.8 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

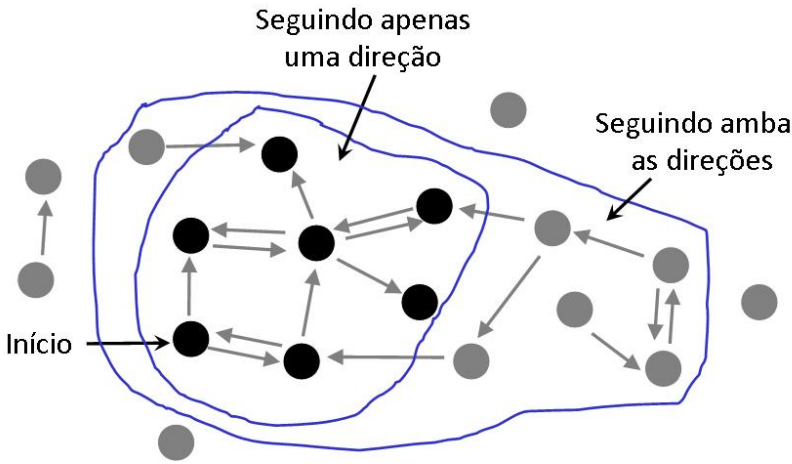


Figura 1.8. Exemplo de Coleta do WCC em um Grafo Direcionado

estudos que envolvem a coleta da Web [Broder et al. 2000]. Tipicamente, a Web é freqüentemente coletada seguindo apenas uma direção dos elos entre páginas, já que não é possível determinar o conjunto de páginas que apontam para uma página.

Uma outra estratégia de coleta por amostragem em redes é baseada em caminhadas aleatórias (*random walks*) [Ribeiro e Towsley 2010]. A idéia básica consiste em, a partir de um nodo semente v, prosseguir selecionando aleatoriamente um dos vizinhos de v, e assim sucessivamente até que um número pre-definido B de nodos tenham sido selecionados. Um mesmo nodo pode ser selecionado múltiplas vezes. Esta é a estratégia de caminhada aleatória mais comum disponível na literatura, embora existam outras abordagens que diferem em termos de como os vizinhos são selecionados [Lovász 1993, Robert e Casella 2005] Assim como a coleta por *snowball*, a coleta através de caminhadas aleatórias também tem suas limitações: a precisão das estimativas depende não somente da estrutura do grafo mas também da característica sendo estimada. Em particular, dependendo da estrutura do grafo, a coleta pode ficar “presa"dentro de um subgrafo. Neste caso, as estimativas podem ser imprecisas se as características do subgrafo não forem representativas da rede como um todo. Para mitigar este problema, Ribeiro e Towsley [Ribeiro e Towsley 2010] propuseram uma estratégia de coleta por caminhos aleatórias chamada *Frontier sampling*, que começa com m nodos sementes. Eles ainda propuseram estimadores sem tendência para várias métricas de redes complexas, incluindo o coeficiente de agrupamento global da rede, usando as arestas coletadas pelo método proposto.

1.4.3.2. 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.

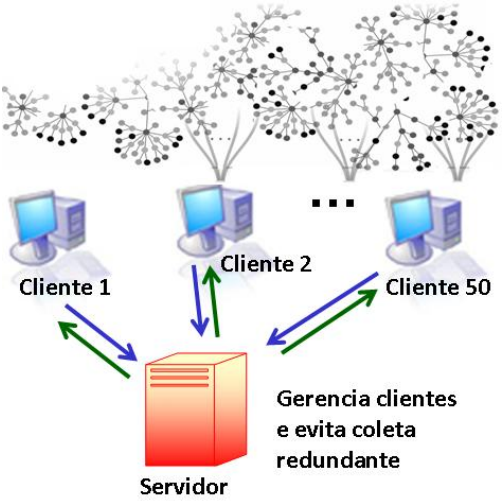


Figura 1.9. Exemplo de Coleta Realizada de Forma Distribuída

Uma forma de se realizar tal coleta, conforme descrito em [Chau et al. 2007], está representada na Figura 1.9. 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, armazenam e processam os dados coletados para identificar novos usuários. Novos usuários identificados pelas máquinas escravas são repassados para a máquina mestre, que, por sua vez, distribui novos usuários a serem coletados para as máquinas escravas.

1.4.3.3. Coleta por Inspeção de Identificadores

Como discutido anteriormente, idealmente, a coleta de uma rede social online deveria incluir a rede completa e não somente 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, realizamos uma coleta do Twitter seguindo essa estratégia. Foi solicitado aos administradores do Twitter a permissão para realizar uma coleta em larga escala. Em resposta, eles adicionaram os endereços IPs de 58 máquinas sob nosso controle em uma lista branca, com permissão para coletar dados. Cada uma das 58 máquinas, localizadas no *Max Planck Institute for Software Systems (MPI-SWS)*, na Alemanha[2], teve permissão para coletar dados a uma taxa máxima de 20 mil requisições por hora. Utilizando a API do Twitter, nosso coletor investigou todos os 80 milhões de IDs de forma seqüencial, coletando todas as informações públicas sobre esses usuários, bem como seus elos de seguidores e seguidos e todos os seus tweets. Dos 80 milhões de contas inspecionadas, 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, não encontramos nenhum identificador acima dos 80 milhões inspecionados, sugerindo que coletamos todos os usuários do sistema na época. Esses dados foram utilizados recentemente em dois trabalhos, um sobre detecção de spammers no Twitter [Benevenuto et al. 2010a] e o outro sobre medição de influência no Twitter [Cha et al. 2010].

Torkjazi e colaboradores [Torkjazi et al. 2009] também exploraram o uso de IDs sequenciais para inspecionar o surgimento de novos usuários no MySpace.

