Uma Fotografia do Instagram: Caracterização e Aplicação

Thiago H. Silva[1], **Pedro O. S. Vaz de Melo**[1], **Jussara M. Almeida**[1], **Antonio A. F. Loureiro**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
31270-010 Belo Horizonte, MG – Brasil

{thiagohs, olmo, jussara, loureiro}@dcc.ufmg.br

Abstract. *Participatory sensing systems (PSS) have the potential to become fundamental tools for supporting the study, in large scale, of urban social behavior and city dynamics. To that end, this work characterizes the photo sharing system Instagram, considered one of the currently most popular PSSs on the Internet. Based on a dataset of approximately 2.3 million shared photos, we characterize user behavior in the system showing that there are several advantages and opportunities for large scale sensing, such as a global coverage at low cost, but also challenges, such as a very unequal photo sharing frequency, both spatially and temporally. Moreover, we present an application based on data obtained from Instagram to identify regions of interest in a city, which illustrates the promising potential of PSSs for the study of city dynamics.*

Resumo. *Sistemas de sensoriamento participativo (SSP) (participatory sensing systems) têm o potencial de se tornarem ferramentas fundamentais para o estudo em larga escala do comportamento social urbano e da dinâmica de cidades. Nessa direção, este trabalho analisa o Instagram, um sistema para compartilhamento de fotos que é considerado um dos mais populares SSPs disponíveis na Internet atualmente. Baseado em um conjunto de dados de aproximadamente 2,3 milhões de fotos compartilhadas, caracterizamos o comportamento de usuários desse sistema, mostrando que existem muitas vantagens e oportunidades para sensoriamento em grande escala, tais como uma abrangência global a baixo custo, mas também desafios, como uma frequência de compartilhamento de fotos espaço-temporal altamente desigual. Além disso, apresentamos uma aplicação baseada em dados obtidos do Instagram para identificar regiões de interesse dentro de uma cidade. Essa aplicação ilustra o potencial promissor de SSPs para o estudo da dinâmica das cidades.*

1. Introdução

Mark Weiser, no seu trabalho clássico intitulado "*The computer for the 21st century*" publicado na *Scientific American* [Weiser 1991], popularizou o conceito da *computação ubíqua*, que prevê o acesso a ambientes de computação por qualquer pessoa, em qualquer lugar, a qualquer instante, com dispositivos computacionais acoplados aos mais triviais objetos, como etiquetas de roupas, copos de café, canetas ou qualquer objeto pessoal. Embora essa ainda não seja a realidade e esse conceito tenha sido estendido para incluir, por exemplo, Internet das Coisas (*Internet of Things*), muito foi feito nessa direção nestes últimos 20 anos após a publicação do trabalho de Weiser. Por exemplo, as Redes de Sensores Sem Fio (RSSFs) [Akyildiz et al. 2002], que são um tipo especial de rede *ad hoc*, são projetadas para coletar dados referentes a grandezas físicas dos mais variados ambientes em que estão inseridas e a fornecer tais informações para o usuário final. Além disso, há o uso crescente de sistemas de sensoriamento participativo

(SSPs) [Burke et al. 2006], que permitem a pessoas conectadas à Internet fornecerem dados de contexto sobre o ambiente em que estão em um determinado momento.

De fato, os SSPs têm o potencial para complementar as RSSFs em diversos aspectos. Enquanto as RSSFs foram projetadas para sensoriar áreas de tamanho limitado, como florestas e vulcões, os SSPs podem alcançar áreas de tamanhos variados e de larga escala, como grandes metrópoles, países ou até mesmo todo o planeta [Silva et al. 2012a]. Além disso, uma RSSF está sujeita a falhas, uma vez que o seu funcionamento depende da correta coordenação das ações dos seus nós sensores, que possuem severas restrições de energia, processamento e memória. Por outro lado, SSPs são formados por entidades autônomas e independentes, os seres humanos, o que torna a tarefa de sensoriamento altamente resiliente a falhas individuais.

O sucesso dos SSPs está diretamente ligado à popularização do *smartphone*, que se tornou o dispositivo computacional pessoal mais amplamente adotado e onipresente [Krumm 2009]. Os *smartphones* possuem um rico conjunto de sensores embutidos, tais como GPS, acelerômetro, microfone, câmera, giroscópio e bússola digital. Entretanto, o sensoriamento não depende apenas dos dados gerados por esses sensores, podendo vir também das observações subjetivas do seu usuário. É possível encontrar vários exemplos de SSPs já implantados e usados através de *smartphones*, como o Waze[1], para relatar condições de tráfego em tempo real, e o Weddar[2], para relatar condições meteorológicas. Além disso, há serviços de compartilhamento de localização, como o Foursquare[3], ou de fotos, como o Instagram[4], nos quais os usuários podem enviar imagens em tempo real para o sistema. Em particular, o Instagram é um dos mais populares SSPs atuais, com quase 100 milhões de usuários e mais de 1 bilhão de fotos recebidas, sendo que, a cada segundo, um novo usuário se registra no sistema e 58 novas fotos são inseridas [Daniells 2012].

O objetivo principal deste trabalho é caracterizar a rede de participação do Instagram, visando mostrar os desafios e as oportunidades que emergem do sensoriamento participativo realizado pelos usuários desta aplicação. Baseado em um conjunto de dados de aproximadamente 2,3 milhões de fotos, mostramos a abrangência planetária da rede, assim como a frequência altamente desigual do compartilhamento de fotos, tanto espacial quanto temporalmente, que é bastante correlacionada com as rotinas de atividades humanas. Além disso, mostramos também como é possível projetar aplicações a partir de sistemas como o Instagram, apresentando uma aplicação para identificar regiões de interesse dentro de uma cidade. Essa aplicação ilustra o potencial de SSPs para o estudo da dinâmica de cidades. Até onde sabemos, este é o primeiro trabalho de caracterização do uso do Instagram, particularmente com o foco no seu potencial, como um sistema de sensoriamento participativo no projeto de novas aplicações e serviços.

O restante deste trabalho é organizado como segue. A seção 2 apresenta os trabalhos relacionados. A seção 3 discute a participação dos seres humanos no processo de sensoriamento, abordando os sistemas participativos de sensoriamento e as redes de sensores participativos (RSPs), advindas de SSPs. A seção 4 apresenta a caracterização da RSP derivada do Instagram. A seção 5 descreve uma aplicação de classificação de regiões usando esta RSP. Finalmente, a seção 6 apresenta as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

O processo de sensoriamento do meio ambiente pode envolver seres humanos como: (i) alvos do processo [Larson et al. 2011], ou (ii) responsáveis por eles a partir do compartilhamento

[1]http://www.waze.com
[2]http://www.weddar.com
[3]http://www.foursquare.com
[4]http://www.instagram.com

de dados locais [Srivastava et al. 2012, Goodchild 2007]. Neste trabalho, focamos no segundo caso, considerando sistemas que utilizam dispositivos móveis do dia a dia, como *smartphones*, para construir uma rede de sensoriamento participativa, que é descrita na seção 3.2. Na literatura, existem diversas propostas de sistemas que consideram a participação de humanos no processo de sensoriamento, como descrito anteriormente. Tais sistemas são chamados de sistemas de sensoriamento participativo (SSPs) e incluem, por exemplo, sistemas de monitoramento de tráfego [Eisenman et al. 2010] e monitoramento de ruídos [Rana et al. 2010].

