MobDatU: A New Model for Human Mobility Prediction Based on Heterogeneous Data

Lucas Maia Silveira*, Jussara M. Almeida*, Humberto Marques-Neto† and Artur Ziviani‡
*Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG – Brazil
Email: {lucasmsil,jussara}@dcc.ufmg.br
†Pontifícia Universidade Católica de Minas Gerais (PUC MINAS)
Belo Horizonte, MG – Brazil
Email: humberto@pucminas.br
‡Laboratório Nacional de Computação Científica (LNCC)
Petrópolis, RJ – Brazil
Email:ziviani@lncc.br

Abstract

Several previous mobility models aim at describing or predicting human behavior in a particular region during a certain period of time. Nevertheless, most of those models have been evaluated using data from a single source, such as data from mobile calls or GPS data obtained from Web applications. Thus, the effectiveness of such models when using different types of data remains unknown. This paper proposes a new model to predict human mobility, called MobDatU, which was designed to use data from mobile calls and data from georeferenced applications (in an isolated or combined way). MobDatU as well as two state-of-the-art models, namely SMOOTH and Leap Graph, are evaluated considering various scenarios with single data source and multiple data sources. The experiments indicate that MobDatU always produces results that are better than or at least comparable to the best baseline in all scenarios, unlike the previous models whose performance is very dependent on the particular type of data used.

Resumo

Atualmente existem diversos modelos de mobilidade humana que visam descrever ou prever o comportamento humano em uma região durante um período de tempo. Entretanto, a maioria desses modelos foram avaliados utilizando dados de uma única fonte, por exemplo, dados de chamadas de telefonia móvel ou dados de GPS obtidos a partir de aplicações Web georreferenciadas. Portanto, a robustez destes modelos a diferentes tipos de dados é ainda desconhecida. Neste artigo, é proposto um novo modelo de previsão de mobilidade humana, chamado MobDatU, que foi projetado para utilizar tanto dados de telefonia móvel quanto dados de aplicativos georreferenciados (isolada e conjuntamente). O MobDatU, assim como dois modelos estado-da-arte, a saber SMOOTH e Leap Graph, são avaliados considerando diversos cenários com dados de fonte única e de múltiplas fontes. Os experimentos indicam que MobDatU atinge resultados melhores ou pelo menos comparáveis ao melhor modelo estado-da-arte em todos os cenários, diferentemente dos modelos alternativos cujo desempenho é bem mais sensível ao tipo de dado utilizado.

Keywords

Human mobility; georeferenced applications; mobile technology

I. Introdução

O entendimento sobre a mobilidade humana pode ajudar governos e empresas a prever e planejar ações para melhorar a qualidade de vida das pessoas de uma determinada região. Tais previsões e planejamentos de ações podem ser feitas a partir de modelos de previsão de mobilidade, que por sua vez podem ser desenvolvidos a partir da análise de dados sobre deslocamentos de pessoas. Esses dados podem ser provenientes de diferentes fontes, incluindo telefonia móvel [1] e aplicativos georreferenciados [2], como o Twitter.

Embora estudos recentes mostrem que a mobilidade humana em áreas urbanas possa ser bem previsível considerando rotinas diárias [3], a vasta maioria dos modelos de previsão de mobilidade disponíveis na literatura [4]–[7] foi proposta ou pelo menos avaliada considerando uma única fonte de dados, tais como dados de chamadas de telefonia móvel ou dados coletados de aplicações Web georreferenciadas (*e.g.*, Twitter). A robustez de tais modelos a dados de fontes diferentes é ainda desconhecida. Mais ainda, o uso combinado de dados de diferentes fontes pode levar a previsões mais precisas ou à cobertura de um maior volume da população. Entretanto, a adequação e a eficácia dos modelos existentes a dados de múltiplas fontes também ainda são desconhecidas.

Neste contexto, este trabalho apresenta um novo modelo de previsão de mobilidade humana, chamado MobDatU, que foi projetado para explorar dados de fontes heterogêneas, especificamente dados de telefonia móvel e dados de aplicações georreferenciadas. O MobDatU utiliza alguns princípios abordados em dois outros modelos considerados estado-da-arte, o Leap Graph [5] e o SMOOTH [4]. Enquanto o Leap Graph foi avaliado com dados de telefonia móvel, tendo conseguido um taxa de acerto de 84% no estudo original, o SMOOTH foi anteriormente avaliado somente com dados de aplicativos georreferenciados.

