Caracterização de Carga de uma Rede Social Baseada em Localização

Theo Silva Lins

Orientador: Fabrício Benevenuto
Universidade Federal de Ouro Preto

Dissertação submetida ao
Instituto de Ciências Exatas e Biológicas da
Universidade Federal de Ouro Preto
para obtenção do título de Mestre em Ciência da Computação

Dedico este trabalho a todos os professores.

Caracterização de Carga de uma Rede Social Baseada em Localização

Resumo

Recentemente, tem ocorrido uma grande popularização das redes sociais baseadas em localização, como o FourSquare e o Gowalla, onde usuários podem criar e compartilhar referências a locais reais, fazer check-in nesses locais e adicionar comentários e dicas a locais do sistema. Parte dessa popularidade é devida à facilidade de acesso à Internet através de dispositivos móveis dotados de GPS. Há uma grande diferença entre publicar conteúdo em redes sociais e redes sociais baseadas em localização (LBSN). LBSNs fornecem uma nova estrutura social em redes composta de indivíduos ligados pelas suas localizações no mundo físico. Apesar do grande interesse, pouco se sabe sobre os padrões de acesso em novos sistemas de redes sociais como LBSNs e como se diferem dos padrões de acesso dos sistemas tradicionais. Este trabalho tem como objetivo dar o primeiro passo no entendimento dessa mudança. Para isso, utilizamos um conjunto de dados obtidos junto ao Apontador, um sistema brasileiro com características semelhantes à do FourSquare e Gowalla, onde usuários compartilham informações sobre localizações e podem navegar por essas localizações. Como resultados, foram identificados modelos que descrevem características das sessões de usuários, padrões com os quais requisições chegam ao servidor, além do perfil de acesso de usuários ao sistema.

Workload Characterization of a Location Based Social Network

Abstract

Recently, there has been a large popularization of location-based social networks, such as FourSquare and Gowalla, in which users can create and share locations, check-in in these places using smart phones, and add comments and tips about places within the system. Part of that popularity is due to easy access to the internet through mobile devices with GPS. There is a big different between publishing content through social networks and through location-based social networks (LBSNs). LBSNs provide a new social structure derived from individuals locations in the physical world. Despite considerable interest, little is known about the patterns of access to new systems of social networks like LBSNs and how they differ from the patterns of traditional systems. This paper aims to take the first step in understanding this change. To that end, we use a dataset obtained from Apontador, a Brazilian system with characteristics similar to FourSquare and Gowalla, where users share information about their locations and can navigate on existent system locations. As results, we identified models that describe unique characteristics of the user sessions on this kind of system, patterns in which requests arrive on the server as well as the user navigation profile within the system.

Agradecimentos

Em primeiro lugar quero agradecer a Deus por ter me dado essa oportunidade.

Agradeço minha família, principalmente minha Mãe pelo apoio, educação e valores que me que permitiram chegar até aqui. A Elô agradeço pela compreensão, carinho e amor incondicional.

Agradeço ao meu orientador Fabrício pela dedicação prestada durante o desenvolvimento desse trabalho e pelo grande conhecimento adquirido ao longo desse mestrado. A todos os professores e técnicos do DECOM agradeço pelo profissionalismo e contribuição para a realização deste trabalho.

Agradeço a todos os técnicos e professores do ICEA por todo o apoio e disponibilidade que me proporcionaram, e aos técnicos e bolsistas do NTI, pela boa disposição que sempre manifestaram.

Agradeço aos amigos que fiz na graduação e no mestrado, que compartilharam os momentos de aprendizado, dúvidas, desespero, distração e solidariedade.

Agradeço a todos os meus amigos, especialmente os do gole e do futebol, que embora não tenham me ajudado diretamente neste trabalho, me proporcionaram preciosos momentos de alegria e distração.

Agradeço ao Apontador pelos dados fornecidos, que tornaram possível a realizacão desse trabalho.

Muito Obrigado a todos.

Sumário

Lista de Figuras

Lista de Tabelas

"Pra quem tem pensamento forte o impossível
é só questão de opinião"

— Alexandre Magno Abrão

Lista de Siglas, Acrônimos e Abreviaturas

WWW	*World Wide Web*
URL	*Uniform Resource Locator*
HTTP	*HyperText Transfer Protocol*
LBSN	*Location-Based Social Network*
GPS	*Global Positioning System*
P2P	*Par-a-Par*
CDF	*Cumulative Distribution Function*
CCDF	*Complementary Cumulative Distribution Function*
PDF	*Probability Density Function*
UBMG	*User Behavior Model Graph*
XML	*eXtensible Markup Language*

Capítulo 1

Introdução

Desde o seu início a Internet recebeu uma grande onda de aplicações, incluindo a Web e Par-a-Par, em que os diferentes padrões de tráfego ajudaram a remodelar a sua infra-estrutura. Recentemente, aplicações de redes sociais online tornaram-se aplicações extremamente populares. Segundo o site Alexa.com, redes sociais como Facebook e Twitter estão entre os 10 *sites* mais visitados no mundo, tanto em termos de usuários distintos, como em termos de tempo gasto nos sites. Com mais de 1 bilhão de usuários, se o Facebook fosse um país, seria o terceiro país mais populoso do mundo [21].

Várias redes sociais online possuem algumas características em comum. Geralmente, elas permitem aos usuários compartilharem informações com amigos e disponibiliza uma página com o perfil do usuário, que pode publicar ou atualizar qualquer conteúdo no seu perfil. Os conteúdos variam de simples mensagens de texto a arquivos multimídias, como fotos ou vídeos. Para incentivar os usuários a compartilharem conteúdo, as redes sociais fazem atualizações disponíveis aos usuários imediatamente após seus amigos compartilhar o conteúdo. Assim, não só os usuários gastam muito tempo nesses sistemas, mas também criam enormes quantidades de conteúdo. Como um exemplo, o serviço de compartilhamento de fotos no Facebook é o maior repositório de fotos da Web, contendo mais de 60 milhões de imagens [20]. O YouTube recebe 24 horas de vídeo por minuto [22].

Em particular, há um tipo especial de sistema de rede social chamado de Rede Sociais Baseadas em Localização (LBSN), que está atraindo novos usuários em ritmo exponencial. LBSN, como Foursquare[1] e Gowalla[2], permitem aos usuários compartilharem sua

[1] www.foursquare.com
[2] http://www.gowalla.com/

localização geográfica com os amigos através de *smartphones* equipados com GPS, busca de lugares interessantes, bem como postagem de dicas sobre os locais existentes. Tem sido relatado que, hoje em dia, quase um em cada cinco donos de smartphones acessam esse tipo de serviço por dispositivos móveis [1].

1.1 Problemas e Objetivos

Nesta seção são descritos os principais problemas causados pela mudança ocorrida na Web, os quais, motivam essa dissertação. Em seguida, nossos objetivos são apresentados.

Intuitivamente, há uma diferença crucial entre a publicação tradicional de conteúdo na Web e compartilhar conteúdo por meio de redes sociais e redes sociais baseadas em localização. Quando as pessoas compartilham conteúdo na Web, elas tipicamente tornam o conteúdo acessível a qualquer usuário da Web. Quando os usuários compartilham o conteúdo em redes sociais online, muitas vezes têm a intenção de atingir um determinado público, como amigos ou seguidores. Finalmente, quando usuários compartilham o conteúdo em LBSNs, muitas vezes tem a intenção de atingir um público local, que pode incluir ou não amigos. Assim, LBSNs fornecem uma nova estrutura social composta de indivíduos ligados pela interdependência derivada de suas localizações no mundo físico.

Esta diferença crucial pode afetar importantes propriedades do tráfego que chega aos sistemas LBSNs, que, por sua vez, podem afetar diferentes aspectos da concepção do sistema, tais como mecanismos de cache e distribuição de conteúdo. Mais importante, dado o crescimento exponencial dos vários sistemas sociais, é razoável considerar que esses sistemas têm o poder de remodelar o tráfego da Internet no futuro. Na verdade, as redes sociais têm sido um importante tópico de discussão na atividade conhecida como a **Internet do Futuro**, um movimento que visa a formulação e avaliação de arquiteturas alternativas para as mudanças que a Internet pode precisar no futuro [24]. Apesar do grande interesse, pouco se sabe sobre os padrões de acesso em novos sistemas de redes sociais como o LBSNs e como eles diferem dos padrões de acesso dos sistemas tradicionais.

Este trabalho tem como objetivo dar o primeiro passo nesta direção, fornecendo uma ampla caracterização de carga de trabalho de uma rede social baseada em localização muito popular no Brasil, chamada Apontador[3]. Apontador inclui as principais carac-

[3]www.apontador.com.br

terísticas dos sistemas como Foursquare e Gowalla. Ele permite aos usuários procurar por lugares, registrar novos locais, postar dicas sobre lugares existentes, e fazer *check-in* em locais utilizando *smartphones*.

1.2 Contribuições do Trabalho

A seguir são apresentadas as principais contribuições dessa dissertação.

Através de uma grande base de dados obtida a partir do sistema Apontador apresentamos uma caracterização de carga de trabalho das sessões e requisições que chegam a esse servidor de LBSN. Obtivemos um conjunto de dados contendo cliques dos usuários, que descrevem no nível de sessão por 64.309.252 de solicitações HTTP extraídas durante um período de um mês.