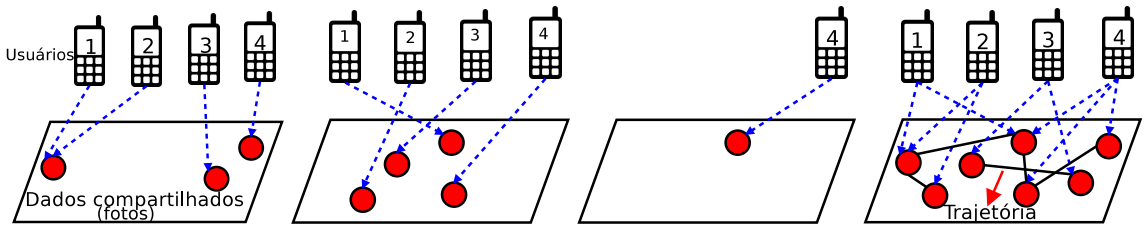
O sucesso de SSPs depende fundamentalmente da participação sustentável dos usuários ao longo do tempo. Em [Reddy et al. 2010], os autores propõem mecanismos de incentivos baseados em micro-pagamentos (*micro-payments*), que são pequenas quantias de dinheiro dadas ao usuário quando ele realiza determinadas atividades no sistema. Além da participação sustentável, é necessário garantir a qualidade dos dados compartilhados pelos usuários [Mashhadi and Capra 2011]. Por exemplo, em diversos SSPs os usuários podem fabricar dados falsos, que supostamente foram sensoriados, a baixo custo. Logo, a integridade dos dados não é sempre garantida [Saroiu and Wolman 2010].

Existem vários trabalhos dedicados ao estudo das características de SSPs específicos. Por exemplo, em serviços de compartilhamento de localização, como o Foursquare, os autores de [Cheng et al. 2011] observaram que os usuários seguem um padrão de mobilidade simples e factível de ser reproduzido. Nessa direção, os autores de [Cho et al. 2011] observaram que viagens de curta distância são periódicas no espaço e no tempo e não são afetadas pela estrutura social da rede, que, por sua vez, influencia somente as viagens de longa distância. De fato, Scellato et al. [Scellato et al. 2011] mostraram que 40% das relações sociais no sistema analisado acontecem a menos de 100 km. Em [Noulas et al. 2011a], os autores analisaram a dinâmica de compartilhamento dos usuários de serviços de compartilhamento de localização mostrando, por exemplo, que a distribuição do número de check-ins é altamente desigual, sendo bem modelada por um comportamento de lei de potência (*power-law*). Em [Vasconcelos et al. 2012], os autores analisaram o uso de *tips* no Foursquare, que são comentários curtos sobre determinado local, caracterizando como os usuários interagem entre si utilizando esta funcionalidade e propuseram um algoritmo para estimar a influência de usuários. Diferentemente dos demais trabalhos, os autores não exploraram características geográficas.

Outros trabalhos propõem utilizar dados derivados de SSPs em novas aplicações, já que esse tipo de dado auxilia no melhor entendimento das fronteiras físicas e noções de espaço [Bilandzic and Foth 2012]. Nessa direção, os autores de [Cranshaw et al. 2012] apresentaram um modelo que classifica regiões de uma cidade a partir de padrões de atividades coletivas, enquanto Noulas et al. [Noulas et al. 2011b] propuseram classificar áreas e usuários de uma cidade usando as categorias dos locais registrados no Foursquare.

Em trabalho anterior [Silva et al. 2012a], analisamos as propriedades de RSPs derivadas de duas aplicações de compartilhamento de localização: Gowalla e Brightkite. Nós analisamos as distribuições espacial e temporal dos *check-ins* realizados pelos usuários desses sistemas, visando levantar evidências relevantes para o projeto de novos serviços e aplicações. Já em [Silva et al. 2012b], propusemos uma nova forma de visualizar a dinâmica de cidades a partir de hábitos e rotinas de usuários, derivados dos *check-ins* no Foursquare.

O trabalho aqui apresentado diferencia-se dos anteriores (incluindo os nossos) por focar em um novo sistema de grande popularidade atualmente – o Instagram. Até onde sabemos, esta é a primeira caracterização de uma aplicação de compartilhamento de fotos acessada, majoritariamente, a partir de *smartphones*. Mais ainda, dando continuidade aos trabalhos recentes [Silva



(a) Tempo 1 (b) Tempo 2 (c) Tempo 3 (d) Tempo total

Figura 1. RSP analisada: serviço de compartilhamento de fotos

et al. 2012a, Silva et al. 2012b], este trabalho também aborda o estudo da dinâmica das cidades através de SSPs, mostrando que sistemas de compartilhamento de fotos, particularmente o Instagram, também podem ser utilizados para este propósito.

3. Humanos no Processo de Sensoriamento

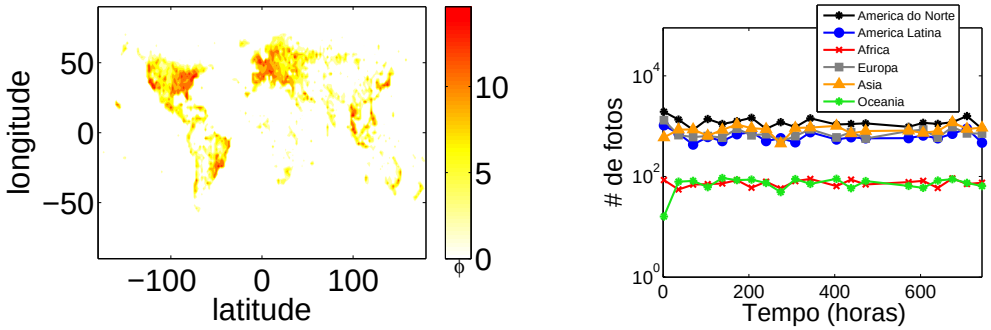
O foco deste trabalho é em sistemas que dependem de humanos no processo de sensoriamento, sendo eles responsáveis pelo compartilhamento de dados locais. Tais dados podem ser obtidos com o auxílio de dispositivos de sensoriamento, como sensores embutidos em celulares (por exemplo, GPS), ou através de sensores humanos (por exemplo, visão), compartilhando informações produzidas por eles próprios.