1.4.3.4. Coleta em Tempo Real

Redes sociais online possibilitam que usuários comuns expressem suas opiniões sobre os mais diversos assuntos, e propaguem informações que consideram relevantes em tempo real. É interessante notar como a interatividade e a avaliação do fluxo de informações em tempo real passou a ser um fator importante para várias aplicações Web, que monitoram fenômenos dinâmicos ocorridos na Web. Como exemplo, Google e Bing já indexam tweets públicos como forma de prover busca por informação em tempo real.

Do ponto de vista científico, informações disponibilizadas em tempo real na Web têm sido utilizadas de diferentes formas. O Google Insights[3], por exemplo, permite que o usuário consulte informações geográficas e temporais relacionadas ao volume de uma determinada consulta, permitndo também que o usuário salve essas informações em formato CSV. De fato, informações extraídas em tempo real do Google Insights foram utilizadas para

[2] Esta coleta foi realizada durante uma visita de 5 meses ao MPI-SWS realizada pelo primeiro autor deste trabalho

[3] http://www.google.com/insights/search

demonstrar que o volume de buscas na Web permite monitorar eventos, tais como o nível de desemprego e a ocorrência de doenças em tempo real [Choi e Varian 2009, Ginsberg et al. 2009]. Em particular, Ginsberg e colaboradores [Ginsberg et al. 2009] propuseram um método para monitorar ocorrências de gripe baseado em buscas realizadas no Google. Eles mostraram que, em áreas com grande população de usuários de máquinas de buscas, o volume de buscas relacionadas à gripe é proporcional à ocorrência da doença.

Em um trabalho recente, Sakaki e colaboradores [Sakaki et al. 2010] mostraram o poder da informação disponibilizada em tempo real nas redes sociais online propondo um mecanismo para detecção de ocorrências de terremotos baseado em monitoramento do Twitter. A abordagem, que consiste em simplesmente identificar tweets relacionados a terremotos por região, foi capaz de enviar alertas sobre terremotos mais rapidamente do que agências meteorológicas. Mais recentemente, Tumasjan e colaboradores [Tumasjan et al. 2010] mostraram que opiniões identificadas em tweets relacionados à eleição federal alemã foi capaz de refletir o sentimento político registrado fora das redes sociais. Mais recentemente, uma análise geográfica e temporal sobre dengue no Twitter foi realizada por Gomide e seus colaboradores [Gomide et al. 2011].

Para o caso específico do Twitter, é bastante simples coletar informações em tempo real, uma vez que o sistema oferece uma API com diversas opções relacionadas à coleta de dados em tempo real. Como exemplo, para coletar em tempo real uma amostra aleatória de tweets públicos, basta executar o seguinte comando em algum terminal UNIX, fornecendo o usuário e senha de um usuário registrado no Twitter.

```
curl http://stream.twitter.com/1/statuses/sample.json -uLOGIN:SENHA
```

1.4.3.5. Utilizando APIs

No contexto de desenvolvimento Web, uma API é tipicamente um conjunto de tipos de requisições HTTP juntamente com suas respectivas definições de resposta. Em aplicações de redes sociais online é comum encontrarmos APIs que listam os amigos de um usuário, seus objetos, suas comunidades, etc.

APIs são perfeitas para a coleta de dados de redes sociais online, pois oferecem os dados em formatos estruturados como XML e JSON. Como exemplo, a Figura 1.10 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 de centenas de requisições, visto que o Twitter só mostra alguns seguidores por página. Além disso, cada página conteria uma quantidade muito grande de dados desnecessários, que deveriam ser tratados e excluídos.

Vários sistemas possuem APIs, incluindo Twitter, Flickr, YouTube, Google

Figura 1.10. Exemplo da API do Twitter: http://twitter.com/users/show/fbenevenuto.xml

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 [Yahoo! Pipes 2010], permite a combinação de diferentes APIs de vários sistemas para a criação automatizada de Mashups.

1.4.3.6. Ferramentas e Bibliotecas

Desenvolver um coletor 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 que, 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.

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 1.11 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 1.12 utiliza a biblioteca URLLIB para realizar a mesma tarefa. O resultado da execução dos coletores é a lista de seguidores de um usuário do Twitter em formato XML, como ilustra a Figura 1.13.

```perl
#!/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 1.11. Exemplo de Uso da Biblioteca LWP em Perl