Usando essa base de dados, fornecemos uma série de análises que definem a sessão do usuário no contexto do tráfego e modela os padrões do tráfego e sessões da carga de trabalho. Particularmente, examinamos com que frequência as pessoas se conectam na LBSNs, por quanto tempo e como os usuários interagem nos locais. Em seguida, identificamos os melhores modelos para uma série de medidas para requisições e sessões, tais como intervalo entre as chegadas de sessões, distribuição do tamanho da sessão e caracterização da navegação do usuários em uma sessão. Dentre as principais características identificadas, podemos destacar:

- Uma sessão típica de um usuário de um sistema de rede social baseada em localização tem 30 minutos, um valor 3 vezes maior em comparação com os sistemas tradicionais da Web.

- A distribuição da popularidade de acessos às localizações segue uma distribuição Lognormal. Outros sistemas avaliados seguem uma distribuição Zipf.

- O *ranking* de atividade dos usuários em função do número de requisições por usuários e sessões criadas por usuários seguem, respectivamente, uma distribuição Weibull e uma lei de potência.

- A chegada de requisições ao servidor segue um padrão com muita intensidade durante o dia e pouca intesidade durante a noite.

- As distribuições do intervalo de requisições e intervalo entre sessões são melhores modeladas com uma distribuição Weibull e Gamma respectivamente.
- Os usuários autenticados (*logged in*) tendem a realizar mais atividades nos locais em uma mesma sessão, enquanto os demais não-logados tendem a não realizar outras atividades.

1.3 Organização dos Capítulos

O restante da dissertação está organizado da seguinte forma. O **Capítulo 2** aborda os trabalhos relacionados, mostrando estudos com caracterizações de cargas de trabalho, redes sociais onlines e redes sociais baseadas em localização. Em seguida, o **Capítulo 3** mostra informações e estatísticas sobre as bases de dados utilizadas neste trabalho, bem como a base de dados coletada da LBSN Apontador. Depois, no **Capítulo 4**, apresentamos uma caracterização de carga de trabalho da base de dados. Finalmente, no **Capítulo 5** concluímos o trabalho e discutimos os trabalhos que poderão ser realizados futuramente.

Capítulo 2

Trabalhos Relacionados

O processo de caracterização de carga é importante para o entendimento e aprimoramento de sistemas Web. Há vários estudos que apresentam caracterizações de carga de trabalho de diferentes tipos. Um estudo seminal sobre a caracterização de servidores web foi apresentado em [4]. Nesse trabalho foram utilizados os logs dos servidores Web da copa do mundo de 1998, onde a maior parte dos acessos eram direcionados a um conjunto pequeno de arquivos estáticos, tornando estratégias de *caching* bastante eficientes. Em [7], Barford e Crovella aplicam uma série de observações de uso de servidores Web para criar uma ferramenta realista de geração de carga de trabalho, que imita um conjunto de usuários reais acessando um servidor. Arlitt e Williamson [5] também realizaram uma caracterização com base nos logs de servidores Web, estudo que mostra como encontrar invariantes que se aplicam a todo conjunto de dados. Estes invariantes são importantes, uma vez que representam modelos para a carga de servidores Web. Com base nos modelos obtidos, os autores ainda propõem melhorias sobre as questões do cache de armazenamento e do desempenho.

Alguns anos mais tarde, surgiram várias abordagens no sentido de caracterizar serviços de comércio eletrônico, onde podemos citar [29], [34] e [28], que caracterizaram as chegadas de requisições e sessões dos usuários, que determinaram o impacto sobre o desempenho e escalabilidade do sistema, mostrando que o cache é vital para garantir a escalabilidade de grandes sistemas de comércio eletrônico. Em [31] é possível observar uma diminuição no contéudo estático acessado pelos usuários, em comparação as caracterizações de servidores Web.

Usando dados de videos sob demanda, podemos citar [14] e [19], onde Veloso e

colaboradores mostraram uma análise da popularidade dos objetos e usuários e também suas diferenças. As análises feitas mostraram que os padrões de acessos de vídeos sob demanda são diferentes dos padrões de acesso dos servidores Web.

Krishnamurthy e colaboradores [15] apresentaram uma abordagem automatizada para a construção de cargas de trabalho sintéticas para sistemas baseados em sessões. Os autores fizeram um estudo experimental que investiga o impacto da carga de trabalho, e várias características que influenciam o desempenho de sistemas baseados em sessão. Outras caracterizações que contribuíram para os estudos foram [16], onde Duarte e colaboradores apresentaram uma caracterização completa dos padrões de acesso em blogs foi concluído que a natureza das interações entre usuários e objetos é fundamentalmente diferente em blogs do que a observada no conteúdo da Web tradicional. Benevenuto e colaboradores [9] fizeram uma análise da carga de trabalho de um serviço de compartilhamento de vídeos, apresentando uma caracterização das sessões e dos perfis de navegação dos usuários. Os resultados provêem um melhor entendimento do padrão de acesso dos usuários aos sistemas de compartilhamento de vídeos e mostram a existência de diferentes perfis de usuários.

Dentre as várias contribuições desses trabalhos, destacamos a criação de valiosos modelos capazes de descrever a carga que chega nesses servidores, essenciais para a geração de carga sintética que, por sua vez, possibilita a realização de experimentação e simulação baseadas em distribuições realistas. Neste trabalho, apresentamos uma caracterização da carga de uma LBSN do ponto de vista do servidor.

No contexto das redes sociais, Benevenuto e colaboradores [10] utilizaram dados de cliques de usuários do Orkut de forma a caracterizar a navegação e as formas de interação dos usuários nesses sistemas. De forma semelhante, Schneider e colaboradores [36] apresentaram um estudo da navegação dos usuários no Facebook. Em um estudo mais recente, Benevenuto e colaboradores [11] mediram a distância física e topológica das interações entre os usuários do Orkut, mostrando que o conteúdo nesses sistemas é em sua maioria produzido e consumido localmente. Em [39] Erramillia e colaboradores fizeram uma caracterização com uma base de dados do Twitter, com isso criaram um *framework* para geração sintética das atividades de escrita do Twitter. Gill e colaboradores [25] caracterizaram sessões dos usuários do Youtube e compararam os resultados com as sessões tradicionais dos usuários da web. Foi identificado que os usuários do YouTube transferem mais dados e tem mais tempo de espera do que as cargas de trabalho Web tradicionais. Essas diferenças têm implicações para as redes e administradores de sistemas responsáveis pelo planejamento de capacidade.

Existem vários trabalhos que caracterizam diferentes aspectos da LBSN. Scellato e colaboradores [35] apresentaram um estudo de três LBSNs, Brightkite, Foursquare e Gowalla. Eles observaram forte heterogeneidade entre os usuários com diferentes escalas geográficas de interação através de laços sociais, com a probabilidade de laço social entre dois utilizadores, em função da distância geográfica entre eles. Em [2] Noulas e colaboradores analisaram a dinâmica dos check-ins, demonstrando os padrões espaço-temporais e a mobilidade dos usuários nos espaços urbanos. Em [40], os autores apresentaram uma caracterização de como os usuários interagem entre si utilizando *tips* e *dones*, através da coleta de seus perfis do Foursquare. *Tips* são dicas sobre um determinado local e podem ser marcadas como *dones* se um usuário concorda com seu conteúdo. Noulas e colaboradores [32] utilizaram um algoritmo de agrupamento (*clustering*) espectral para agrupar os usuários baseado nos padrões de check-ins. Baseados nos atributos das regiões e usuários de duas cidades metropolitanas, puderam identificar grupos de usuários que visitam categorias similares de lugares e caracterizar o tipo de atividade que acontece em cada região da cidade. Cho e colaboradores [18] estudaram o Gowalla, Brightkite e dados de telefone celular, relatando que viagens de longa distância são mais influenciadas pela amizade social, enquanto movimentos com distâncias curtas não são influenciado pelas redes sociais.

Diferentemente de todos esses esforços, esse trabalho visa caracterizar e entender como as requisições chegam a um servidor, um tipo de sistema que ainda não foi investigado sob essa perspectiva.

Capítulo 3

Conjunto de Dados

Este capítulo apresenta as diferentes bases de dados utilizadas ao longo deste trabalho. Grande parte das bases de dados descritas a seguir já foram utilizadas em trabalhos anteriores [27] [37] . Sendo assim, apenas as características das bases importantes para o trabalho serão discutidas.

3.1 Dados do Apontador

Em nosso estudo, analisamos a carga de trabalho do *site* Apontador[1]. O Apontador é uma rede social brasileira baseada em localização que possui uma base georeferenciada com aproximadamente sete milhões de locais. Cada local possui uma página no *site* onde são apresentadas informações, tais como: o nome, endereço, latitude, longitude, categoria e telefone do local. Os usuários que acessam estas informações podem fazer isto de forma anônima ou registrada (logados). Além de procurar e visualizar as informações desses locais, os usuários também podem recomendar, avaliar, inserir fotos e cadastrar novos locais. No entanto, para que um usuário possa cadastrar um novo local, avaliar um existente ou associar uma foto ao local, é preciso estar logado no *site*. Os mesmos locais disponíveis no *site* também estão disponíveis nas aplicações para dispositivos móveis das plataformas iPhone, Android ou BlackBerry. Nessas aplicações, um usuário cadastrado pode fazer check-in num lugar, tirar uma foto e associá-la ao lugar.