3.1. Sensoriamento Participativo

Sensoriamento participativo é o processo em que seres humanos usam dispositivos móveis e serviços de computação em nuvem para compartilhar dados sensoriados [Burke et al. 2006]. Usualmente, sistemas de sensoriamento participativo consideram que o compartilhamento dos dados é gerado automaticamente, ou passivamente, por sensores embutidos no dispositivo móvel. Porém, neste trabalho consideramos também observações geradas pelos usuários de forma manual, ou proativa. Sensoriamento participativo com essas características também pode ser chamado de *crowdsourcing* ubíquo (*ubiquitous crowdsourcing*) [Mashhadi and Capra 2011]. A popularidade de sistemas para sensoriamento participativo cresceu rapidamente com o aumento do uso de celulares com sensores embutidos e capacidade de acesso à Internet, ou seja, os chamados *smartphones*. Esses dispositivos se tornaram uma poderosa plataforma que inclui capacidades de sensoriamento, computação e comunicação.

Um dado sensoriado em uma aplicação de sensoriamento participativo é: (i) obtido através de sensores físicos (por exemplo, GPS) ou observações humanas (por exemplo, congestionamento na rodovia); (ii) definido no tempo e no espaço; (iii) obtido automaticamente ou manualmente; (iv) estruturado ou não; (v) compartilhado voluntariamente ou não. Para ilustrar esse tipo de sistema, considere uma aplicação para monitoramento de trânsito, como o Waze. Usuários podem compartilhar observações sobre acidentes ou congestionamentos manualmente. Uma aplicação poderia ainda calcular e compartilhar automaticamente a velocidade de um carro com o auxílio de dados obtidos com o GPS. Com as medidas da velocidade de diferentes veículos amostrados em uma área particular, é possível inferir, por exemplo, congestionamentos. Como nesse caso específico usuários operam uma aplicação que foi criada para esse propósito, o dado sensoriado é estruturado. Mas, caso os usuários usem um serviço de *microblogging*, como o Twitter, o dado sensoriado seria não estruturado. Por exemplo, o usuário "João" envia uma mensagem "trânsito agora está muito lento próximo da portaria do campus".

Serviços de compartilhamento de fotos, como o Instagram, são exemplos de aplicações para sensoriamento participativo. O dado sensoriado é uma foto de um lugar específico. Pode-



(a) Número de fotos n por pixel dado pelo valor de ϕ mostrado na figura, em que $n = 2^{\phi} - 1$.

(b) Variação temporal do número de fotos compartilhadas por continente.

Figura 2. Cobertura da RSP do Instagram.

mos extrair informação desse tipo de dado de diversas maneiras. Por exemplo, é possível visualizar em tempo real como está a situação de uma certa área da cidade.

3.2. Rede de Sensoriamento Participativo

Em uma rede de sensoriamento participativo (RSP), o dispositivo móvel do usuário é uma peça fundamental. Indivíduos carregando esses dispositivos são capazes de sensoriar o ambiente e fazer observações relevantes. Assim, cada nó em uma RSP consiste de um usuário com o seu dispositivo móvel. De forma similar às RSSFs, o dado sensoriado é enviado para o servidor, ou "nó sorvedouro". Mas, diferentemente das RSSFs, RSPs têm as seguintes características: (a) nós são entidades móveis autônomas, mas uma pessoa com um dispositivo móvel; (b) o custo da rede é distribuído entre os nós, proporcionando uma escalabilidade global; (c) o sensoriamento depende da vontade das pessoas participarem no processo de sensoriamento; (d) nós transmitem o dado sensoriado diretamente para o sorvedouro; (e) nós não sofrem de severas limitações de energia; (f) o nó sorvedouro só recebe dados e não tem controle direto sobre os nós.

A figura 1 mostra um exemplo de RSP formada a partir de serviços de compartilhamento de fotos, que é a RSP analisada nas seções seguintes. As figuras 1-a, 1-b e 1-c representam quatro usuários em três diferentes momentos. Fotos compartilhadas pelos usuários a cada momento são marcadas por uma seta pontilhada. Observe que nem todos os usuários realizam atividades em todos os momentos. Depois de um certo intervalo, podemos analisar os dados de diversas maneiras. Por exemplo, a figura 1-d mostra um grafo onde os vértices representam os locais onde as fotos foram compartilhadas e as arestas conectam fotos compartilhadas pelo mesmo usuário. Com esse grafo é possível extrair várias informações interessantes de diferentes partes do mundo, fornecendo uma notável escala global a uma infraestrutura de baixo custo, como ilustrado na figura 2-a.

4. Caracterização do Instagram

Nesta seção nós analisamos uma rede de sensores participativa (RSP) derivada do Instagram.

4.1. Descrição dos Dados

A RSP analisada é derivada de um conjunto de dados do Instagram, que é um serviço de compartilhamento de fotos online. Os dados do Instagram foram coletados através do Twitter[5], que é um serviço de *microblogging*, ou seja, ele permite que os seus usuários enviem e recebam atualizações pessoais de outros contatos em textos de até 140 caracteres, conhecidos como

[5]http://www.twitter.com

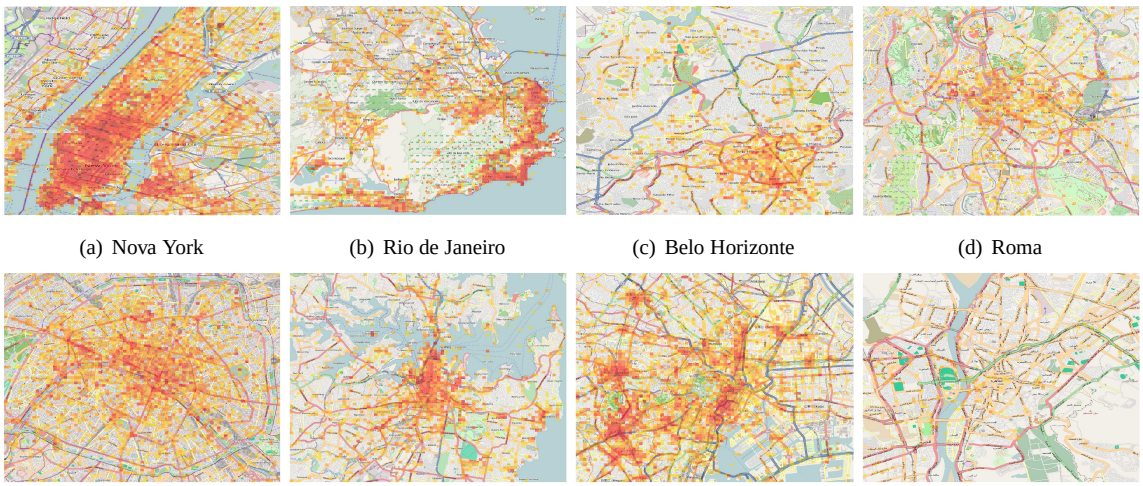
"*tweets*". Além de *tweets* de texto simples, os usuários também podem compartilhar fotos a partir de uma integração com o Instagram. Neste caso, fotos do Instagram anunciadas no Twitter passam a ficar disponíveis publicamente, o que por padrão não acontece quando a foto é publicada unicamente no sistema do Instagram.