```python
#!/usr/bin/python

import urllib

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

print content
```

Figura 1.12. Exemplo de Uso da Biblioteca URLLIB em Python

1.4.3.7. Ética dos Coletores

É importante ressaltar que o uso de ferramentas automáticas de coleta (coletores ou robôs), se feito sem o devido cuidado, pode causar problemas de sobrecarga nos servidores alvos da coleta, o que, em última instância, pode afetar a qualidade de serviço da aplicação alvo. Para evitar que isto aconteça, muitos servidores adotam um protocolo conhecido como Robots.txt, no qual sítios Web regulamentam o que pode e o que não pode ser coletado do 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. Ele se aplica a qualquer tipo de coleta, seja ela parte de uma pesquisa sobre redes sociais online, ou um dos componentes centrais de uma máquina de busca como o Google [Brin e Page 1998].

O Robots.txt nada mais é do que um arquivo que fica no diretório raiz dos sítios e contém regras para coleta. Estas regras podem ser direcionadas a robôs específicos ou podem ser de uso geral, tendo como alvo qualquer robô. Ao visitar um site, os robôs devem primeiramente buscar pelo arquivo Robots.txt a fim de verificar suas permissões. Exemplos desses arquivos são:

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

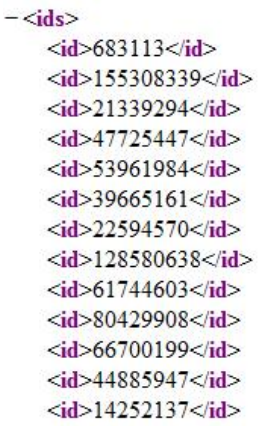


Figura 1.13. Resultado da Execução dos Coletores em Perl e Python

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 online, é importante verificar não só o arquivo Robots.txt dos sistemas, mas também os termos de uso do sistema.

1.4.4. 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: abriu sua plataforma para desenvolvedores de aplicações [Facebook 2010b]. Com esta inovação, desenvolvedores são capazes de criar diferentes tipos de aplicações. Com o sucesso no Facebook, outros sistemas, como Orkut e MySpace também adotaram essa estratégia. Como consequência, o número e a variedade de aplicações criadas nestes sistemas cresceram significativamente. O Facebook sozinho possui atualmente mais de 81,000 aplicaçõess[4] [DeveloperAnalytics 2010]. Empresas como a Zynga, especializadas em desenvolver essas aplicações, contam com mais de 80 milhões de usuários registrados em suas aplicações [DeveloperAnalytics 2010].

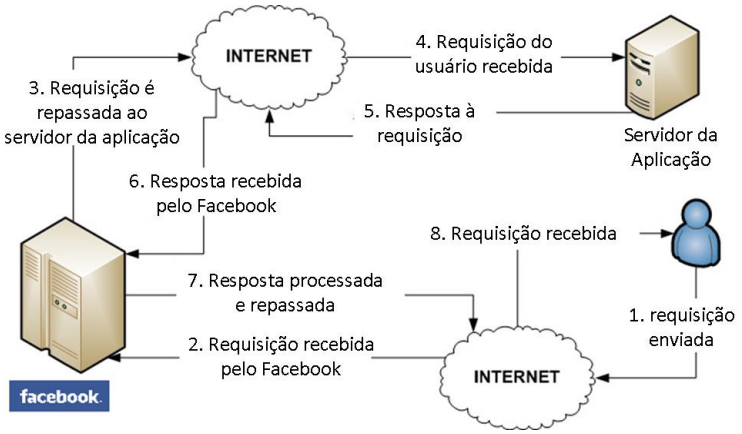


Figura 1.14. Funcionamento de Aplicações no Facebook e no Orkut

A Figura 1.14 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 a repassa ao servidor da aplicação. Então, o servidor da aplicação manda de volta a resposta ao servidor da rede social, que a repassa ao cliente [Nazir et al. 2009].

[4] Uma grande lista de aplicações do Facebook pode ser encontrada na seguinte referência [Facebook 2010a].

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, tais como lista de amigos e atividades executadas durante uma sessão. Alguns trabalhos fizeram uso dessa estratégia para estudar os usuários de redes sociais online. Nazir e colaboradores [Nazir et al. 2008] 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, os mesmos autores estudaram várias características relacionadas ao desempenho de suas aplicações no Facebook [Nazir et al. 2009].

1.5. Extração de Informação

A extração e o tratamento de informação a partir dos dados coletados são etapas essenciais para permitir diferentes análises incluindo identificação de padrões de comportamento de usuários em diferentes sistemas, identificação de tópicos de interesse dos usuários, etc. Nesta seção, discutimos brevemente os principais desafios relacionados a essas etapas. Em particular, a identificação de entidades nomeadas, tais como pessoas, organizações e produtos, encontradas em porções de texto, assim como a derivação de relacionamentos entre estas entidades, representam importantes problemas, com diversas aplicações interessantes. O nosso objetivo nesta seção não é detalhar todas as particularidades e soluções para estes problemas, que são tipicamente abordados por pesquisadores de outras áreas tais como Bancos de Dados, Recuperação de Informação, Mineração de Dados e Processamento de Linguagem Natural. Entretanto, reconhecendo a necessidade de expertises complementares para o estudo de redes sociais online, achamos benéfico para o leitor o contato, ainda que superficial, com estes tópicos.

1.5.1. Visão Geral

As redes sociais online são atualmente importantes plataformas para produção, processamento e fluxo de informação. Tal informação pode se originar dentro das redes ou fora delas e pode ser utilizada como fonte primária ou complementar para derivar conhecimento sobre a própria rede, seus membros, seus temas, suas comunidades, etc. No entanto, esta informação quase sempre ocorre em um formato textual, não estruturado, em linguagem natural e até mesmo em estilo telegráfico ou informalmente codificado. Isso é um grande empecilho para que esta informação possa ser processada de maneira automática e para que dela se possa derivar conhecimento latente. Como exemplo, uma mesma entidade (p.ex: uma pessoa ou uma localidade) pode ser referenciada de várias formas, devido às variações introduzidas por diferentes formas de grafia, aspectos regionais ou culturais, uso de abreviaturas, erros tipográficos e outras razões associadas com o uso convencional.

Em contextos já bastante explorados como sítios e páginas da Web, técnicas de áreas como Recuperação de Informação, Mineração de Dados e Pro-

cessamento de Linguagem Natural têm sido aplicadas com muito sucesso para extrair informação e derivar conhecimento de *corpus* textuais. No entanto, no contexto de redes sociais online, este *corpus* tem uma natureza totalmente diversa. De fato, nestas redes, o *corpus* textual é formado por micro-documentos como *tweets*, *blog posts*, *comentários*, *feeds*, *tags*, cujo tratamento deve ser necessariamente diferente do que é normalmente aplicado a, por exemplo, páginas Web. Além disto, dado o caráter informal de várias informações disponibilizadas em redes sociais online, técnicas de processamento de linguagem natural são difíceis de serem aplicadas devido à ausência de padrões linguísticos. Além disso, nos *corpora* textuais encontrados em redes sociais online existe muito lixo e ruído informacional (palavras escritas erradas, de baixa qualidade sintática ou semântica) dificultando esta tarefa ainda mais.