Os registros (logs) utilizados correspondem ao período de um mês, de 01/10/2011

[1]http://www.apontador.com.br

Descrição	**Distintos**	**Requisições**
Usuários Logados	38.053	603.696
Usuários Não Logados	51.876.168	63.705.556
Usuários Totais	51.914.221	64.309.252
Locais acessados	**2.679.533**	**27.499.263**

Tabela 3.1: Características da Base de Dados do Apontador

a 31/10/2011, a tabela 3.1 mostra que foram contabilizados um total de **64.309.252** requisições, vindas de 51.914.221 usuários diferentes. Cada registro da carga de trabalho representa uma requisição enviada por um usuário ao Apontador. As seguintes informações estão disponíveis para cada requisição: *timestamp*, *usuário*, *objeto*, *tipo* e *local*. O campo timestamp é o momento em que a requisição foi recebida pelo servidor. O campo usuário corresponde a um identificador do *cookie* do navegador do usuário que gerou a requisição. O objeto é o código único para identificar a requisição. O campo tipo são as ações que uma pessoa pode realizar em um local. O campo local é o local solicitado na requisição pelo usuário.

Como pode ser visto na tabela 3.2, são várias as ações que uma pessoa pode realizar em um local e que são monitoradas pelo sistema de log. Estas ações são: acessar a página de um local (*visit*); clicar no telefone do local (*phone*)[2]; clicar no botão "recomendo" do local (*thumbs up*); clicar no botão "não recomendo" do local (*thumbs down*); clicar no botão ir para o site do local (*site*); fazer o upload de uma foto relacionada com o local (*send photo*); clicar no link que compartilha o local no Facebook (*facebook*); clicar no link que compartilha o local no Orkut (*orkut*); clicar no link que compartilha o local no Twitter (*twitter*); clicar no e-mail do local (*email*) e; quando a pessoa solicita o widget com o mapa do local (*widget*). Além das ações descritas acima, existem outras ações que são monitoradas quando o local é patrocinado. Estas ações são: momento em que a pessoa solicita a impressão de um cupom promocional (*focus coupon*); quando a pessoa visualiza o telefone do local (*focus phone*), e; quando a pessoa visualiza o e-mail do local (*focus email*).

[2]Propositadamente o número do telefone do local é parcialmente ocultado. Para que a pessoa possa visualizar o número completo do telefone ela precisa clicar no número.

Grupo	**# Requisições**	Porcentagem
Visit	53.623.387	83,3800
Phone	9.225.458	14,3400
Site	1.160.655	1,8000
Thumbs up	242.937	0,3700
Thumbs down	49.604	0,0770
Send photo	3.941	0,0060
Focus email	669	0,0010
Facebook	655	0,0010
Email	630	0,0009
Focus phone	547	0,0008
Orkut	343	0,0005
Wigdet	235	0,0003
Focus copoun	125	0,0001
Twitter	66	0,0001

Tabela 3.2: Tipos de Ações

3.2 Coleta de Locais

Os dados com cliques dos usuários obtidos junto ao Apontador contêm apenas o identificador dos locais armazenados no sistema. Sendo assim, informações como endereço, geo-localização e categoria do local não estão disponíveis nos logs dos servidores do Apontador. Entretanto, a partir do identificador do local é possível coletar tais informações através da API do Apontador[3].

Para realizar tal coleta desenvolvemos um coletor em Python que recuperou as informações de todos os locais disponíveis em nossa base de cliques dos usuários.

A Tabela 3.3 apresenta as características da base de dados coletada. No total, foi possível recuperar informações de 99,8% dos locais distintos acessados. Cada local no formato XML (*eXtensible Markup Language*) possui as seguintes informações: identificação única, nome, descrição, contador de *clicks*, número de avaliações, número de recomendações, categoria do local, endereço, telefone, latitude, longitude, endereço do

[3]http://api.apontador.com.br/pt/

Descrição	**Distintos**	Porcentagem
Locais acessados	2.679.533	100
Locais coletados em XML com sucesso	2.672.353	99,8

Tabela 3.3: Características da Base de Dados Coletada

site do local e informações do usuário criador do local.

Através do campo endereço, conseguimos listar os Estados mais frequentes dos locais distintos acessados no período de um mês, conforme mostrado na Tabela 3.4. Observamos que três dos seis Estados mais frequentes pertencem à região sudeste do país e os outros três à região sul.

Estado	**Número de Locais Distintos**	**Porcentagem**
São Paulo	796.181	29,79
Minas Gerais	279.772	10,47
Rio de Janeiro	251.029	9,39
Rio Grande do Sul	224.546	8,40
Paraná	195.554	7,32
Santa Catarina	146.524	5,48
Bahia	121.633	4,55
Pernambuco	88.383	3,31
Ceará	76.121	2,85
Goias	74.561	2,79
Espirito Santo	53.533	2,00
Mato Grosso	41.134	1,54
Distrito Federal	40.255	1,51
Mato Grosso do Sul	39.138	1,47
Pará	34.820	1,30
Rio Grande do Norte	32.976	1,23
Paraíba	29.901	1,12
Outros	146.292	5,48

Tabela 3.4: Estados com Maior Número de Locais Acessados

O campo categoria identifica qual é o tipo de estabelecimento ou serviço oferecido pelo local. A Tabela 3.5 mostra as categorias mais frequentes dos locais únicos acessados no período de um mês.

Categoria	**Número de Locais Distintos**	**Porcentagem**
Endereços Empresariais	254.468	9,52
Automóveis e Veículos	82.677	3,09
Confecções e Vestuário	77.130	2,89
Construção	67.927	2,54
Beleza	54.168	2,03
Móveis e Decoração	53.703	2,01
Medicina e Saúde	52.579	1,97
Bancos e Instituições Financeiras	44.900	1,68
Alimentos	44.251	1,66
Associações e Sindicatos	43.663	1,63
Postos de Combustível	43.483	1,63
Restaurantes	41.931	1,57

Tabela 3.5: Categorias Mais Frequentes

A Tabela 3.6 mostra as categorias dos 10 locais com os maiores de números de sessões no período de um mês.

3.3 Outros Sistemas

Estamos listando outros sistemas para que possamos fazer uma comparação com a popularidade dos objetos do apontador.

3.3.1 Servidor Web da Copa do Mundo de 1998

Idealmente, gostaríamos de comparar dados obtidos de redes sociais atuais com dados da Web 1.0, constituída em sua maioria por servidores contendo páginas estatícas onde usuários da Web eram meros expectadores. Um conjunto de dados que atende tais re-

Categoria	**#Sessões**
Serviços Gerais	5.660
Laboratórios	5.283
Consulados e Embaixadas	4.684
Alimentos	3.782
Correios	3.688
Confecções e Vestuário	3.427
Transporte	3.403
Escolas Públicas	3.146
Transporte	3.009
Transporte	2.979

Tabela 3.6: Categorias dos 10 Locais com mais Sessões

quisitos e se encontra publicamente disponível, consiste de dados anonimizados públicos do servidor da Web da Copa do Mundo de 1998 [4], que teve em média 11.000 visitas por minuto e 40MB de dados transferidos por minuto aos usuários . Em particular, nós utilizamos 32 dias do log (de 24 de maio a 24 de junho de 1998), contendo 69.747 objetos únicos e 681.469.425 requisições registradas para esses objetos.

A Tabela 3.7 mostra que em quase todos os pedidos dos usuários (98%) eram para HTML ou para arquivo de imagem. Essa é uma característica típica observada em cargas de trabalho de servidores Web.

Tipo	**% de requisições**
Imagens	88,16
HTML	9,85
Java	0,82
Compactados	0,08
Audio	0,02
Video	0,00
Dinâmicos	0,02
Outros	1,05

Tabela 3.7: Distribuição por Tipo de Arquivo - Copa do Mundo 1998

3.3.2 Orkut

Foram utilizados dados do Orkut coletados e caracterizados em um trabalho anterior [10]. Esses dados foram coletados de um agregador de redes sociais e possui o registro de todos os objetos acessados de diferentes redes sociais por 36.309 usuários que utilizaram o sistema durante o período monitorado. Para realizarmos nossas análises, vamos utilizar apenas os acessos a fotos do Orkut de modo a medir a popularidade de fotos compartilhadas nesse sistema. No total essa base de dados contém 23.764 fotos em nossos logs, acessadas 121.939 vezes.

3.3.3 YouTube

Dentre os sistemas sociais atuais, um dos maiores tráfegos está associado à distribuição de vídeos. Com o intuito de comparar a popularidade de vídeos à popularidade de outros objetos da Web 2.0 e a objetos da Web 1.0, vamos utilizar uma base de dados do YouTube contendo 1.666.226 vídeos coletada em dezembro de 2006 [12]. Para cada vídeo, essa base contém o número de visualizações dos vídeos, sendo que no total os vídeos dessa base receberam 369.762.000.000.000 acessos.

3.3.4 Uol Mais

Nossa base de vídeos de YouTube contém apenas números relativos à popularidade dos vídeos. Entretanto, sistemas de compartilhamento de vídeos recebem outras requisições relativas às imagens que representam os vídeos ou mesmo requisições de busca e navegação pelos sistemas. Os tipos de requisições são apresentados na tabela 3.8. Para estudar a popularidade de todos os objetos acessados e não só dos vídeos, vamos utilizar também uma base de dados do UOL Mais, um sistema de compartilhamento de vídeos do UOL. Uma descrição detalhada dos dados dessa base pode ser obtido em [9]. O log utilizado nesse trabalho foi obtido no período de 12 de dezembro de 2007 a 07 de janeiro de 2008, possui 109.239 objetos e 3.613.935 requisições de acessos a esses objetos.