Entre 30 de junho e 31 de julho de 2012, foram coletados 2 272 556 *tweets* contendo fotos georeferenciadas, postadas por 482 629 usuários. Cada *tweet* é composto de coordenadas GPS (latitude e longitude) e o horário do compartilhamento da foto.

4.2. Cobertura da Rede

Nesta seção, analisamos a cobertura da RSP do Instagram em diferentes granularidades espaciais, começando por todo o planeta, depois por continentes e cidades e, por fim, até áreas específicas de uma cidade. A figura 2-a mostra a cobertura no planeta da RSP do Instagram na forma de um mapa de calor da participação dos usuários: cores mais escuras representam um maior número de fotos compartilhadas em determinada área. Apesar da cobertura ser bastante abrangente na escala planetária, ela não é homogênea. A figura 2-b mostra o número de fotos compartilhadas por continente ao longo do tempo. Note que a atividade de sensoriamento nas Américas, Europa e Ásia é pelo menos uma ordem de magnitude maior que na África e Oceania. Além disso, pode-se observar ainda que a participação dos usuários da América do Norte é levemente superior a dos usuários da América Latina, Europa e Ásia.

Avaliamos agora a participação dos usuários do Instagram em oito grandes e populosas cidades localizadas em cinco continentes: Nova York, Rio de Janeiro, Belo Horizonte (BH), Roma, Paris, Sydney, Tokyo, e Cairo. A figura 3 mostra o mapa de calor da atividade de sensoriamento para cada uma dessas cidades. Mais uma vez, cores mais escuras representam um maior número de fotos em determinada área. Pode-se observar uma alta cobertura para algumas cidades, como mostrado nas figuras 3-a (Nova York), 3-e (Paris) e 3-g (Tokyo). No entanto, pode-se observar na figura 3-h que o sensoriamento no Cairo, que também possui um número elevado de habitantes, é significativamente mais baixo. Tamanha diferença na cobertura pode ser explicada por diversos fatores. Além dos aspectos econômicos, diferenças na cultura dos habitantes desta cidade quando comparadas com as culturas presentes nas outras cidades estudadas podem ter um impacto significativo na adoção e uso do Instagram [Barth 1969].



(a) Nova York (b) Rio de Janeiro (c) Belo Horizonte (d) Roma

(e) Paris (f) Sydney (g) Tokyo (h) Cairo

Figura 3. Cobertura espacial da RSP do Instagram em 8 cidades: todas as fotos compartilhadas. O número de fotos em cada área é representado por um mapa de cores, onde a escala vai de amarelo a vermelho (atividade mais intensa).

Além disso, pode-se observar que a cobertura no Rio de Janeiro e em Sydney é bem mais heterogênea quando comparada com a cobertura em Paris, Tóquio e Nova York. Isto ocorre provavelmente por causa dos aspectos geográficos que estas cidades têm em comum, ou seja, grandes áreas verdes e grandes porções d'água. Rio de Janeiro, por exemplo, tem a maior floresta urbana do mundo, localizada no meio da cidade, além de muitas colinas de difícil acesso humano. Estes aspectos geográficos limitam a cobertura do sensoriamento. Além disso, em ambas as cidades os pontos de interesse público, tais como pontos turísticos e centros comerciais, são distribuídos de forma desigual pela cidade. Há grandes áreas residenciais com poucos pontos desse tipo, enquanto outras áreas têm grande concentração desses pontos. Estes resultados são qualitativamente semelhantes aos reportados em [Silva et al. 2012a, Silva et al. 2012b] para RSPs derivadas de três aplicações de compartilhamento de localização e para diferentes cidades, o que demonstra o potencial do Instagram como ambiente para sensoriamento participativo em grandes regiões.

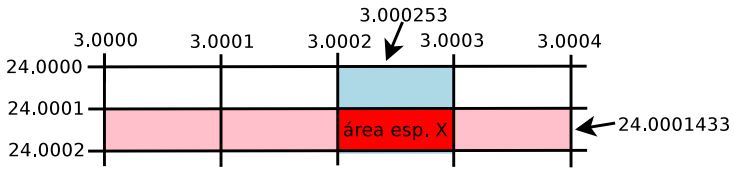


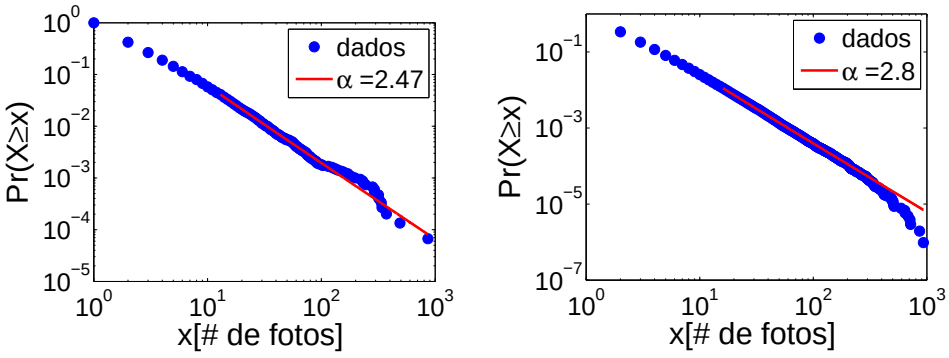
Figura 4. Exemplo de identificação de uma área específica

Uma vez que a atividade de participação pode ser bastante heterogênea dentro de uma cidade, propomos dividir a área de cidades em espaços retangulares menores, como em uma grade[6]. Chamaremos cada área retangular de uma *área específica* dentro de uma cidade e, a partir disso, analisaremos o número de fotos compartilhadas nessas áreas específicas. Neste trabalho, consideramos que uma área específica possui a seguinte delimitação: $1{\cdot}10^{-4^\circ}$ (latitude) $\times$ $1{\cdot}10^{-4^\circ}$ (longitude). Isso representa uma área de aproximadamente 8×11 metros em NY e 10×11 metros no Rio de Janeiro. Para outras cidades, as áreas também podem variar um pouco, mas não a ponto de afetar significativamente as análises realizadas. A figura 4 ilustra o processo de divisão da área de uma cidade em áreas específicas e de como é feita a associação da coordenada geográfica (24,0001433; 3, 000253) a uma área específica X.