Por exemplo, técnicas de extração aberta de informações na Web [Etzioni et al. 2008] que se baseiam em modelos linguísticos gerais de como relacionamentos são expressos em um idioma, são o estado da arte em extração de entidades e relacionamentos de páginas Web. No entanto, como os modelos utilizados nestas técnicas são construídos a partir de características linguísticas através da identificação de partes de discurso nos documentos-alvo, elas dificilmente poderiam ser aplicadas em micro-documentos (por exemplo, *tweets*), onde estas características podem não se apresentar.

Apesar das dificuldades mencionadas, a extração de informações contidas no *corpus* textual de redes sociais pode ajudar a responder de forma automática e efetiva questões como:

- Quem fala com quem sobre que assunto?
- Quem são os atores principais nas redes? Onde estão localizados?
- Que assuntos e tópicos emergem, se disseminam e desaparecem nos eco-sistemas sociais digitais?
- Que indivíduos e grupos promovem e suprimem estes assuntos e tópicos?
- Qual a polaridade (positiva, negativa ou neutra) das opiniões emitidas na rede sobre assuntos, pessoas, empresas, etc.?

Um Exemplo: Observatório da Web

Um exemplo de como a identificação de entidades é uma tarefa importante para a análise de redes sociais online é o *Observatório das Eleições 2010*[5], que é uma instância do *Observatório da Web* [Santos et al. 2010], projeto desenvolvido no âmbito do InWeb (Instituto Nacional de Ciência e Tecnologia para a Web).

Este observatório foi desenvolvido para monitorar, em tempo real, o que estava sendo veiculado sobre as eleições de 2010 nas várias mídias e pelos

[5] `http://www.observatorio.inweb.org.br/eleicoes2010`.

vários usuários da Web. O seu objetivo era ajudar a traçar um panorama do cenário eleitoral do ponto de vista das informações e das opiniões que circulavam na Web e nas redes sociais online, incluindo jornais, revistas, portais e o Twitter. Foi implementado como um portal utilizando dezenas de ferramentas inéditas de captura e análise de dados baseadas em código livre ou aberto.

As entidades correlatas ao contexto das eleições foram o alvo principal do monitoramento. Isso incluía, além dos candidatos à presidência, políticos com influência na eleição, como o ex-presidente Lula. Em muitos casos, o monitoramento era concentrado em eventos, ou seja, acontecimentos importantes no contexto observado, tais como um debate, por exemplo, e que pudessem ter um grande efeito no conteúdo das fontes observadas.

A partir da identificação das entidades no textos coletados em tempo real, é possível gerar uma série de produtos de análise e visualização. Um exemplo de um destes produtos é apresentado na Figura 1.15.

Figura 1.15. Exemplo de Visualização de Dados Gerados a Partir da Identificação de Entidades no Observatório das Eleições.

No observatório, antes da extração propriamente dita, é realizado um pré-processamento dos textos coletados, incluindo a padronização da codificação dos caracteres, a eliminação de código HTML, cabeçalhos e anúncios de páginas coletadas através de *feeds*, e o uso de métodos tradicionais de pré-processamento de textos [Manning et al. 2008] tais como a remoção de *stop words* (palavras de pouco valor informacional como artigos, preposições e conjunções) e o *stemming*, que consiste na extração dos radicais das palavras do texto. A identificação de entidades nos textos é feita através da ferramenta

Illinois Named Entity Tagger [Ratinov e Roth 2009], que utiliza técnicas de processamento de linguagem natural para identificar referências a entidades (pessoas, organizações, locais, etc.) em texto livre. Apos a fase de identificação, segue-se uma fase de desambiguação de entidades. Isso é necessário porque os métodos identificação de entidades têm dificuldade, em geral, de diferenciar "José Serra" de "Serra da Piedade", ou "Lula presidente" de "Lula Molusco". Para isso, um método de classificação foi utilizado para aprender a associar entidades a determinados contextos.

A Figura 1.16 ilustra um RSS *feed* processado para identificação de entidades pelo Observatório das Eleições. As *tags* são usadas como identificadores para distinguir as duas entidades em todos os textos processados.

A pré-candidata do PT à Presidência da República <Person2> Dilma Rousseff </Person2> , quer juntar ao seu redor o maior número de legendas que hoje estão na base aliada do presidente <Person1> Luiz Inácio Lula da Silva </Person1>

Figura 1.16. Exemplo de RSS Feed com Entidades Identificadas no Observatório das Eleições

1.5.2. Identificação de Entidades

O problema de identificação de entidades nomeadas (*Named Entity Recognition – NER*) consiste em encontrar palavras que ocorrem em um documento ou trecho de texto não estruturado e que façam referência a entidades do mundo real. Este problema tem sido estudado em vários contexto como identificação de nomes de pessoas e companhias em notícias, identificação de genes e proteínas e publicações biomédicas, etc. [Cohen e Sarawagi 2004]. A Figura 1.17(a) ilustra o resultado típico da aplicação de um método de extração de entidades.

Uma abordagem comum para o problema de NER é a de *rotulamento de sequências*. Um documento é representado como uma sequência **x** de tokens $x_1, \ldots, x_n$ e um *classificador* associa **x** a uma sequência paralela de rótulos **y** = $y_1, \ldots, y_N$, onde cada y_i é um rótulo pertencente a um conjunto Y. Uma atribuição correta dos rótulos permite a identificação direta das entidades. Por exemplo, na sequência ilustrada na Figura 1.17(b), cada token recebe um rótulo, sendo que um rótulo especial outro é associado aos tokens que não são partes de nomes de entidades.

A construção do classificador pode ser feita usando técnicas de aprendizagem de máquina, ou seja, utilizando dados de treinamento que representam exemplos de mapeamento de sequências **x** para sequências **y**. Isso é feito, em geral, através de documentos manualmente anotados de forma similar ao que está ilustrado na Figura 1.17(b). Estes classificadores são gerados pelo no aprendizado de modelos sequenciais tais como