Neste artigo, nós comparamos a eficácia da previsão dos três modelos em vários cenários consistindo de: (i) dados homogêneos provenientes de fonte única, sejam logs de chamadas de telefonia móvel, sejam dados de GPS coletados da aplicação Twitter; e (ii) dados heterogêneos a partir da combinação de ambas as fontes. Os nossos resultados experimentais indicam que o novo modelo MobDatU alcança as maiores taxas de acerto em todos os cenários avaliados. Já o SMOOTH atinge resultados comparáveis ao do nosso modelo quando dados de GPS são utilizados, mas tem desempenho muito inferior nos outros cenários. O Leap Graph, por sua vez, obtém resultados similares ao do MobDatU para os cenários com dados de telefonia móvel (para os quais ele foi projetado) e também quando os dados das duas fontes são combinados. Entretanto, seu desempenho quando apenas dados de GPS são utilizados é inferior ao do SMOOTH e ao do MobDatU.

Portanto, as contribuições deste artigo são: (i) a avaliação de dois modelos de previsão de mobilidade estado-da-arte, SMOOTH e Leap Graph, em diferentes cenários com dados homogêneos e heterogêneos; e (ii) a proposição de um novo modelo, MobDatU, que produz resultados melhores ou pelo menos comparáveis ao melhor modelo tanto para os cenários de dados de telefonia móvel, quanto para os de Twitter.

O resto do artigo está organizado como segue. A Seção II discute a literatura sobre modelos de mobilidade humana. A Seção III formalmente apresenta a tarefa de previsão de mobilidade humana, explica o funcionamento dos dois modelos de referência e apresenta o novo modelo MobDatU. A metodologia de avaliação adotada, incluindo as coleções de dados utilizadas nos experimentos, é descrita na Seção IV, enquanto os resultados são discutidos na Seção V. Conclusões e trabalhos futuros são apresentados na Seção VI.

II. Trabalhos Relacionados

A literatura contém vários trabalhos que propõem modelos de previsão de mobilidade humana. Esses modelos utilizam diferentes tipos de dados e exploram diversas estratégias para tentar predizer com maior precisão a movimentação das pessoas (*i.e.*, usuários) [8]. Alguns destes modelos buscam prever as trajetórias das pessoas utilizando dados de GPS [9] ou dados de telefonia móvel [5]. Outros modelos realizam a previsão a partir de distribuições estatísticas que capturam padrões específicos identificados nos dados, tais como a distribuição da distância percorrida por um usuário [10], [11].

De forma similar, outros trabalhos, como [12] e [13], exploram padrões diversos observados nos dados, tais como os lugares visitados pelas pessoas, a frequência de visitação e a regularidade dos padrões de locomoção de um usuário. Em [2], os autores exploram os locais visitados bem como as distâncias de locomoção de um usuário dentro de uma mesma região para prever sua localização em um momento futuro. Já em [14], os autores abordam a previsão de mobilidade considerando tanto trajetórias longas quanto trajetórias curtas.

Utilizando dados de GPS coletados de aplicações de redes sociais georreferenciadas, os modelos propostos em [15] e [16] exploram não somente os padrões de distância percorrida e locais visitados, mas também as relações de amizade entre os usuários (inferidas a partir dos dados coletados). Ou seja, os modelos pressupõem que uma pessoa terá maior chance de ir para um local se um grande número de seus amigos frequentarem esse local.

A previsão da mobilidade humana pode auxiliar na prevenção de doenças e desastres bem como no planejamento urbano. Por exemplo, os trabalhos de [17], [18] mostram que o conhecimento sobre a mobilidade das pessoas em uma determinada região pode auxiliar na tomada de decisões para melhor e mais rapidamente evitar a disseminação de uma doença, amenizar os danos causados por desastres e até mesmo evitar engarrafamentos durante as horas de pico. Além disso, tal conhecimento também pode auxiliar as operadoras a aperfeiçoarem seus serviços de telefonia móvel [19].