A figura 5 apresenta a função de distribuição acumulada complementar (CCDF) do número de fotos compartilhadas por área específica da cidade de Nova York (figura 5-a) e de todas as localidades em nossa base de dados (figura 5-b). Primeiramente, observe que, em ambos os casos, uma lei de potência[7] descreve bem esta distribuição. Isso implica que, na maioria das áreas específicas, há poucas fotos compartilhadas, enquanto existem algumas poucas áreas com centenas de fotos compartilhadas. Estes resultados estão consistentes com os resultados apresentados em [Noulas et al. 2011a, Silva et al. 2012a], que estudaram a participação de usuários em sistemas de compartilhamento de localização. Em sistemas para compartilhamento de fotos, assim como em sistemas de compartilhamento de localizações, é natural que algumas áreas possuam mais atividade que outras. Por exemplo, em áreas turísticas o número de fotos compartilhadas tende a ser maior do que em um supermercado, apesar de um supermercado ser geralmente um local bastante popular. Se uma determinada aplicação requer uma cobertura mais abrangente, é necessário incentivar os usuários a participarem em locais que eles usualmente não o fariam. Micro-pagamentos ou sistemas de pontuação são exemplos de alternativas que poderiam funcionar nesse caso.

[6]Note que nas áreas selecionadas não é considerado fronteiras.

[7]Matematicamente, uma quantidade x segue uma lei de potência se ela pode ser obtida de uma distribuição de probabilidade $p(x) \propto x^{-\alpha}$, onde α é um parâmetro constante conhecido como expoente ou parâmetro escalar, e é um valor tipicamente entre $2 < \alpha < 3$.



(a) NY (b) Todas as localidades

Figura 5. Distribuição do número de fotos em áreas específicas

Como foi mostrado anteriormente, uma RSP pode ter uma cobertura em escala planetária. No entanto, foi mostrado também que essa cobertura pode ser bastante heterogênea, em que grandes áreas ficam praticamente descobertas. A figura 6 mostra a cobertura da rede total considerando a dimensão temporal, ou seja, o número de localidades que estão ativas (i.e., sensoriadas) em um determinado intervalo de tempo, considerando todos os dados disponíveis. O número máximo de áreas específicas sensoriadas por hora corresponde a aproximadamente somente 0,2% do número total de áreas em nossa base de dados (1 030 558). Em outras palavras, a cobertura instantânea da RSP do Instagram é muito limitada quando consideramos todas as localidades que poderiam ser sensoriadas no planeta[8]. Isso significa que a probabilidade de uma área específica aleatória ser sensoriada em um horário aleatório é bem baixa.

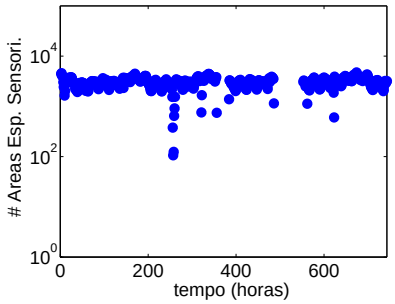


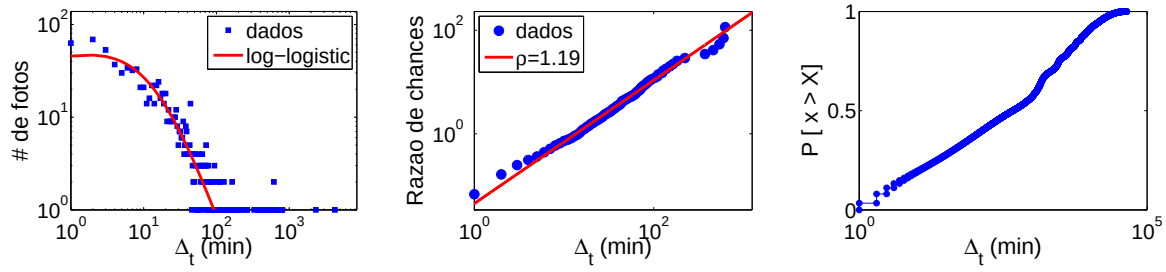
Figura 6. Variação temporal do número de áreas específicas sensoriadas.

4.3. Intervalo de Sensoriamento

Redes de sensoriamento participativo são bastante escaláveis porque seus nós são autônomos, ou seja, os usuários são responsáveis pela sua própria operação e funcionamento. Como o custo da infraestrutura é distribuído entre os participantes, esta enorme escalabilidade e cobertura é alcançada mais facilmente. O sucesso desse tipo de rede consiste em ter participação sustentável e de alta qualidade. Em outras palavras, o sensoriamento é eficiente desde que os usuários sejam mantidos motivados a compartilharem seus recursos e dados sensoriados frequentemente.

Investigamos agora a frequência com que usuários do Instagram realizam o compartilhamento de fotos. A figura 7-a mostra o histograma do intervalo de tempo Δ_t entre compartilhamentos de fotos consecutivos em uma determinada área específica tipicamente popular. Note que o histograma é bem ajustado por uma distribuição *log-logistic* [Fisk 1961]. Observe as rajadas de atividade e os longos períodos de inatividade: há momentos em que muitas fotos são

[8]Considerando nesse caso todas as localidades já sensoriadas pelo menos uma vez.



(a) Histograma (b) Razão de chances (c) Distribuição acumulada

Figura 7. Distribuição do intervalo de tempo entre compartilhamentos de fotos em uma área específica popular.

compartilhadas em intervalos de poucos minutos e momentos em que não há compartilhamento por horas. Isso pode indicar que a maioria do compartilhamento de fotos, nesta área popular (assim como em outras), acontece em intervalos específicos, provavelmente relacionados ao horário em que as pessoas usualmente as visitam. Por exemplo, o compartilhamento de fotos em restaurantes tende a acontecer mais nos horários de almoço e jantar. Aplicações baseadas nesse tipo de sensoriamento devem considerar que a participação do usuário pode variar significativamente ao longo do tempo.