<pessoa> Odorico Paraguaçu <pessoa> foi <cargo> prefeito </cargo> eterno de <local> Sucupira </local>, cidadezinha localizada em algum lugar da <local> Bahia </local>, a qual governou com muita sabedoria e inteligência. Dotado de uma habilidade incrível com as palavras fruto de seu curso na <organização> Universidade de Sourbone </organização>), cativava as massas numa época em que os comícios não tinham nem fanfarra, quanto mais bandas completas.[6]

(a)

Odorico	*Paraguaçu*	*foi*	*prefeito*	*eterno*	*de*	*Sucupira*
pessoa	pessoa	outro	cargo	outro	outro	local

(b)

Figura 1.17. Exemplo de um trecho de texto com entidades identificadas (a) e como uma sequência rotulada (b)

Hidden Markov Models [Freitag e McCallum 2000] ou *Conditional Random Fields* [Lafferty et al. 2001] ou suas variações (por exemplo, [Cortez et al. 2010]).

A maioria dos métodos de aprendizado explora de alguma forma a característica sequencial do processo de classificação, de forma que rotulo atribuído a um token dependerá dos rótulos atribuídos aos tokens em sua vizinhança. Por exemplo, considerando que nomes de pessoas no Brasil têm usualmente entre três e quatro tokens, se y_i recebe o rótulo Pessoa, então a probabilidade de y_{i+1},. . . ,y_{i+3} receberem o mesmo rótulo aumenta, enquanto que a probabilidade de y_{i+4} em diante receberem o este rótulo diminui.

Em outras palavras, seja **x** uma variável aleatória sobre uma sequencia de dados a serem rotulados. Cada x_i é chamado de uma *observação*. Seja **y** uma variável aleatória sobre uma sequencia de rótulos correspondentes a estes dados. Desta forma, o problema de identificação de entidades se reduz ao problema de atribuir os rótulos em **y** à sequência de token **x** de forma a maximizar a probabilidade condicional P(**y**|**x**).

Assim, abordagem baseadas em aprendizagem de máquina para NER consistem no aprender modelos sequenciais tais como *Hidden Markov Models* [Freitag e McCallum 2000] ou *Conditional Random Fields* [Lafferty et al. 2001] ou suas variações (por exemplo, [Cortez et al. 2010]).

De forma geral, o modelo de *Conditional Random Fields* representa o estado da arte para extração de entidades. Como descrito em [Lafferty et al. 2001], P(**y**|**x**) neste modelo é estimada da seguinte forma:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp(\sum_{i=1}^{n}(\sum_{k} \lambda_k f_k(y_i, y_{i-1}, \mathbf{x})) + (\sum_{\ell} \mu_\ell g_\ell(y_i, \mathbf{x})))$$
$$= \frac{1}{Z(\mathbf{x})} \exp(\sum_{i=1}^{n}(\lambda^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}) + \mu^T \mathbf{g}(y_i, \mathbf{x})))$$

onde

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp(\sum_{i=1}^{n}(\lambda^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}) + \mu^T \mathbf{g}(y_i, \mathbf{x})))$$

e **f**,**g** são conjuntos de funções pré-definidas que computam valores binários para *features* ou características observadas entre um par de rótulos na sequência (**f**) e em um determinado token na sequência (**g**). Os valores de $\lambda_1, \lambda_2, \lambda_3, \ldots, \mu_1, \mu_2, \mu_3, \ldots$ são parâmetros que indicam o "peso" de cada feature. Exemplos de features são: (1) uma função g que retorna verdadeiro se o token em x_i começa com maiúsculo e o rótulo y_i é pessoa; (2) f retorna verdadeiro se o rótulo y_{i-1} é pessoa e o rótulo y_i é outro. Também é comum utilizar dicionários ou *gazeteers* para gerar features. Assim, por exemplo, determinada função g pode retornar verdadeiro se token em x_i se encontra em um *gazeteers* e y_i é lugar;

O treinamento de um modelo CRF consiste em fornecer um conjunto de sequências de treino corretamente rotuladas e determinar os valores dos parâmetros $\lambda_1, \lambda_2, \lambda_3, \ldots, \mu_1, \mu_2, \mu_3, \ldots$ que maximizam $P(\mathbf{y}|\mathbf{x})$ para estes exemplos. Existem várias propostas na literatura para ajustar estes parâmetros. Uma delas é o algoritmo de gradiente descendente [Dietterich et al. 2004].

O processo de identificação de entidades propriamente dito se dá pela execução do processo de inferência do modelo CRF treinado. Assim, dados os parâmetros λ e μ, deseja-se encontrar a instância de rótulos $\mathbf{y}^*$ que maximiza $P(\mathbf{y}|\mathbf{x})$, ou seja,

$$\mathbf{y}^* = \arg\max_{y} \exp(\sum_{i=1}^{n}(\lambda^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}) + \mu^T \mathbf{g}(y_i, \mathbf{x}))).$$

Para isso é geralmente utilizado o Algoritmo de Viterbo que utiliza programação dinâmica fazendo somas parciais das matrizes de forma incremental [Freitag e McCallum 2000].

Como já mencionado, no Observatório das Eleições utilizamos um sistema chamado *Illinois Named Entity Tagger* [Ratinov e Roth 2009] para fazer o reconhecimento de entidades. Este sistema se baseia no modelo CRF e tem atualmente o melhor desempenho comparativo entre os sistemas de NER. Neste sistema, o treinamento é feito para que o sistema reconheça o entidades com vários tokens, distinguindo os tokens que ocorrem no ínicio, no final da ocorrência da entidade e também os tokens ocorrendo dentro e fora desta ocorrência. Os autores demonstram que esta configuração é melhor que o esquema de

treinamento tradicional (início, dentro e fora). O sistema é capaz de reconhecer quatro tipos de entidades: pessoas, organizações, locais e miscelânea. Utiliza features baseadas em fontes externas, tais como *gazeteers* e dados obtidos da wikipedia.

1.5.3. Resolução de Entidades

Uma vez que as referências a entidades são identificadas em um certo *corpus*, algumas aplicações podem requerer ainda que sejam estabelecidas correspondências entre referências distintas que se refiram a uma mesma entidade do mundo real. Este problema é conhecido como *Resolução de Entidades* [Bhattacharya e Getoor 2007].

Existem na realidade dois subproblemas relacionados ao problema de resolução de entidades, os quais ilustramos pelos trechos de texto apresentados na Figura 1.18.