Grupo	Tipo de Requisição	#Requisições	Porcentagem
Visualização	Visualizações de vídeos	2.758.883	74,94 %
Usuário	Listagem de vídeos de certo usuário	218.335	5,93%
	Listagem de vídeos de certo usuário com certa tag	75.583	2,05%
Listas	Listagem de "top"vídeos	55.307	1,50%
	Listagem de relacionados de um vídeo	32.838	0,89 %
Interações	Avaliações de vídeos	22.038	0,60%
	Postagem de comentário para vídeo	14.131	0,38%
	Adição de vídeo como favorito	10.774	0,29%
Busca	Busca	1.625	0,04%
	Listagem de vídeos com certa tag	421.700	11,46%
Outros	Página principal	2.679	0,07%
	Requisições de erro ou mal formatadas	67.339	1,82%

Tabela 3.8: Tipos de Requisições do Uol Mais

Capítulo 4

Caracterização da Carga de Trabalho

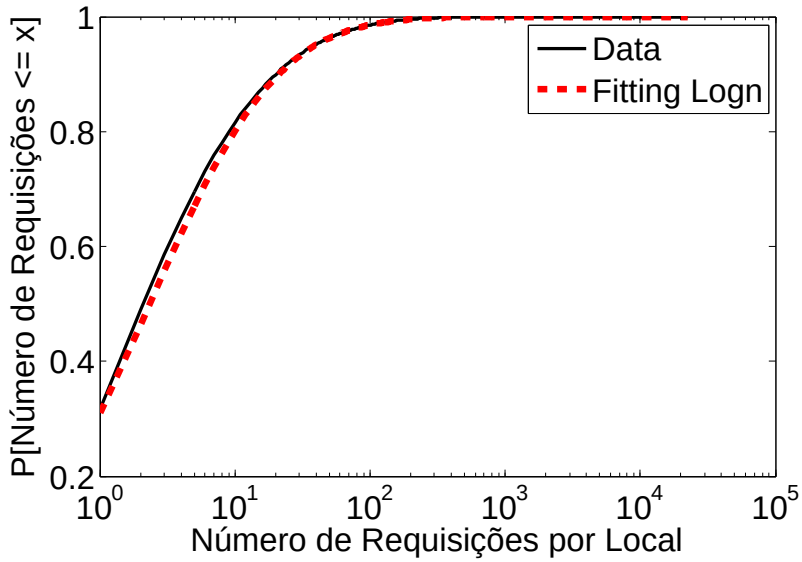
Neste capítulo, apresentamos uma caracterização da carga de trabalho do Apontador sob diferentes perspectivas, mostrando vários aspectos e distribuições. Para verificar a acurácia dos modelos propostos, medimos o fator R^2 da regressão linear [38] para cada distribuição analisada. Em todos os modelos apresentados no trabalho, os valores de R^2 estão acima de 0,96. Sendo que quando o valor de R^2 é igual a 1 significa que não há diferenças entre o modelo e a carga de trabalho real.

4.1 Popularidade dos Locais

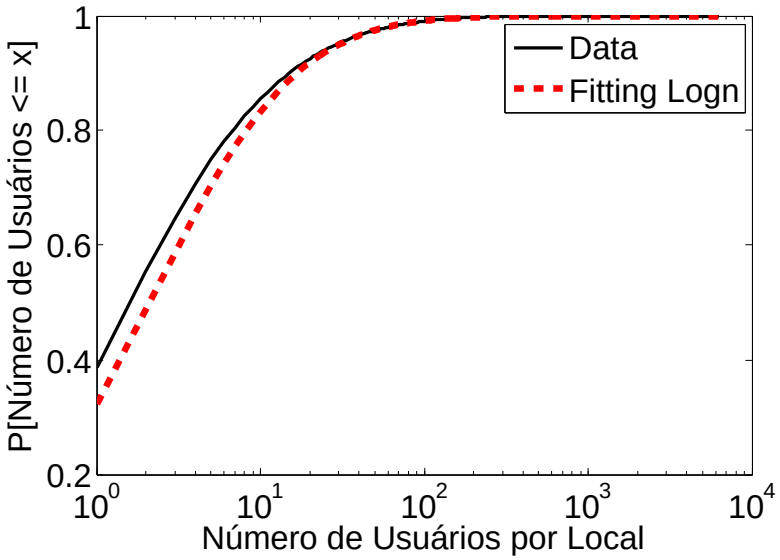
Primeiramente avaliamos a popularidade dos locais, com o objetivo de verificar se a mesma segue uma distribuição conhecida.

A Figura 4.1(a) mostra a distribuição de probabilidade acumulada (CDF) do número de requisições por locais. Podemos notar que existe uma pequena quantidade de locais com muitos acessos e uma grande quantidade locais com poucos. Por exemplo mais de 80% dos locais possuem até 10 requisições. Tal observação é importante pois mostra o grande potencial para *caching* de locais que o sistema possui. De fato, essa distribuição é bem modelada com uma distribuição Lognormal, com $\mu = 0,849$, $\sigma = 1,720$ e $R^2 = 0,996$.

Assim como a distribuição de requisições por local, a Figura 4.1(b) mostra uma distribuição de probabilidade acumulada (CDF) que segue uma distribuição Lognormal, sendo a 4.1(b) o número de usuários por local (quantidade de usuários distintos que



(a) CDF - Popularidade dos Locais



(b) CDF - Número de Usuário por Local

Figura 4.1: Numero de Requisições e Usuários por Local

acessaram cada local) com $\mu = 0,741$, $\sigma = 1,617$ e $R^2 = 0,979$.

Na Web, a idéia de haver uma grande concentração de popularidade em poucos objetos é a base para a construção de sistemas hierárquicos de cache e foi amplamente aplicado no projeto de sistemas de caches em um passado bastante recente [6, 8, 23, 42]. Nossa hipótese com base em [27] é que a popularização das redes sociais possa contribuir

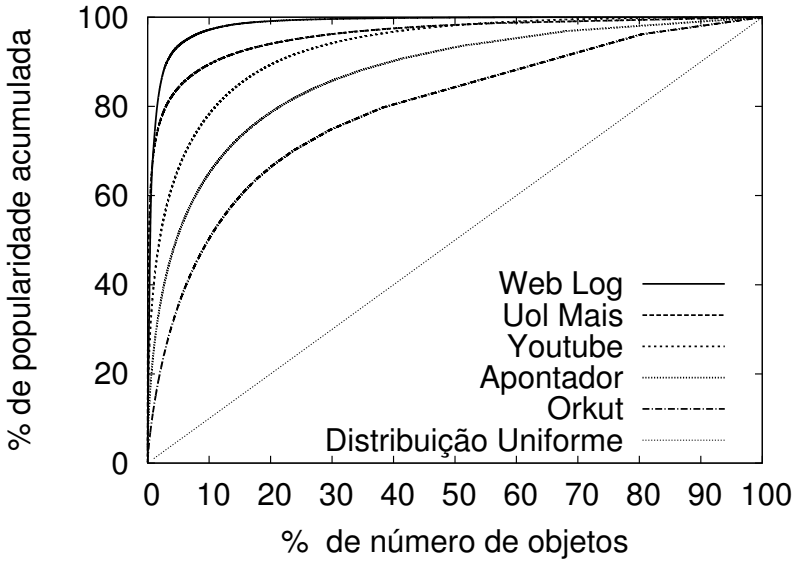


Figura 4.2: Gráfico Normalizado de Popularidade

para uma menor concentração de popularidade em poucos objetos.

A seguir vamos analisar as características da popularidade de conteúdo em diferentes sistemas como uma tentativa de quantificar como padrões de interações de redes sociais afetam a popularidade de conteúdo nesses sistemas. A Figura 4.2 mostra essas distribuições normalizadas para diferentes sistemas discutidos no Capítulo 3. O eixo x representa o ranking do conteúdo em porcentagem, onde o ranking 10% representa os primeiros 10% dos objetos de cada base de dados analisada. O eixo y representa a porcentagem de popularidade acumulada, ou seja, para os 10% primeiros objetos do ranking, o eixo y mostra qual a fração dos acessos que esses 10% receberam. Podemos notar a grande diferença de concentração de popularidade que cada curva apresenta e que as curvas sociais são bem mais distribuídas em comparação com a concentração de popularidade dos objetos dos dados do servidor Web da Copa do Mundo de 98. Como exemplo, enquanto 10% dos objetos mais populares do servidor Web da Copa do Mundo concentram 97,18% dos acessos, 10% dos objetos do Orkut receberam apenas 50,33% dos acessos.

Nas demais redes podemos ver que a concentração de popularidade também é sempre menor se comparada ao servidor da Copa do Mundo. O Uol-Mais, por ser um servidor de vídeos que também recebe requisições relativas às imagens (*thumbnails*) que representam os vídeos ou mesmo requisições de busca e navegação pelo sistema, é o que possui a curva mais próxima do servidor Web da Copa do Mundo de 98. Nos dados do YouTube,

que contabilizam apenas a popularidade de acesso à vídeos, podemos notar um maior espalhamento dos acessos aos objetos. No Apontador, os objetos analisados são localizações e a curva representa a popularidade de acesso a diferentes localizações. Podemos notar que a concentração de popularidade é ainda menor, o que reflete o interesse local por diferentes objetos nesse tipo de sistema. No Orkut, as fotos e suas popularidades são analisadas. A concentração de popularidade se mostrou a menor, visto que usuários do Orkut normalmente acessam apenas fotos de seus amigos, o que dificulta a formação de objetos muito populares no sistema.