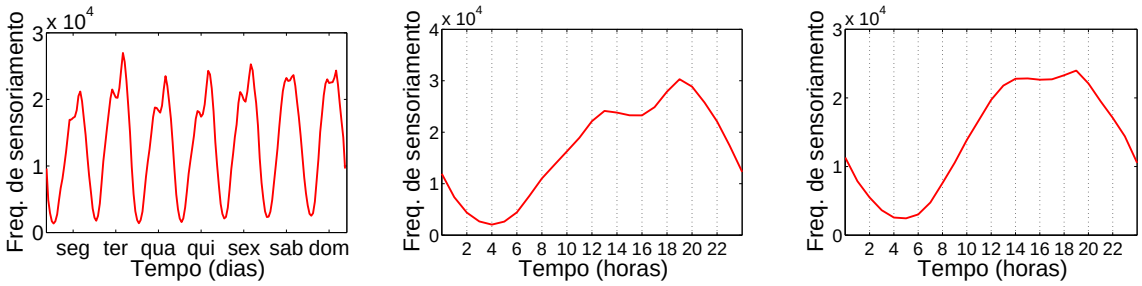
Outra observação interessante relacionada ao intervalo de tempo Δ_t entre compartilhamentos pode ser extraída da figura 7-b, que mostra a função razão de chances[9] (RC) desses intervalos. A RC é uma função acumulada que mostra claramente o comportamento cumulativo de uma dada distribuição na cabeça quanto na cauda. Sua fórmula é $RC(\Delta_t) = \frac{CDF(\Delta_t)}{1-CDF(\Delta_t)}$, onde $CDF(\Delta_t)$ é a função de densidade acumulada da distribuição sendo analisada, no caso a distribuição dos intervalos de tempo Δ_t entre compartilhamentos. Como em [Vaz de Melo et al. 2011], a RC do intervalo de tempo entre fotos compartilhadas também mostra um comportamento de lei de potência com inclinação $\rho \approx 1$. Isso sugere que os mecanismos por trás das atividades humanas podem ser mais simples e gerais do que aqueles propostos na literatura, pois dependem de uma grande quantidade de parâmetros [Malmgren et al. 2008]. A figura 7-c mostra a distribuição do intervalo entre eventos. Podemos observar que uma fatia significativa dos usuários realiza compartilhamento consecutivo de fotos em um curto intervalo de tempo. Cerca de 20% de todo compartilhamento observado acontece em até 10 minutos. Como será discutido na seção 4.5, isso sugere que os nós tendem a compartilhar mais de uma foto na mesma área.

4.4. Sazonalidade

Analisamos agora como a rotina dos humanos afeta o compartilhamento dos dados. A figura 8-a mostra o padrão semanal de compartilhamento de fotos do Instagram[10]. Como esperado, a atuação da rede de participação apresenta um padrão diurno, o que implica que durante a madrugada a atividade de sensoriamento é bastante baixa. Considerando dias de semana, é possível observar um ligeiro aumento da atividade ao longo da semana, com exceção de terça-feira, quando há um pico de atividade. No trabalho [Cheng et al. 2011], que analisou sistemas para compartilhamento de localização, foi observado esse mesmo comportamento, sem nenhum dia como exceção. Isso sugere que durante o período de coleta pode ter ocorrido um evento atípico durante a terça-feira que gerou muitos compartilhamento de fotos. Por fim, pode-se

[9] *Odds ratio function.*

[10] O horário do compartilhamento foi normalizado de acordo com o local onde a foto foi tirada, utilizando para isso a informação geográfica do local.



(a) Padrão semanal (b) Agregado - Seg. a Sex. (c) Agregado - Sáb. a Dom.

Figura 8. Padrão do compartilhamento de fotos durante os dias da semana

observar dois picos de atividades ao longo do dia, por volta dos horários de almoço e jantar. Diferentemente do comportamento observado para o compartilhamento de localizações [Cheng et al. 2011], não foi observado picos de atividade no compartilhamento de fotos por volta do horário do café da manhã.

Analisamos ainda os diferentes padrões de comportamento para dias de semana e final de semana. A figura 8-b mostra o número médio de fotos compartilhadas por hora, de segunda-feira a sexta-feira. A figura 8-c também mostra a mesma informação para sábado e domingo. Como podemos observar, os picos durante os dias de semana acontecem por volta de 13:00 (almoço) e 19:00 (jantar). Já no final de semana não é observado um pico de atividade claro durante o horário do almoço. Pelo contrário, a atividade permanece intensa durante toda a tarde até o início da noite, com um ligeiro aumento por volta das 19:00.

4.5. Comportamento dos Nós

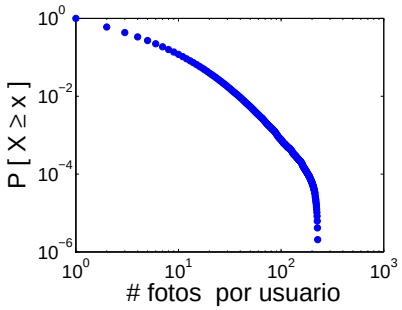
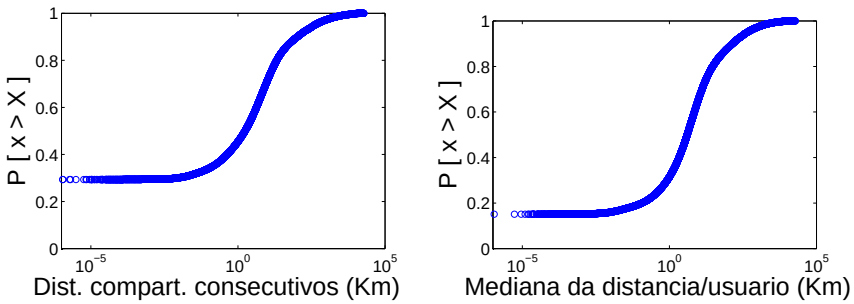


Figura 9. Distribuição do número de fotos compartilhadas pelos usuários



(a) Todas as distâncias (b) Mediana da distância por usuário

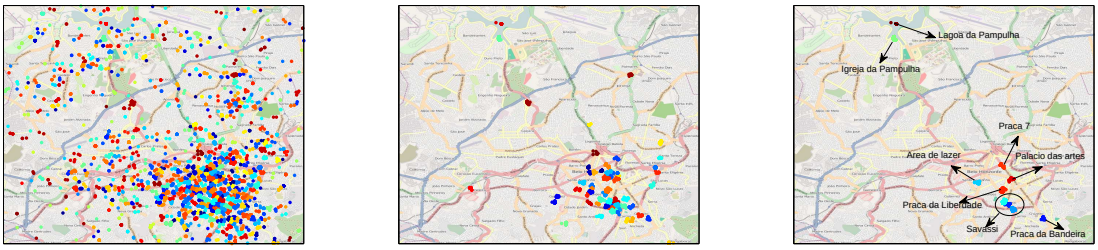
Figura 10. Distribuição da distância geográfica entre fotos consecutivas de um mesmo usuário.

Nesta seção é analisado o desempenho dos nós da RSP (i.e., dos usuários) quanto ao

compartilhamento de fotos. A figura 9 mostra que a distribuição do número de fotos compartilhadas por cada usuário da nossa base de dados possui cauda pesada, significando que a participação dos usuários pode variar muito. Por exemplo, aproximadamente 40% dos usuários contribuíram com apenas uma foto no período considerado, enquanto que somente 17% e 0.1% dos usuários contribuíram com mais que 10 e 100 fotos, respectivamente. É natural que essa variabilidade aconteça por diversos motivos. Por exemplo, alguns usuários podem dar mais importância para quesitos de privacidade do que outros.