D_1 *O Acre é o estado mais à oeste do Brasil, seu território é inteiramente recoberto pela Floresta Amazônica. É também berço de grandes nomes como <pessoa>Marina Silva</pessoa>, política, e <pessoa>Glória Perez</pessoa>, novelista.*

D_2 *Ao lado da então ministra da Casa Civil, <pessoa>Dilma Rousseff</pessoa>, <pessoa>Lula</pessoa> acabou recebendo uma amostra do óleo na <pessoa>Marina</pessoa> da <pessoa>Glória</pessoa>, no Rio, para onde o evento foi transferido.*

D_3 *A candidata <pessoa>Dilma</pessoa>, vencedora do primeiro turno das eleições, telefonou hoje para a candidata <pessoa>Marina</pessoa> do PV e a parabenizou pelo seu desempenho nas urnas.*

Figura 1.18. Exemplo de um Trecho de Texto com Entidades Identificadas (a) e como uma Sequência Rotulada (b).

O primeiro subproblema consiste em determinar o conjunto de referências distintas no corpus que são utilizadas para se referir a mesma entidade no mundo real. Este é caso de "Marina Silva" em D_1 e "Marina" em D_3, e também de "Dilma Rousseff" em D_2 e "Dilma" em D_3. Este subproblema é conhecido como *Identificação* [Bhattacharya e Getoor 2007] ou *Coreferência* [Hobbs 1979]. Note que este problema também pode ser causado por erros de escrita, formatação, etc., ou mesmo pelo uso de apelidos (p.ex., "Dilma") e anáforas (p.ex., "a ministra").

O segundo subproblema consiste em distinguir quando referências muito similares, ou até mesmo exatamente iguais, se referem a diferentes entidades

do mundo real. Esse é o caso de "Marina" em D_3 e em D_2 e também de "Glória Perez" em D_3 e "Glória" em D_2. Este problema é chamado de *Desambiguação* [Bhattacharya e Getoor 2007]. Note que neste caso específico o problema foi causado por uma falha do processo de identificação de entidades em D_2. Tais problemas são comuns e, dependendo do domínio de aplicação, podem ser exacerbados pelo uso de abreviações.

A solução do problema de resolução de entidades pode ser determinante para a qualidade dos resultados obtidos a partir da análise das entidades extraídas. Por outro lado, se neglicenciado, este problema pode comprometer o conhecimento derivado destes resultados. Por exemplo, as análises realizadas pelo Observatório das Eleições poderiam ser seriamente afetadas se referências ao candidato "José Serra" e à "Serra da Piedade" não sofressem desambiguação.

Na literatura recente, o problema de resolução entidades tem sido tratado através do cálculo da similaridade entre os atributos associados às entidades (no caso de bancos de dados) [de Freitas et al. 2010] ou utilizando, quando possível, grafos de co-ocorrência de entidades [Bhattacharya e Getoor 2007]. Em *corpora* textuais, ferramentas linguísticas têm sido usadas para solução deste problema [Ng 2007]. No caso do Observatório das Eleições, o problema foi tratado através de uma solução simples baseada em classificação usando centroides [Han e Karypis 2000], que neste caso são usado para representar o contexto em que determinada entidade é tipicamente encontrada em termos de vocabulário recorrente.

Seja r_i uma referência a entidade identificada no corpus que está sendo processado. Seja D_i o conjunto de documentos (i.e., *tweets*, comentários do *YouTube*, etc.) nos quais r_k ocorre. Para cada documento $d_{i,k} \in D_i$, extraímos os termos (palavras) que nele se encontram e pré-processamos fazendo remoção de *stop-word* e *steming*. Tomamos então estes termos e geramos vetores $\vec{d_{i,k}}$ de acordo com o modelo vetorial [Baeza-Yates e Ribeiro-Neto 1999], considerando que o vocabulário do corpus corresponde ao espaço de termos.

Um centroide $\vec{c_i}$ que representa D_i é definido como

$$\vec{c_i} = \frac{1}{|D_i|} \sum_{\vec{d_{i,k}} \in D_j} \vec{d_{i,k}}$$

Chamamos o centroide $\vec{c_i}$ de *Contexto* da referência r_i. Como sugerido em [Han e Karypis 2000], podamos $\vec{c_i}$ para utilizar somente os K termos mais frequentes[7], evitando assim a introdução de ruído e acelerando o processamento.

Para fins de resolução de entidades, consideramos que duas referências r_i e r_j são co-referentes, ou seja, se referem a mesma entidade, somente se r_i é *similar* a r_j e o cosseno entre os vetores $\vec{c_i}$ e $\vec{c_j}$, $\cos(\vec{c_i}, \vec{c_j})$, é menor ou igual a

[7] Em experimentos de calibração determinamos $K = 100$ como um valor adequado.

um limiar ϵ. Para isso, a similaridade entre r_i (e.x., "Marina Silva") e r_j ("Marina da Glória") pode ser calculada usando uma função de similaridade de strings e valor de ϵ pode ser determinado empiricamente.

1.6. Bases de Dados Disponíveis na Web

Vários trabalhos que coletaram dados de redes sociais online oferecem disponibilizam os dados coletados para a comunidade acadêmica. A seguir, algumas bases contendo dados públicos disponíveis na Web são relacionadas.

- Dados sobre **Orkut**, **Flickr**, **LiverJournal** e **YouTube**. Foram utilizados no trabalho [Mislove et al. 2007] e estão disponíveis em `http://socialnetworks.mpi-sws.org/data-imc2007.html`
- Dados sobre os vídeos de duas categorias inteiras do **YouTube**. Coletados em 2007 e utilizados no artigo [Cha et al. 2007]. Disponível em `http://an.kaist.ac.kr/traces/IMC2007.html`.
- Dados do **Flickr** coletados ao longo do tempo. Esses dados foram utilizados na referência [Mislove et al. 2008] e estão disponíveis em `http://socialnetworks.mpi-sws.org/data-wosn2008.html`.
- Dados sobre a popularidade de vídeos do **YouTube** com registro do crescimento e fontes dos acessos ao longo do tempo. Foram utilizados na referência [Figueiredo et al. 2011] e estão disponívis em `http://vod.dcc.ufmg.br/traces/youtime/`.
- Grafo completo contendo 55 milhões de usuários do **Twitter** e cerca de 1.8 bilhões de tweets. Esses dados foram utilizados nas referências [Cha et al. 2010, Benevenuto et al. 2010a] e estão disponíveis no endereço `http://twitter.mpi-sws.org/`.
- Dados sobre o grafo de amizade e postagens de amostra de usuários do **Facebook**. Utilizados na referência [Viswanath et al. 2009a] e disponíveis em `http://socialnetworks.mpi-sws.org/data-wosn2009.html`.