Com essa análise podemos concluir que o sistema Apontador em nível de popularidade de objetos fica entre os sistemas de compartilhamento de vídeos e o orkut. Possivelmente isso ocorre devido à localidade espacial dos usuários que acessam os objetos do apontador.

Para examinar mais a fundo as diferenças de popularidade, vamos medir a disparidade entre essas medidas. A medida de disparidade é bastante conhecida na economia para medir diferenças entre ricos e pobres em um país. Tipicamente, o 95º e o 5º percentis são comparados. A Tabela 4.1 mostra as medidas de disparidade para as diferentes distribuições. A disparidade entre o 95º e o 5º percentis é 20 para o Orkut e 45.831 para o servidor da copa do mundo de 98. Mesmo quando comparamos a disparidade das outras distribuições com a distribuição da Web, podemos notar que a disparidade na Web é ordens de grandeza maior do que a de sistemas sociais.

Ratio	*Web Copa98*	*UOL Mais*	*YouTube*	*Apontador*	*Orkut*
1º / 99º	703.959	334	15.410,5	128	46
5º / 95º	45.831	52	979,62	39	20
10º / 90º	15.119	24	214,61	21	12

Tabela 4.1: Disparidade de Popularidade

Nossas observações de que distribuições de acessos a objetos em sistemas sociais são bem menos concentradas do que em dados de um servidor típico da Web 1.0 levantam importantes questionamentos sobre a efetividade da infraestrutura tradicional para distribuição de conteúdo atualmente e, principalmente no futuro, caso as expectativas de crescimento e ainda maior popularização de sistemas sociais se confirme. Isso porque a atual infraestrutura é baseada em caching de uma pequena fração de objetos que dominam o conteúdo. A falta de objetos extremamente populares em sequencias

de requisições na Web sugere que pode ser necessário reexaminar a infraestrutura para distribuição de conteúdo social no futuro. De fato, não é de se estranhar que trabalhos recentes já mostraram que o conteúdo do Facebook poderia ser processado 79% mais rápido e consumir 91% a menos de largura de banda com a implantação de servidores e caches regionais. [43].

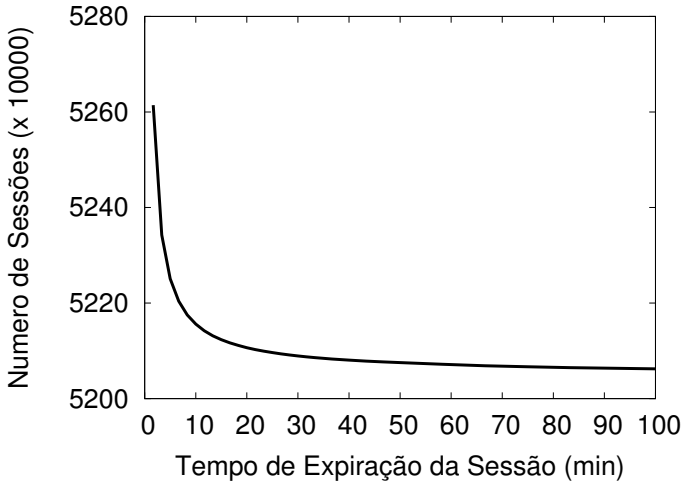
4.2 Definição de Sessões

Uma sessão de um usuário é definida como um série de requisições realizadas pelo usuário a um *site* durante um determinado período de tempo [3, 30]. Em ambientes das LBSN, uma sessão de usuário pode incluir acesso ao local, acesso ao *site*, acesso ao telefone e as ações citadas na Capítulo 3. Tais tipos de requisições diferem bastante das sessões de usuários de *sites* convencionais, os quais não dispõem do mesmo grau de interação dos usuários de sistemas da *Web 2.0*.

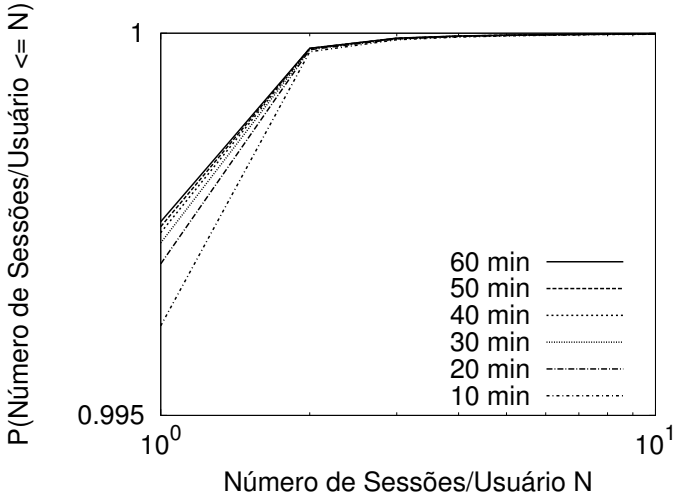
A determinação do início e término de uma sessão em aplicações LBSN requer uma análise específica dos tempos entre requisições a fim de medir a inatividade do usuário, uma vez que a maioria das sessões não apresenta um registro explícito de operações de *login* e *logout*. Portanto, é necessário realizar uma análise para identificar um valor limite de tempo entre requisições para que sejam consideradas como sendo de uma mesma sessão. Assim, duas requisições consecutivas são consideradas da mesma sessão se o tempo entre elas é menor do que esse limite, denominado tempo de expiração da sessão.

É importante escolher um tempo de expiração adequado para não gerarmos sessões que não representam o uso do serviço pelos usuários, evitando unir diferentes momentos de uso do serviço ou fragmentar uma navegação realizada pelo usuário. Seguindo a metodologia proposta em [30], realizamos uma avaliação do tempo de expiração da sessão mais adequado para nossa aplicação.

A Figura 4.3(a) apresenta o número total de sessões para diferentes valores de tempo de expiração. Um valor extremamente pequeno (ex., 1 minuto) resulta em um volume de sessões extremamente alto (mais de 52 milhões de sessões), gerando praticamente somente sessões com uma requisição. À medida que o valor do tempo de expiração aumenta, o número de sessões reduz continuamente, até que essa diminuição se torna mais estável. Essa estabilidade ocorre por volta dos 30 minutos, indicando que esse valor



(a) Tempo de Expiração das Sessões



(b) CDF - Número de Sessões por Usuário

Figura 4.3: Definição de Sessões

é um limite adequado para ser adotado como tempo de expiração da sessão.

A fim de testar esse valor geramos a distribuição de probabilidade acumulada (CDF) do número de sessões por usuário para vários valores de tempo de expiração de sessão, conforme ilustra a Figura 4.3(b). A diferença entre as distribuições para os diferentes valores de tempo de expiração é maior para os valores menores, tornando-se mais consistente a partir de 30 minutos. Sendo assim, adotamos 30 minutos como tempo de expiração das sessões para nossas análises, obtendo um total de 52.089.255 de sessões de

usuários em nossa carga de trabalho.

É interessante observar na Tabela 4.2 que esse resultado é similar às análises realizadas no trabalho [16], um pouco menor do que nos trabalhos [9, 26]. Quando comparado com os resultados que caracterizam sessões em *sites* Web tradicionais [3, 33], o valor de tempo de expiração da sessão aqui obtido é 3 vezes maior do que os 10 minutos tipicamente observados. Isso ocorre devido ao tempo maior que o usuário gasta para visualizar os locais com seus detalhes e serviços relacionados, que podem levar o usuário a ficar mais tempo em sua navegação pelo sistema.

Sistema/Descrição	**Tempo de Expiração(min)**	**Ano Coleta**
Servidor do Site da Copa do Mundo [3]	10	1998
Servidor de Compras via Web [28]	15	2001
LBSN Apontador	**30**	**2011**
Servidor de Weblog [16]	30	2006
Youtube, Compartilhamento de Vídeos [26]	40	2007
Uol Mais, Compartilhamento de Vídeos [9]	40	2008
Servidor de Vídeos sob Demanda [19]	60	2002
Twitter, Escrita de Tweets [13]	167	2009

Tabela 4.2: Tempo de Expiração da Sessão (min)

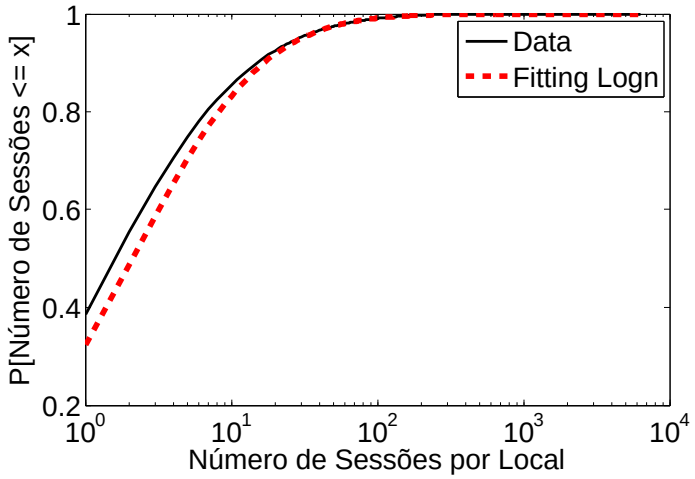


Figura 4.4: CDF - Número de Sessões por Local

A Figura 4.4 apresenta uma distribuição Lognormal para o número de sessões por local com $\mu = -4,524$, $\sigma = 3,018$ e $R^2 = 0,979$.