Entretanto, os modelos de mobilidade humana disponíveis na literatura foram propostos ou avaliados considerando apenas uma fonte única de dados, sejam chamadas de telefonia móvel [1], [5], [18] sejam dados de GPS coletados de aplicações georreferenciadas [2], [4], [10], [11], [14]–[16]. A robustez desses modelos a dados de fontes diversas, seja isolada ou conjuntamente, é o foco deste trabalho. Para tanto são selecionados dois modelos de referência, o SMOOTH [4] e o Leap Graph [5]: enquanto o primeiro foi avaliado anteriormente somente a partir de dados de GPS, o segundo foi avaliado para dados de telefonia móvel. Mais ainda, é proposto um novo modelo que herda princípios dos dois modelos de referência, mas é projetado para explorar dados de fontes heterogêneas. Os dois modelos de referência, bem como o novo modelo proposto, chamado MobDatU, são descritos na próxima seção.

III. Modelos de Mobilidade

Esta seção primeiramente apresenta formalmente a tarefa de previsão de mobilidade humana abordada neste artigo (Seção III-A). A seguir, ela descreve brevemente o funcionamento dos dois modelos de referência, SMOOTH e Leap Graph (Seção III-B) e introduz o novo modelo de previsão proposto, o MobDatU (Seção III-C).

A. Previsão de Mobilidade Humana

A tarefa de previsão de mobilidade abordada neste artigo pode ser definida como segue. Sejam uma área alvo R, consistindo de um conjunto de regiões $r_i \in R$, um conjunto de usuários $u_i \in U$ e um intervalo de tempo $t \in \{0..T\}$ minutos. Sejam ainda um conjunto de treino $\mathcal{D}$ e um conjunto de teste $\mathcal{T}$, consistindo de tuplas $< u_i, r_i, t >$ que indicam que u_i estava na região r_i em t. Deseja-se construir um modelo para prever em que região r_i um dado usuário u_i estará em um dado momento t utilizando apenas os dados de treino $\mathcal{D}$. Deseja-se ainda avaliar o modelo utilizando os dados

de teste, ou seja, utilizar o modelo desenvolvido para prever a localização (*i.e.*, região) r_i para cada tupla $< u_i, ?, t >$ no conjunto de teste $\mathcal{T}$.

Note que a definição das regiões r_i pode ser feita de diversas maneiras. Por exemplo, cada região pode ser definida por um ponto central x_i, y_i (e.g, as coordenadas de uma estação rádio-base para o caso de dados de telefonia móvel) e um raio d. Alternativamente, cada região pode representar uma área quadrada com centro x_i, y_i e lado d (área total d^2). A definição das regiões adotada por cada modelo é explicitada nas seções seguintes. Note também que, neste trabalho, o intervalo de tempo total T é discretizado em janelas consecutivas Δt para fins de computação dos deslocamentos dos usuários. Em outras palavras, a localização de um usuário é predita considerando a granularidade de tempo de Δt minutos.

A divisão dos dados disponíveis em treino e teste também pode ser feita de forma diversa, dependendo da disponibilidade dos dados. Entretanto, tal divisão deve respeitar restrições temporais, ou seja, os dados de treino devem preceder os dados de teste (*i.e.*, $\forall <*, *, t_i> \in \mathcal{D}$ and $\forall <*,*,t_j> \in \mathcal{T}$, $t_i < t_j$). Mais ainda, considerando que os padrões de mobilidade humana variam ao longo do dia ou mesmo em dias diferentes (*e.g.*, dias de semana e fim de semana) [19], é desejado que os dados de treino tenham sido obtidos em períodos comparáveis com aqueles do conjunto de teste. Por exemplo, se deseja prever a localização de usuários entre 8 e 9 da manhã, deve-se utilizar dados coletados do mesmo período em dias anteriores, ou dados coletados em um período imediatamente anterior. A definição dos conjuntos de treino e teste em nossos experimentos será discutida na Seção IV-A.

B. Modelos de Referência

Esta seção descreve os principais componentes dos dois modelos de referência adotados neste trabalho, o SMOOTH [4] e o Leap Graph [5]. Em nossos experimentos, nós utilizamos as implementações de ambos os modelos disponibilizadas pelos autores.[1]

1) SMOOTH: O modelo SMOOTH, proposto por [4], captura a locomoção de um grupo de usuários U em uma área bidimensional simulada, composta por um conjunto de regiões de interesse. Cada região r_i é definida por coordenadas x_i, y_i dentro da área simulada e probabilidades p_i de um usuário se deslocar para cada uma delas. A probabilidade p_i portanto captura a popularidade da região r_i, isto é, o número esperado de pessoas que visitam r_i.