Analisamos também a distância geográfica entre dois compartilhamentos de fotos consecutivos pelo mesmo usuário, usando, para tal, as coordenadas geográficas associadas a cada foto. A figura 10-a mostra a função de densidade acumulada da distância geográfica entre cada par de fotos consecutivas compartilhadas por cada usuário do nosso conjunto de dados. Pode-se observar que uma fatia significativa (aproximadamente 30%) das distâncias entre fotos consecutivas são muito curtas (menos de 1 metro). Isso indica que os usuários tendem a compartilhar várias fotos no mesmo local. Essa hipótese é reforçada pela significativa fatia de intervalos de tempo entre fotos consecutivas de curta duração mostrada na figura 7-c: 20% destes intervalos (Δ_t) não ultrapassam 10 minutos. Isso não foi observado na mesma proporção para o compartilhamento de localização. Em [Noulas et al. 2011a], por exemplo, foi observado que 20% dos compartilhamentos de localizações acontece em até 1 km de distância. Para o compartilhamento de fotos, esse valor chega a aproximadamente 45%. Esse resultado pode ser explicado pelo simples fato de que uma foto pode conter muito mais informações que uma localização. Por exemplo, em um restaurante os usuários poderiam compartilhar fotos dos amigos presentes, da comida, ou de uma situação particular, mas tenderiam a compartilhar sua localização apenas uma única vez.

Por fim, analisamos cada usuário separadamente. A figura 10-b mostra a distribuição das medianas das distâncias entre compartilhamentos consecutivos computadas para cada usuário. Note que pelo menos 50% das fotos consecutivas de uma parcela significativa de usuários (aproximadamente 20%) são tiradas a uma distância muito pequena ($\approx$1 metro).



(a) Todos os clusters (b) Pontos de interesse (c) Pontos turísticos

Figura 11. Pontos de Interesse de Belo Horizonte

5. Aplicação

É bastante comum existirem áreas em uma cidade que despertam um maior interesse dos residentes ou visitantes, os aqui denominados *pontos de interesse* (PDI). Dentre os PDIs mais visitados, podemos mencionar os pontos turísticos da cidade. No entanto, nem todos os PDIs de uma cidade são pontos turísticos. Por exemplo, uma área de bares pode ser bastante popular entre os residentes da cidade, mas sem atrativos turísticos. Além disso, PDIs são dinâmicos, ou seja, áreas que são populares hoje podem não o ser amanhã.

Assim, uma aplicação que emerge naturalmente a partir da análise de dados do Instagram é de identificação de PDIs em uma cidade. Isso é possível porque cada foto representa, implicitamente, um interesse de um indivíduo em um determinado instante. Com isso, quando muitas fotos de um determinado local são compartilhadas em um determinado instante, pode-se inferir que esse local é um PDI (observe a figura 5). Mais especificamente, o processo de identificação de PDIs envolve os seguintes passos:

1. Considera-se que cada par i de coordenadas (longitude, latitude) $(x, y)_i$ está associada a um ponto p_i;
2. calcula-se a distância [Sinnott 1984] entre cada par de pontos (p_i, p_j);
3. agrupa-se todos os pontos p_i que possuem uma distância inferior a 250 metros em um *cluster* C_k. Essa distância limite foi obtida através do método Acoplamento Completo (Complete-Linkage) [Sørensen 1948]. O resultado desse procedimento é exibido na figura 11-a. Nessa figura cores diferentes[11] representam diferentes clusters C_k para a cidade de Belo Horizonte;
4. para cada cluster C_k, consideramos apenas um ponto (foto) por usuário. Com isso, a popularidade de um cluster se baseia no número de diferentes usuários que compartilharam uma foto na área do cluster. Este procedimento evita considerar as áreas visitadas por poucos usuários, por exemplo casas, como populares;
5. para cada cluster C_k, cria-se um cluster alternativo C_r, assim $r = k$. A princípio cada cluster alternativo não possui nenhuma foto associada a ele. Em seguida, para cada foto f_i, escolhemos um cluster alternativo C_r de forma aleatória e atribuímos f_i a C_r. Após distribuir todas as fotos f_i seguindo esse procedimento, comprovamos que a distribuição do número de fotos atribuídas para cada cluster C_r é explicada por uma distribuição normal com média μ e desvio padrão σ;
6. por fim, dos clusters originais C_k encontrados a partir do passo 3, excluímos todos aqueles em que o número de fotos do mesmo está a uma distância 2σ da média μ, ou seja, está no intervalo $[\mu - 2\sigma; \mu + 2\sigma]$. De acordo com a regra dos três sigmas (*three-sigma rule*) esse intervalo representa ≈95% da distribuição de fotos aleatórias nos clusters C_r. A ideia deste passo é excluir aqueles clusters que podem ter sido gerados por situações aleatórias, ou seja, clusters que provavelmente não refletem um PDI da cidade.

Os PDIs obtidos através desse processo são mostrados na figura 11-b. Além de identificar PDIs em uma cidade, podemos separar dos PDIs os pontos turísticos. Para isso, primeiramente geramos um grafo $G(V, E)$, onde os vértices $v_i \in V$ são todos os PDIs e uma aresta (i, j) existe do vértice v_i para o vértice v_j se em um determinado momento um usuário compartilhou uma foto em um PDI v_j, logo após ter compartilhado uma foto no PDI v_i. O peso $w(i, j)$ de uma aresta representa o número total de transições realizadas do PDI v_i para o PDI v_j, considerando as transições de todos os usuários. Para identificar pontos turísticos, consideramos que grande parte dos turistas seguem uma trajetória bem conhecida dentro da cidade, sendo guiada pelos principais pontos turísticos da mesma. Além disso, em cada ponto turístico ele tira uma ou mais fotos e parte para o ponto turístico seguinte. Dessa maneira, consideramos que arestas (i, j) com pesos $w(i, j)$ altos denotam essas transições frequentes de um ponto turístico para outro em uma cidade.

Feito isso, excluímos de G todas as arestas (i, j) com peso $w(i, j)$ menor que um limiar t, que é dado pela probabilidade de gerar $w(i, j)$ aleatoriamente em um grafo aleatório $G_R(V, E_R)$. A identificação do valor que separa arestas de peso alto das de peso baixo é feita da seguinte maneira. Primeiro, criamos um grafo aleatório $G_R(V, E_R)$ contendo os mesmos

[11] Devido ao grande número de clusters algumas cores se repetem, mas isso não possui nenhum significado especial.

nós de G. Depois, para cada sequência de n_u fotos $f_u^1, f_u^2, ..., f_u^{n_u}$ de cada usuário u, atribuímos PDIs aleatoriamente para cada foto e geramos as arestas aleatórias E_R de G_R a partir dessa nova atribuição. Assim, a sequência das fotos é aleatória, o que gera arestas e pesos de arestas aleatórios em G_R, mas preserva o número total de fotos que foi tirada em um determinado local. A ideia é simular trajetórias aleatórias na cidade. Desta maneira aleatória, a distribuição dos pesos das arestas segue uma distribuição normal $N_w(\mu_w, \sigma_w)$ com média μ_w e desvio padrão σ_w.