4.3 Nível de Atividade dos Usuários

A seguir analisamos o nível de atividade dos usuários. Sabemos que usuários podem acessar o serviço de busca local repetidas vezes dentro da mesma sessão ou retornar ao sistema constantemente, gerando um grande número de sessões. Sendo assim, para modelarmos o nível de atividade dos usuários, caracterizamos o *ranking* dos usuários em termos do número de requisições enviadas e em termos do número de sessões criadas no sistema. Chamamos de usuário cada endereço *IP* anonimizado da carga de trabalho.

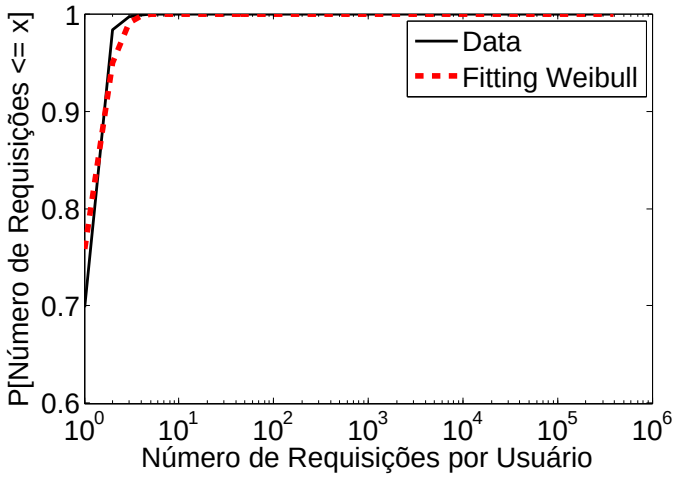
A Figura 4.5(a) mostra a distribuição de probabilidade acumulada (CDF) do número de requisições enviadas ao servidor por usuário. Podemos notar que existe uma pequena quantidade de usuários que fazem muitas requisições ao servidor e uma grande quantidade de usuários que fazem poucas requisições. Ou seja 69% dos usuários possuem 1 requisição e mais de 99% dos usuários possuem até 5 requisições. Com isso foi utilizada uma distribuição Weibull para obtermos uma modelagem que represente bem os dados. Sendo $\alpha = 0,345$, $\beta = 2,683$ e $R^2 = 0,967$.

Em termos das sessões criadas no servidor visto na figura 4.5(b), foi utilizada uma função que segue a Lei de Potência para modelar a distribuição do *ranking* de sessões com $\alpha = 0,0007$ e $R^2 = 0,984$. Esse resultado enfatiza o comportamento de que poucos usuários possuem muitas sessões, enquanto muitos possuem poucas sessões.

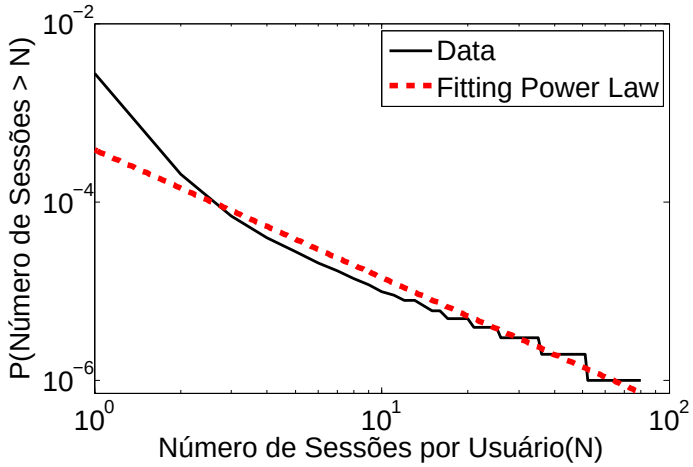
Em comparação a outros trabalhos já realizados [9, 19], no ranking de requisições por usuários, as distribuições seguem Zipf. No ranking de sessões por usuários, eles seguem um distribuição Zipf e Exponencial, respectivamente.

4.4 Padrões Temporais do Acesso

Nesta seção analisamos o número de requisições que chegam ao servidor ao longo do tempo. A Figura 4.6(a) mostra o número de requisições que chega ao servidor em intervalos de uma hora. A curva apresenta um padrão periódico, com maior intensidade de acessos durante o dia e menor intensidade durante a noite. Podemos notar que durante os finais de semanas e nos feriados, como por exemplo, o feriado de 12 de outubro ocorrem quedas de acesso ao sistema. Como pode ser analisado, os picos que normalmente passam de 250.000 requisições em 1 hora, em dias de semana, nos finais de semana e feriados ficam em torno de 100.000 requisições em 1 hora, uma queda de mais



(a) CDF - Número de Requisições por Usuário



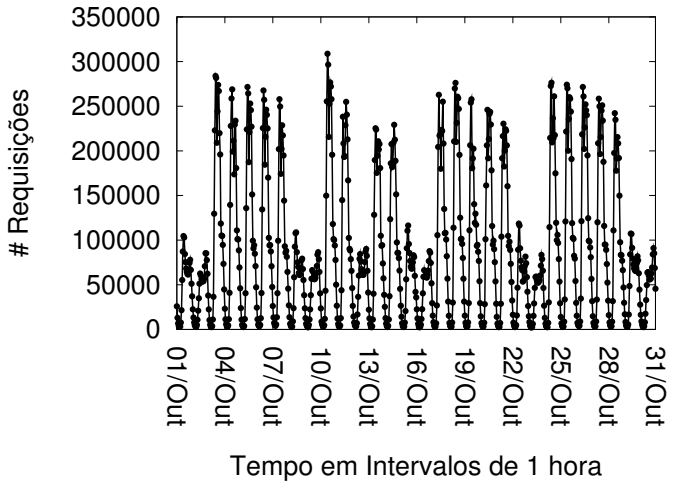
(b) CCDF - Número de Sessões por Usuário

Figura 4.5: Nível de Atividade dos Usuários

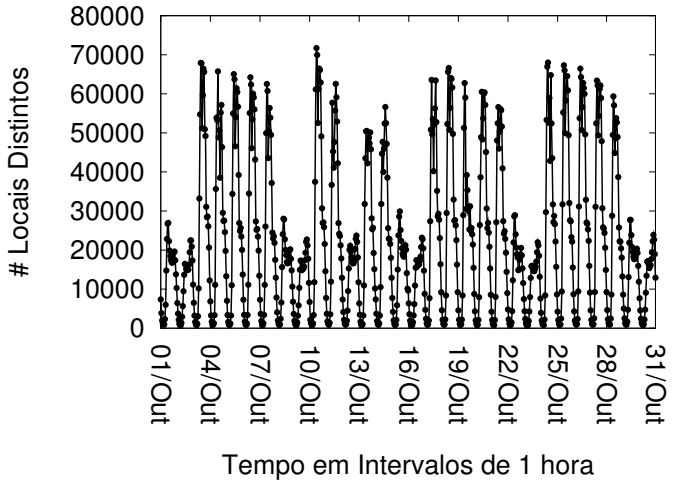
de 60%. Os locais únicos acessados nessas requisições seguem o mesmo padrão como pode ser visto na Figura 4.6(b).

Esses padrões são similares aos descritos em estudos sobre servidores tradicionais da *Web* [5, 41] e também a outros tipos de servidores como o de weblogs [16], compartilhamento de vídeos [9], comércio eletrônico [28] e vídeos sob demanda [19]. Existe uma diferença nos padrões apenas nas datas especiais, quando pode ocorrer um grande aumento de requisições, como por exemplo [4], ocorreu um aumento de demanda em jogos chaves da Copa do Mundo de 1998, assim como eventos especiais que podem afetar os

sites de comércio eletrônico como as campanhas publicitárias, promoções especiais, ou a aproximação de feriados como o Dia dos Namorados, Páscoa, Dia das Mães, Dia dos Pais e Natal.



(a) Requisições em Intervalos de 1h

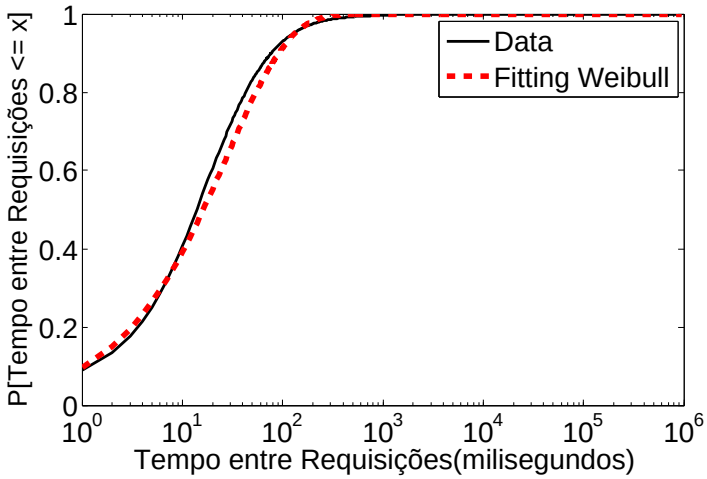


(b) Locais em Intervalos de 1h

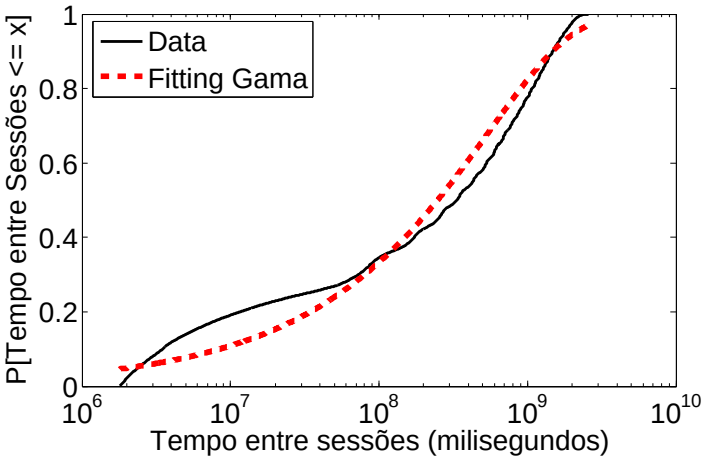
Figura 4.6: Número de Requisições e Locais em Intervalos de 1h