A ideia básica do SMOOTH é simular a locomoção dos usuários U utilizando cada probabilidade p_i de forma interativa, onde cada passo representa uma janela de tempo Δt minutos. Em cada passo, o modelo simula a locomoção de cada usuário $u_i \in U$ a partir de duas distribuições específicas extraídas do conjunto de treino: a distribuição das distâncias percorridas f_{dist} e a distribuição dos tempos de pausa f_{pausa}. Como em [4], observamos que tais distribuições seguem leis de potência para todos os cenários simulados. Tais distribuições são caracterizadas pelo parâmetro α e por valores mínimo (min) e máximo (max)[2].

Em cada passo, para cada usuário que não está em pausa (explicado posteriormente), primeiramente é computada a direção de deslocamento em função da sua localização atual e das probabilidades associadas a cada região $r_i \in R$. Em seguida, é selecionada aleatoriamente uma distância de deslocamento utilizando f_{dist}, e o deslocamento é simulado. Por fim, é escolhido aleatoriamente um tempo de pausa utilizando f_{pausa}. O usuário permanecerá neste local durante o intervalo de tempo f_{pausa} selecionado.

Conforme proposto em [4], a cada usuário é associado um raio de cobertura d que definirá a região na qual o usuário se encontra em cada passo de execução do modelo. Quando o usuário para em um ponto (x_i, y_i), é verificado se ele se encontra a uma distância d de pelo menos uma região (ou seja, se ele cobre uma região). O usuário é associado à região mais próxima coberta por ele. Caso ele não cubra nenhuma região, é introduzida uma nova região r_i em R definida pela sua localização atual. Ao final de cada passo, as probabilidades associadas a cada região (inclusive as novas) são recomputadas.

Logo, o treinamento do modelo consiste em extrair as distribuições f_{dist} e f_{pausa} bem como as regiões $r_i \in R$ do conjunto $\mathcal{D}$. Todas as regiões que aparecem no conjunto de treinamento são inicialmente introduzidas no conjunto R com as probabilidades correspondentes. Estas probabilidades são utilizadas para determinar as localizações iniciais dos usuários para a fase de simulação dos deslocamentos, durante a qual são aprendidas novas regiões visitadas e suas probabilidades.

Durante a fase de teste, os deslocamentos dos usuários do conjunto $\mathcal{T}$ são simulados utilizando o modelo aprendido, mantendo o conjunto R fixo e considerando a localização inicial de cada usuário dada pela sua primeira aparição em $\mathcal{T}$. As regiões visitadas pelos usuários durante a simulação são comparadas com os demais dados do conjunto $\mathcal{T}$ para avaliar a precisão do modelo.

2) Leap Graph: Em [5], os autores investigaram como utilizar dados de telefonia móvel para prever a mobilidade de usuários. Estes dados tipicamente correspondem a um conjunto de chamadas telefônicas. A cada chamada estão associados um identificador único do usuário, os instantes de início e fim da chamada, bem como as coordenadas

[1] toilers.mines.edu e www.cs.utexas.edu/ wdong86/.

[2] Distribuição Acumulada Complementar dada por $f(x) = \frac{(x*max^{\alpha} - x*min^{\alpha} - max^{\alpha})}{(x*max^{\alpha})*(x*min^{\alpha})}^{-\alpha}$.

(latitude e longitude) das antenas onde a chamada foi iniciada e finalizada e as identificações dos setores utilizados nestas antenas[3].

No modelo proposto em [5], chamado Leap Graph, cada região $r_i \in R$ corresponde a uma antena, sendo definida pelas suas coordenadas e por um raio d que representa sua área de cobertura. Assim, uma chamada do usuário u_i associada à antena correspondente à região r_i em um instante t indica a presença de u_i naquela região na janela de tempo que inclui t. O modelo tenta inferir os deslocamentos de cada usuário a partir de um grafo que captura a trajetória do usuário entre as regiões de R.

A fase de treinamento, portanto consiste primeiramente em criar um grafo de trajetórias para cada usuário a partir dos dados em $\mathcal{D}$. No grafo G_i criado para o usuário u_i, cada vértice corresponde a uma região. Uma aresta entre r_i e r_j é adicionada toda vez que: (i) u_i fez uma chamada que foi iniciada em r_i e finalizada em r_j; ou (ii) u_i fez duas chamadas consecutivas, a primeira em r_i e a segunda em r_j.