- Coleção de usuários do **YouTube**, manualmente classificados como spammers, promoters ou usuários legítimos. Esses dados foram utilizados nos seguintes trabalhos [Benevenuto et al. 2009a, Benevenuto et al. 2008b, Langbehn et al. 2010]. A base de dados está disponível em `http://homepages.dcc.ufmg.br/~fabricio/testcollectionsigir09.html`.
- Coleção de dados do **Del.icio.us** e do **Digg**. Coleção utilizada em diferentes artigos e disponível em `http://www.public.asu.edu/~mdechoud/datasets.html`.

1.7. Conclusões

Redes sociais online se tornaram extremamente populares e parte do nosso dia a dia, causando o surgimento de uma nova onda de aplicações dis-

poníveis 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 online 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, aplicações de 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 (InWeb), pelo CNPq, FAPEMIG e UOL (www.uol.com.br).

Referências

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

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

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

[Albert et al. 1999] Albert, R., Jeong, H. e Barabasi, A. (1999). Diameter of the world wide web. *Nature*, 401:130–131.

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

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

[Baeza-Yates e Ribeiro-Neto 1999] Baeza-Yates, R. e Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

[Baluja et al. 2008] Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D. e Aly, M. (2008). Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *World Wide Web Conference (WWW)*, pp. 895–904.

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

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

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

[Benevenuto et al. 2010b] Benevenuto, F., Pereira, A., Rodrigues, T., Almeida, V., Almeida, J. e Gonçalves, M. (2010b). Characterization and analysis of user profiles in online video sharing systems. *Journal of Information and Data Management*, 1(2):115–129.

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

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

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

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

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

[Bhattacharya e Getoor 2007] Bhattacharya, I. e Getoor, L. (2007). Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, 1.

[Binder et al. 2009] Binder, J., Howes, A. e Sutcliffe, A. (2009). 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)*, pp. 965–974.

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

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

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

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

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

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

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

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

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

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

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

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

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

[Choi e Varian 2009] Choi, H. e Varian, H. (2009). Predicting the present with google trends. `http://bit.ly/2iujV3`. Accessed in Jan/2011.

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

[Cohen e Sarawagi 2004] Cohen, W. e Sarawagi, S. (2004). Exploiting Dictionaries in Named Entity Extraction: Combining Semi-Markov Extraction Processes and Data Integration Methods. In *Proc. 10th ACM SIGKDD Intl. Conf. on Knowl. Discov. and Data Mining*, pp. 89–98.

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

[Cortez et al. 2010] Cortez, E., da Silva, A. S., Gonçalves, M. A. e de Moura, E. S. (2010). Ondux: on-demand unsupervised learning for information extraction. In *Proceedings of the 2010 international conference on Management of data*, SIGMOD '10, pp. 807–818.

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

[de Freitas et al. 2010] de Freitas, J., Pappa, G. L., da Silva, A. S., Gonçalves, M. A., de Moura, E. S., Veloso, A., Laender, A. H. F. e de Carvalho, M. G. (2010). Active learning genetic programming for record deduplication. In *IEEE Congress on Evolutionary Computation*, pp. 1–8.

[DeveloperAnalytics 2010] DeveloperAnalytics (2010). Developer analytics. `http://www.developeranalytics.com`. Acessado em Março/2010.

[Dietterich et al. 2004] Dietterich, T. G., Ashenfelter, A. e Bulatov, Y. (2004). Training conditional random fields via gradient tree boosting. In *Proceedings of the Twenty-first International Conference on Machine Learning*.

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

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

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

[eMarketer 2007] eMarketer (2007). Social network marketing: ad spending and usage. *EMarketer Report*, 2007. `http://tinyurl.com/2449xx`. Acessado em Março/2010.

[Etzioni et al. 2008] Etzioni, O., Banko, M., Soderland, S. e Weld, D. S. (2008). Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

[Facebook 2010a] Facebook (2010a). Facebook application directory. `http://www.facebook.com/apps`. Acessado em Março/2010.

[Facebook 2010b] Facebook (2010b). Facebook platform. http://developers.facebook.com. Acessado em Março/2010.

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

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

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

[Figueiredo et al. 2011] Figueiredo, F., Benevenuto, F. e Almeida, J. (2011). The tube over time: Characterizing popularity growth of youtube videos. In *Proceedings of the 4th ACM International Conference of Web Search and Data Mining (WSDM)*.

[Freitag e McCallum 2000] Freitag, D. e McCallum, A. (2000). Information Extraction with HMM Structures Learned by Stochastic Optimization. In *Proc. of the 17th Nat. Conf. on Art. Intell. and 12th Conf. on Innov. Appl. of Art. Intell.*, pp. 584–589.

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

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

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

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