Para analisarmos a participação dos usuários do sistema, caracterizamos o intervalo de tempo entre chegadas de requisições e sessões ao sistema. Apresentamos nas Figuras 4.7(a) e 4.7(b) a probabilidade acumulada (CDF) para os intervalos de tempo entre requisições e sessões, respectivamente. Podemos notar que a probabilidade do intervalo

de tempo entre requisições ser maior do que 500 milisegundos é menor do que 3%, sendo que 78% das requisições chegam ao servidor com intervalos menores do que 100 milisegundos. Da mesma forma, cerca de 99% dos intervalos entre requisições são menores do que 1 segundo. E analisando o intervalo entre sessões notamos que a probabilidade de ser menor que 1h é de 20%.



(a) CDF - Intervalos de Tempo entre Requisições



(b) CDF - Intervalos de Tempo entre Sessões

Figura 4.7: Padrões Temporais do Acesso

As distribuição do intervalo entre requisições é melhor aproximada por uma distribuição Weibull onde $\alpha = 0,049$, $\beta = 0,710$ e $R^2 = 0,983$. Para a distribuição do intervalo de tempo entre sessões foi utilizada uma distribuição Gama com $\alpha = 0,360$,

$\beta = 1023168222$ e $R^2 = 0,961$.

Comparando com outros trabalhos, a Tabela 4.3 mostra que a distribuição do intervalo entre requisições é similar a [7, 16, 28], onde todos seguem uma distribuição Weibull. Diferentemente a Tabela 4.4 mostra que quando comparamos a distribuiçao do intervalo entre sessões temos [16] com uma distribuição Weibull, [9] com uma distribuição Exponencial e [10] que segue uma distribuição Lognormal.

Sistema	**Distribuições**
LBSN Apontador	**Weibull**
Servidor de Compras via Web [28]	Weibull
Servidor de Weblogs [16]	Weibull
Servidor Web [7]	Weibull

Tabela 4.3: Distribuições dos Intervalos entre Requisições

Sistema	**Distribuições**
LBSN Apontador	**Gama**
Uol Mais, Compartilhamento de Vídeos [9]	Exponencial
Servidor de Weblogs [16]	Weibull
Orkut [10]	Lognormal

Tabela 4.4: Distribuições dos Intervalos entre Sessões

4.5 Modelo de Comportamento do Usuário

Esta seção descreve o modelo de comportamento do usuário, representando as atividades de um visitante a uma LBSN. Como primeiro passo, o comportamento típico dos visitantes de um LBSN pode ser descrito, do seguinte modo: um usuário inicia uma nova sessão solicitando um acesso a uma página do local. Em seguida, o usuário pode manter-se dentro do local, visitando um ou mais *links* nesse local, bem como acessar o telefone do local, recomendar esse local, ou o usuário pode acessar um novo local. Em algum momento o usuário pode terminar a sessão saindo do site.

De forma a modelar o comportamento de um visitante da LBSN, e descrever padrões de solicitação para os vários locais visitados dentro de uma sessão, propomos usar um Grafo do Modelo de Comportamento do Usuário (UBMG), que é um grafo de transição de estados. Neste grafo, nós representamos os estados possíveis. A probabilidade é atribuída a cada transição entre dois estados. É possível definir diferentes tipos de usuários usando UBMGs, que são diferenciados pela probabilidade na transição de estado. Determinamos os seguintes estados de um visitante da busca local durante uma sessão:

Novo Local: O usuário acessa esse estado quando ele faz seu primeiro acesso a um determinado local ou quando faz o acesso a um local estando anteriormente em outro local ou de detalhes de outro local.

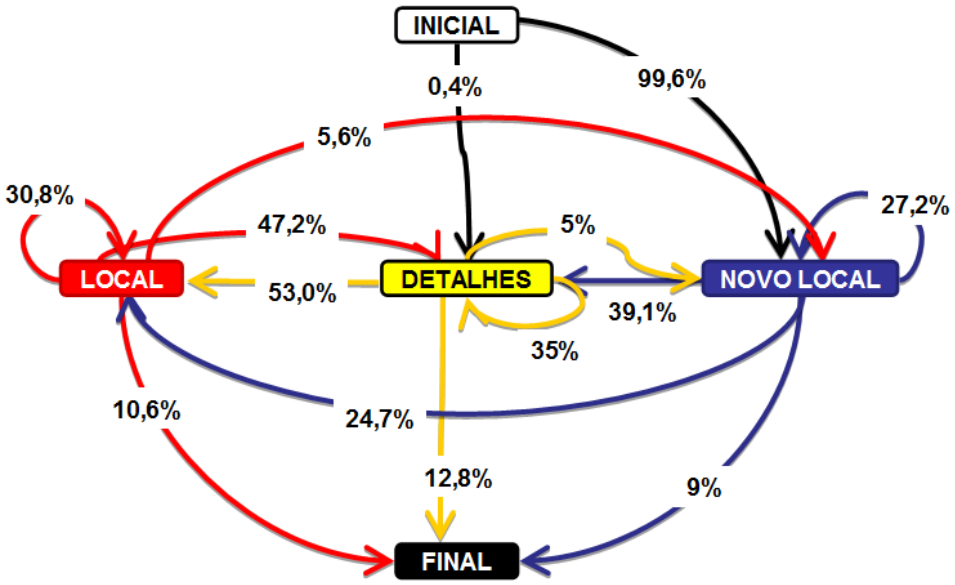
Local: O usuário acessa esse estado quando ele volta ao mesmo local que ele estava visitando e sai desse estado quando visita details ou um local diferente.

Detalhes: Quando o usuário acessa os detalhes de um local.

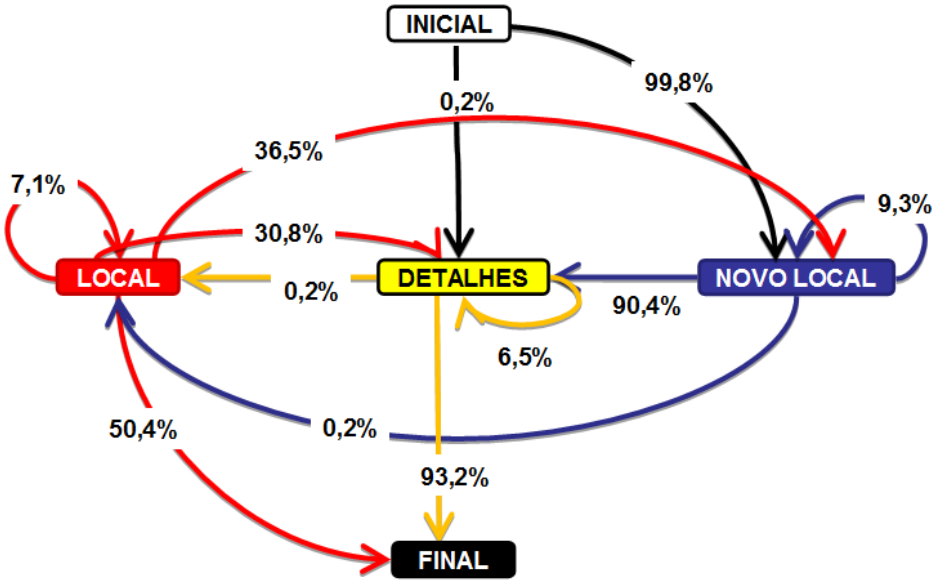
Final: A sessão termina quando o tempo desde o último acesso excede um valor de tempo limite, o qual é assumido como sendo de 30 minutos.

Os visitantes do LBSN foram classificados em dois perfis, de acordo com seus padrões de acessos. Um perfil é dos usuários logados, como pode ser visto na Figura 4.8(a) e o outro perfil é o dos usuários não logados, que pode ser visto na Figura 4.8(b). Podemos observar que usuários logados realizam muito mais atividades dentro de um mesmo local (ex. 30,8% de chance de visitar o mesmo local e 35% de chance de acessar mais detalhes do local) em comparação com usuários não logados (apenas 7% de chance de visitar o mesmo local e 6,5% de acessar o mais detalhes do local alguma). Além disso, usuários logados navegam entre locais no sistema, o que praticamente não acontece com os usuários não logados.