Como proposto originalmente, o Leap Graph objetiva prever a próxima região em que um usuário estará dado a sua localização atual. Portanto, ele não considera a dimensão tempo e explora apenas as transições entre antenas feitas por cada usuário. Para torná-lo comparável ao SMOOTH e ao MobDatU (que consideram o tempo) e aplicá-lo à tarefa de predição alvo, nós adicionamos laços (arestas da região r_i para ela mesma) para capturar os períodos entre chamadas consecutivas de um mesmo usuário, quando sua localização é desconhecida. Durante tais períodos, foi assumido que o usuário permaneceu a metade do tempo em uma região e a metade seguinte na outra. Ou seja, dadas duas chamadas consecutivas nas regiões r_i e r_j nos tempos t_1 e t_2 e considerando a discretização do tempo em janelas de duração Δt, foi computado o número w de janelas entre t_1 e t_2. Foram então adicionados laços em r_i e r_j, cada um com peso $w/2$.

Os grafos criados são então combinados em um grafo único ponderado G que representa os deslocamentos da população de usuários em $\mathcal{D}$. Para tal, os grafos de usuários G_i são ordenados pelo instante da primeira chamada de cada usuário em $\mathcal{D}$ e processados conforme esta ordem. As arestas de todos os grafos são combinadas em G, sendo que o peso de uma aresta corresponde ao número de grafos de usuários em que ela aparece. Porém, para trajetórias cobrindo n ou mais arestas que aparecem em múltiplos grafos, são considerados apenas a trajetória e os trechos que a sucedem no primeiro grafo processado. Em outras palavras, suponha que o grafo G_1 contenha a trajetória $\{r_1, r_2, r_3, r_4\}$ e que grafo G_2 contenha a trajetória $\{r_1, r_2, r_3, r_5\}$. Para $n = 2$, o grafo G conterá as arestas $\{r_1, r_2, r_3, r_4\}$, todas com peso 1, já que a trajetória $\{r_1, r_2, r_3\}$ ($n = 2$ arestas) aparece nos dois grafos. Conforme os autores, esta medida é tomada para evitar contabilizar duplamente as mesmas trajetórias. Ao final da combinação, os pesos de todas as arestas são normalizadas de forma que os pesos de todas as arestas de saída de cada vértice r_i totalizem 1.

A aplicação do modelo Leap Graph, durante a fase de teste, consiste em simular o grafo G produzido durante o treinamento como uma cadeia de Markov. Para cada usuário, a sua posição inicial é extraída da sua primeira chamada no conjunto $\mathcal{T}$ e que corresponde a um estado da cadeia. A cadeia é então simulada para inferir a posição do usuário em sucessivos passos. Em nossos experimentos, utilizamos $n = 2$ pois esta escolha levou aos melhores resultados em [5].

C. MobDatU: Um Novo Modelo de Predição de Mobilidade Humana

O novo modelo MobDatU tem como objetivo explorar dados de fontes heterogêneas para capturar a movimentação dos usuários entre as regiões de R. A área total simulada pelo modelo é divida em regiões quadrangulares não sobrepostas r_i (como em um *grid*). Cada região r_i é definida por um centro x_i, y_i e um lado d[4].

O MobDatU herda alguns aspectos dos dois modelos de referência. Por exemplo, assim como no SMOOTH, a cada região r_i é associada uma medida de popularidade, que, no MobDatU, representa o número de usuários que visitou r_i no conjunto de treino $\mathcal{D}$ (podendo ser 0). De forma similar ao Leap Graph (e diferentemente do SMOOTH), o MobDatU simula a movimentação dos usuários entre as regiões a partir de um grafo de transições. Entretanto, diferentemente do Leap Graph, a criação deste grafo não parte das trajetórias individuais de cada usuário e também não inclui o descarte de trajetórias com prefixo comum durante o processo de combinação dos grafos de usuários.

Ao invés disto, o MobDatU cria um grafo de transições onde o peso associado a uma aresta (r_i, r_j) representa o número de pessoas que fizeram a transição entre as duas regiões no conjunto $\mathcal{D}$. Este número pode ser inferido a partir tanto de dados de aplicativos georreferenciados quanto de dados de telefonia móvel. No segundo caso, assim como no Leap Graph, transições *self loop* são introduzidas para capturar os períodos entre chamadas sucessivas de um mesmo usuário.