[Ginsberg et al. 2009] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. e Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–4.

[Gnuplot 2010] Gnuplot (2010). Gnuplot homepage. http://www.gnuplot.info/. Acessado em Agosto/2010.

[Gomes et al. 2007] Gomes, L., Almeida, J., Almeida, V. e Meira, W. (2007). Workload models of spam and legitimate e-mails. *Performance Evaluation*, 64(7-8).

[Gomide et al. 2011] Gomide, J., Veloso, A., Jr., W. M., Benevenuto, F., Almeida, V., Ferraz, F. e Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *ACM SIGWEB Web Science Conference (WebSci)*.

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

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

[Han e Karypis 2000] Han, E.-H. e Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 424–431.

[Hobbs 1979] Hobbs, J. (1979). Coherence and coreference*. *Cognitive science*, 3(1):67–90.

[Joinson 2008] Joinson, A. (2008). 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)*, pp. 1027–1036.

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

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

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

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

[Lafferty et al. 2001] Lafferty, J. D., McCallum, A. e Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML'01, pp. 282–289.

[Langbehn et al. 2010] Langbehn, H., Ricci, S., Gonçalves, M., Almeida, J., Pappa, G. e Benevenuto, F. (2010). A multi-view approach for detecting spammers and content promoters in online video social networks. *Journal of Information and Data Management*, 1(3):1–16.

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

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

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

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

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

[Lovász 1993] Lovász, L. (1993). Random Walks on Graphs: A Survey. *Combinatorics*, 2:1–46.

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

[Manning et al. 2008] Manning, C. D., Raghavan, P. e Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

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

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

[Mislove 2009] Mislove, A. (2009). *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. Tese de Doutorado, Rice University, Department of Computer Science.

[Mislove et al. 2008] Mislove, A., Koppula, H., Gummadi, K., Druschel, P. e Bhattacharjee, B. (2008). Growth of the flickr social network. In *ACM SIGCOMM Workshop on Social Networks (WOSN)*, pp. 25–30.

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

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

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

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

[New York Times 2010a] New York Times (2010a). Uploading the avant-garde. http://www.nytimes.com/2009/09/06/magazine/06FOB-medium-t.htm. Acessado em Julho/2010.

[New York Times 2010b] New York Times (2010b). 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.

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

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

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

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

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

[Ng 2007] Ng, V. (2007). Shallow semantics for coreference resolution. In *Int'l Joint Conference on Artificial Intelligence*, pp. 1689–1694.

[Nielsen Online 2010] Nielsen Online (2010). Social networks & blogs now 4th most popular online activity, 2009. http://tinyurl.com/cfzjlt. Acessado em Março/2010.

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

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

[Ratinov e Roth 2009] Ratinov, L. e Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pp. 147–155.

[Ribeiro e Towsley 2010] Ribeiro, B. e Towsley, D. (2010). Estimating and Sampling Graphs with Multidimensional RandomWalks. In *Proceedings ACM SIGCOMM Internet Measurement Conference*.

[Robert e Casella 2005] Robert, C. P. e Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer-Verlag, second edição.

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

[Sakaki et al. 2010] Sakaki, T., Okazaki, M. e Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pp. 851–860.

[Santos et al. 2010] Santos, W., Pappa, G., Jr., W. M., Guedes, D., Veloso, A., Almeida, V., Pereira, A., Guerra, P., Silva, A., Mourão, F., Magalhães, T., Machado, F., Cherchiglia, L., Simões, L., Batista, R., Arcanjo, F., Brunoro, G., Mariano, N., Magno, G., Ribeiro, M., Teixeira, L., Silva, A., Reis, B. e Silva, R. (2010). Observatório da web: Uma plataforma de monitoração, síntese e visualização de eventos massivos em tempo real. In *Anais do XXXVII Seminário Integrado de Hardware e Software*, SEMISH'10, pp. 110–120.

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

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

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

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

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

[Tumasjan et al. 2010] Tumasjan, A., Sprenger, T., Sandner, P. e Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment.

[Viswanath et al. 2009a] Viswanath, B., Mislove, A., Cha, M. e Gummadi, K. (2009a). On the evolution of user interaction in facebook. In *ACM SIGCOMM Workshop on Social Networks (WOSN'09)*.

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

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

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

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

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

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

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

[Yahoo! Pipes 2010] Yahoo! Pipes (2010). Yahoo! pipes. `http://pipes.yahoo.com/pipes`. Acessado em Agosto/2010.

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

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

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