Esse tipo de análise de mudança de estado, depende muito da interface e ferramentas que o sistemas possuem. Por exemplo [30] apresentou um grafo do modelo de comportamento do cliente de um comércio eletrônico, que tinha como nós: site, navegador, busca, selecionar, adicionar e pagar, que são as principais funções de um comércio eletrônico. Podemos citar também [17] que apresentou um grafo do modelo de comportamento de visitantes de um Blog, que tinha como opções de navegação: iniciar leitura em novo blog, continuar lendo o mesmo blog e fazer comentários. Em uma caracterização de compartilhamento de vídeo podemos citar [9] com um grafo do modelo de comportamento do usuário com os seguintes estados de transição: visualização, usuário, lista, avaliação e



(a) Perfil dos logados



(b) Perfil dos não logados

Figura 4.8: Perfis dos Usuários - UBMGs

busca. Na caracterização do orkut feita em [10] foi usado os seguinte estados de transição: recados, depoimentos, buscas, mensagens, fotos, profiles e amigos, comunidades e vídeos.

Capítulo 5

Conclusão e Trabalhos Futuros

Desde o lançamento das primeiras redes sociais online, esses sistemas têm ganhado popularidade continuamente. Seguir atualizações de amigos é hoje uma das mais populares atividades da Internet. Este novo paradigma de acesso a dados na Web está mudando a forma como conteúdo é consumido na Web. Utilizando dados de diferentes redes sociais, neste trabalho, nós investigamos propriedades de acesso aos objetos desses sistemas e discutimos implicações futuras para a Internet.

Nossos resultados mostram que objetos de redes sociais possuem suas popularidades de acesso bem mais distribuídas quando comparados a objetos da Web tradicional. Nossas observações indicam que novas estruturas de caches desenhadas para lidar especificamente com dados de redes sociais online podem ser mais adequadas para a Internet do Futuro.

Além disso, neste trabalho utilizamos uma carga de trabalho real e representativa para caracterizar os padrões de acesso ao servidor de uma LBSN, de forma a caracterizar e modelar os padrões de acessos dos usuários a esses sistemas. Como resultados, fornecemos modelos estatísticos para várias características de acesso, como popularidade dos locais e dos usuários, tempo entre chegada de requisições e sessões, etc. O estudo apresentado é inovador por ser o primeiro a analisar uma rede social baseada em localização sob o ponto de vista do servidor. Os modelos apresentados são úteis não só para a geração de carga sintética, mas também para o projeto e criação de novas infra-estruturas para esse tipo de serviço.

Quanto aos modelos apresentados, no ranking de requisições por usuário nosso trabalho segue uma distribuição Lognormal, diferentemente de outros trabalhos estudados

que seguem Zipf, comparando o ranking de usuários por sessão enquanto nosso trabalho segue uma Lei de Potência, apresentamos estudos que também seguem a Lei de Potência e outro que segue uma Exponencial. Nos padrões temporais do acesso vários trabalhos mostraram as mesma características que o nosso, com picos de requisições diurnas e no intervalo entre requisições seguindo a distribuição Weibull. Também foi apresentado um Grafo do Modelo de Comportamento do Usuário de uma LBSN.

Como trabalhos futuros, planejamos construir um gerador de carga sintética que possibilite realizar experimentação e simulação baseadas em distribuições realistas. O que possiblita melhor gerência de recursos computacionais e de rede, seja através de políticas de controle de qualidade de serviço (QoS) ou planejamento de capacidade, além de permitir a identificação de práticas comuns e oferecer serviços personalizados aos usuários, como forma de fidelização. Pretendemos também investigar formas de distribuir conteúdo publicado em redes sociais de maneira eficiente.

Referências Bibliográficas

[1] Nearly 1 in 5 smartphone owners access check-in services via their mobile device. `http://bit.ly/mgaCIG`.

[2] C. Mascolo A. Noulas, S. Scellato and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *International Conference on Weblogs and Social Media*, 2011.

[3] M. Arlitt. Characterizing web user sessions. *SIGMETRICS Performance Evaluation Review*, 28(2):50–63, 2000.

[4] M. Arlitt and T. Jin. Workload characterization of the 1998 world cup web site. In *Technical Report HPL-1999-35R1*, 1999.

[5] M. Arlitt and C. Williamson. Web server workload characterization: the search for invariants. *SIGMETRICS Performance Evaluation Review*, 24(1):126–137, 1996.

[6] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web client access patterns: Characteristics and caching implications. *World Wide Web*, 2:15–28, 1999.

[7] P. Barford and M. Crovella. Generating representative web workloads for network and server performance evaluation. In *ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, volume 26, pages 151–160, 1998.

[8] F. Benevenuto, F. Duarte, V. Almeida, and J. Almeida. Web Cache Replacement Policies: Properties, Limitations and Implications. In *Proc. of Latin American Web Congress*, November 2005.

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

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

[11] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user navigation and interactions in online social networks. *Information Sciences*, 195(15):1–24, 2012.

[12] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM Internet Measurement Conference*, 2007.

[13] G. Comarela, M. Crovella, and V. Almeida F. Benevenuto. Understanding factors that affect response rates in twitter. In *Proceedings of the 23rd ACM conference on Hypertext and social media(HT 12)*, pages 123–132, 2012.

[14] C. Costa, I. Cunha, A. Vieira, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto. Analyzing client interactivity in streaming media. In *World Wide Web Conference (WWW)*, pages 534–543, 2004.

[15] J. Rolia D. Krishnamurthy and S. Majumdar. A synthetic workload generation technique for stress testing session-based systems. In *IEEE Trabsactions on software engineering*, volume 32, pages 868–882, 2006.

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

[17] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, J. Almeida, and M. Curiel. Hierarchical characterization and generation of blogosphere workloads. In *Boston University Computer Science Department*, 2008.

[18] J. Leskovec E. Cho, S. Myers. Friendship and mobility: user movement in location-based social networks. In *ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1082–1090, 2011.

[19] W. Meira A. Bestavros E. Veloso, V. Almeida and S. Jin. A hierarchical characterization of a live streaming media workload. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment (IMW)*, pages 117–130, 2002.

[20] Needle in a Haystack: Efficient Storage of Billions of Photos, 2009. Facebook Engineering Notes, `http://tinyurl.com/cju2og`.

[21] Key Facts, Facebook Newsroom, 2012. `http://newsroom.fb.com/Key-Facts`.

[22] YouTube Fact Sheet. `http://www.youtube.com/t/fact_sheet`. Acessado em Dezembro/2012, 2011.

[23] L. Fan, P. Cao, J. Almeida, and A. Broder. Summary Cache: a Scalable Wide-area Web Cache Sharing Protocol. *IEEE / ACM Transactions on Networking*, 8(3):281–293, 2000.

[24] A. Gavras, A. Karila, S. Fdida, M. May, and M. Potts. Future internet research and experimentation: the fire initiative. *SIGCOMM Comput. Commun. Rev.*, 37:89–92, July 2007.

[25] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *ACM SIGCOMM conference on Internet measurement (IMC)*, 2007.

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

[27] T. Lins, F. Benevenuto, W. Dores, and F. Barth. Object popularity distributions in online social networks. In *ACM SIGWEB Web Science Conference (WebSci)*, 2012.

[28] D. Krishnamurthy M. Arlitt and J. Rolia. Characterizing the scalability of a large web-based shopping system. In *ACM Transactions on Internet Technology*, pages 44–69, 2001.

[29] D. Menascé and V. Almeida. *Scaling for E Business: Technologies, Models, Performance, and Capacity Planning.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[30] D. Menascé, V. Almeida, R. Fonseca, and M. Mendes. A methodology for workload characterization of e-commerce sites. In *ACM Conf. on Electronic Commerce (EC)*, 1999.

[31] D. Menascé, V. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. Meira Jr. In search of invariants for e-business workloads. In *ACM conference on Electronic commerce (EC)*, pages 56–65, New York, NY, USA, 2000. ACM.

[32] A. Noulas, C. Mascolo S. Scellato, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *SMW 2011*, 2011.

[33] Ad. Oke and R. Bunt. Hierarchical workload characterization for a busy web server. In *Int'l Conf. on Computer Performance Evaluation, Modelling Techniques and Tools (TOOLS)*, 2002.

[34] A. Pereira, L. Silva, and W. Meira Jr. Evaluating the impact of reactive workloads on the performance of web applications. In *Proceedings of the 25th IEEE International Performance, Computing, and Communications Conference (IPCCC)*, Phoenix, Arizona, USA, 2006. IEEE CS.

[35] S. Scellato. Beyond the social web: the geo-social revolution. *SIGWEB Newsletter*, pages 5:1–5:5, September 2011.

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

[37] F. Benevenuto T. Lins, H. Costa. Caracterização e modelagem do tráfego e da navegação dos usuários do apontador. *WPerformance - SBC 2012*, 2012.

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

[39] P. Rodriguez V. Erramilli, X. Yanga. Explore what-if scenarios with song: Social network write generator. `http://arxiv.org/abs/1102.0699`, 2012.

[40] M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Caracterizacão e influência do uso de tips e dones no foursquare. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, 2012.

[41] E. Veloso, V. Almeida, W. Meira Jr., A. Bestavros, and S. Jin. A hierarchical characterization of a live streaming media workload. *IEEE/ACM Transactions on Network*, 14(1):133–146, February.

[42] J. Wang. A Survey of Web Caching Schemes for the Internet. *ACM Computer Communication Review*, 25(9):36–46, 1999.

[43] M. Wittie, V. Pejovic, L. Deek, K. Almeroth, and B. Zhao. Exploiting locality of interest in online social networks. In *ACM Int'l Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, pages 1–12, 2010.