De acordo com $N_w(\mu_w, \sigma_w)$, quando a probabilidade de gerar $w(i, j)$ em $G_R(V, E_R)$ é próxima de zero, ou seja, quando for pouco provável que o grafo aleatório $G_R(V, E_R)$ tenha um peso de aresta $w(i, j)$, então a transição de $v_i \rightarrow v_j$ é uma transição popular na cidade sendo, de acordo com a nossa conjectura, uma transição entre pontos turísticos. Para o nosso conjunto de dados, o valor de t que fornece uma probabilidade próxima de 0 é $t = 10$. Como podemos observar na figura 11-c, os vértices (PDIs) do grafo resultante representam a maioria dos pontos turísticos de Belo Horizonte. As áreas dos PDIs resultantes cobrem sete de todos os oito pontos recomendados pelo TripAdvisor[12], sendo as áreas culturais e de lazer mais importantes de BH.

É interessante notar a diferença entre as figuras 11-b e 11-c, a primeira contendo todos os PDIs e a segunda somente os pontos turísticos da cidade de BH. Novamente, esta aplicação é interessante porque ela é capaz de identificar os PDIs em um contexto espaço-temporal, o que é fundamental, uma vez que os PDIs são dinâmicos e mudam ao longo do tempo.

6. Conclusão e Trabalhos Futuros

Neste trabalho apresentamos o que é, no melhor de nosso conhecimento, a primeira caracterização do Instagram. A análise do sistema foi feita tratando-o como uma rede de sensoriamento participativa. Assim, discutimos a cobertura espacial e temporal desta rede mostrando a sua abrangência planetária bem como uma frequência de compartilhamento de fotos espaço-temporal muito desigual e correlacionada com as rotinas de atividades humanas. Também discutimos uma aplicação que demonstra o potencial de uma RSP derivada do Instagram para o estudo da dinâmica de cidades.

Como trabalhos futuros, pretendemos analisar outras RSPs e desenvolver novas aplicações que exploram estas redes. Por exemplo, podemos imaginar aplicações que consideram conjuntamente dados provenientes de outros sistemas de sensoriamento participativo como o Waze (condições de tráfego) e o Weddar (condições meteorológicas), considerando inclusive, diferentes categorias/interesses das pessoas.

Agradecimentos

Este trabalho é parcialmente financiado pelo INCT-Web (MCT/CNPq 57.3871/2008-6), e pelas bolsas e projetos individuais dos autores financiados pelo CNPq, CAPES (bolsa 7356/12-9), e FAPEMIG.

Referências

Akyildiz, I., Su, W., Sankarasubramaniam, Y., and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer Networks*, 38(4):393 – 422.

Barth, F. (1969). *Ethnic groups and boundaries: the social organization of culture difference*. Scandinavian university books. Little, Brown.

Bilandzic, M. and Foth, M. (2012). A review of locative media, mobile and embodied spatial interaction. *International Journal of Human-Computer Studies*, 70(1):66–71.

[12]www.tripadvisor.com

Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., and Srivastava, M. B. (2006). Participatory sensing. In *Proc. Workshop on World-Sensor-Web (WSW)*.

Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011). Exploring Millions of Footprints in Location Sharing Services. In *Proc. 5th Int'l Conference on Weblogs and Social Media*.

Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proc. 17th ACM Int'l Conference on Knowledge Discovery and Data Mining*.

Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proc. 6th International Conference on Weblogs and Social Media*.

Daniells, K. (2012). Infographic: Instagram statistics 2012. *Digital Buzz Blog*.

Eisenman, S. B., Miluzzo, E., Lane, N. D., Peterson, R. A., Ahn, G.-S., and Campbell, A. T. (2010). Bikenet: A mobile sensing system for cyclist experience mapping. *ACM Transactions on Sensor Networks*, 6(1).

Fisk, P. R. (1961). The graduation of income distributions. *Econometrica*, 29(2):171–185.

Goodchild, M. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4):211–221.

Krumm, J. (2009). *Ubiquitous Computing Fundamentals*. Chapman & Hall/CRC, 1st ed.

Larson, E. C., Lee, T., Liu, S., Rosenfeld, M., and Patel, S. N. (2011). Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proc. 13th International Conference on Ubiquitous Computing*.

Malmgren, R. D., Stouffer, D. B., Motter, A. E., and Amaral, L. A. N. (2008). A poissonian explanation for heavy tails in e-mail communication. *Proc. National Academy of Sciences*, 105(47):18153–18158.

Mashhadi, A. J. and Capra, L. (2011). Quality Control for Real-time Ubiquitous Crowdsourcing. In *Proc. 2nd International Workshop on Ubiquitous Crowdsouring*.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011a). An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proc. of the Fifth Int'l Conf. on Weblogs and Social Media (ICWSM'11)*.

Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011b). Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proc. 5th Int. Conf. on Weblogs and Social Media*.

Rana, R. K., Chou, C. T., Kanhere, S. S., Bulusu, N., and Hu, W. (2010). Ear-phone: an end-to-end participatory urban noise mapping system. In *Proc. 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*.

Reddy, S., Estrin, D., Hansen, M., and Srivastava, M. (2010). Examining micro-payments for participatory sensing data collections. In *Proc. 12th ACM International Conference on Ubiquitous Computing*.

Saroiu, S. and Wolman, A. (2010). I am a sensor, and i approve this message. In *Proc. 11th Workshop on Mobile Computing Systems and Applications*.

Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. (2011). Socio-spatial Properties of Online Location-based Social Networks. In *Proc. 5th International Conference on Weblogs and Social Media*.

Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2012a). Uncovering Properties in Participatory Sensor Networks. In *Proc. 4th ACM International Workshop on Hot Topics in Planet-scale Measurement*.

Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., and Loureiro, A. A. F. (2012b). Visualizing the invisible image of cities. In *Proc. IEEE International Conference on Cyber, Physical and Social Computing*.

Sinnott, R. W. (1984). Virtues of the Haversine. *Sky and Telescope*, 68(2):159+.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 5(4).

Srivastava, M., Abdelzaher, T., and Szymanski, B. (2012). Human-centric sensing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):176–197.

Vasconcelos, M., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012). Caracterização e Influência do Uso de Tips e Dones no Foursquare. In *Proc. XXX SBRC*.

Vaz de Melo, P. O. S., Faloutsos, C., and Loureiro, A. A. (2011). Human dynamics in large communication networks. In *Proc. SDM*.

Weiser, M. (1991). The Computer in the 21st Century. *Scientific American*, 265(3):94–104.