A fase de treinamento do modelo consiste, portanto, em aprender o grafo de transições (incluindo os pesos das arestas) e as popularidades de cada região a partir do processamento do conjunto $\mathcal{D}$. Ao final, os pesos de todas as arestas são recomputados para capturar as popularidades de cada região destino. Em outras palavras, o peso da aresta (r_i, r_j) é multiplicado pela popularidade p_j. Os pesos das arestas saindo de cada vértice são então normalizados para refletir probabilidades de transição (ou seja, devem somar 1). Este é um ponto em que o MobDatU difere tanto do Leap Graph, que considera somente as probabilidades de transição, quanto do SMOOTH, que explora somente as popularidades de

[3]Cada antena é dividida em 3 setores de 120^o, cada um responsável por cerca de um terço da área de cobertura da antena.

[4]O modelo também foi executado utilizando regiões circulares de raio d, apresentando resultados similares aos reportados neste artigo.

Tabela I: Coleções de Dados Utilizadas

	Chamadas		*Tweets*		Intervalo de Tempo
	# Chamadas	# Usuários	#*Tweets*	#Usuários	
Fortaleza - 29/06/14	7185	2372	13453	4236	14h - 21h
Recife - 29/06/14	13335	4923	13577	3981	14h - 21h
Belo Horizonte - 02/03/13	15630	9354	14332	4870	12h - 19h
Belo Horizonte - 11/09/13	14023	4532	15635	5103	17h - 23h
Rio de Janeiro - 29/06/14	5120	1132	14033	3643	14h - 21h
Rio de Janeiro - 13/07/14	5340	1038	15860	4572	14h - 21h

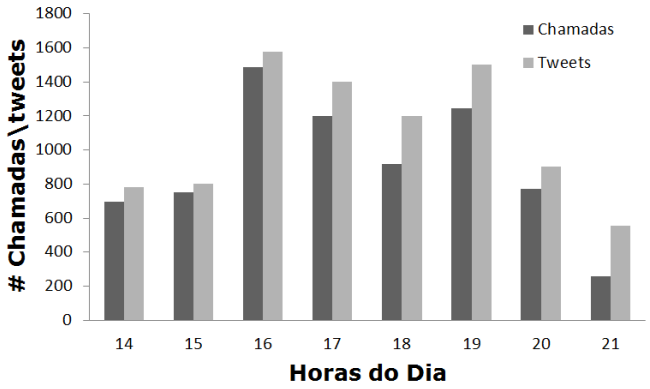

Figura 1: Quantidade de Chamadas/*Tweets* por Hora - Rio de Janeiro 29/06/14

cada região. O MobDatU considera que ambos aspectos podem influenciar a trajetória de um usuário: se por um lado os usuários tendem a visitar locais específicos dependendo da sua localização atual (como mostrado em [5]), por outro, a popularidade de uma região também influencia a movimentação dos usuários [2], [4].

Durante a fase de teste, é simulada uma cadeia de Markov utilizando as probabilidades de cada transição, como no Leap Graph, e considerando a primeira posição de cada usuário como sendo sua posição inicial em $\mathcal{T}$.

IV. Metodologia

Nessa seção serão apresentadas as coleções de dados que foram utilizadas pelos modelos (Seção IV-A) e a metodologia de avaliação adotada (Seção IV-B).

A. Coleções de Dados

Para avaliar os modelos de predição foram utilizados dados de telefonia móvel e dados coletados do Twitter. Os dados de chamadas foram fornecidos por uma grande companhia de telefonia móvel brasileira e correspondem às chamadas realizadas em algumas cidades brasileiras durante intervalos de tempo pré-especificados. Já os dados do Twitter consistem em *tweets* georreferenciados. Eles foram coletados pela *Stream API* da aplicação [20] utilizando o filtro *location* que restringe a área de coleta para uma determinada região. A coleta pela API é realizada em tempo real. Logo, após a companhia de telefonia móvel informar os dados que seriam coletados, a coleta de *tweets* foi planejada para os mesmos períodos de tempo e locais. Os dados coletados são: o identificador do usuário que realiza a chamada/*tweet* e a posição geográfica da antena (em caso de chamada) ou do *tweet*. Como as duas coletas são independentes, não é possível identificar a mesma pessoa em coleções diferentes. Logo, os usuários nas duas coletas são tratados como diferentes.

Os dados coletados foram filtrados para retirar os usuários que realizaram somente uma chamada ou publicaram somente um *tweet* em todo o período de tempo analisado. A Tabela I apresenta um sumário das coleções filtradas, informando local e período de cobertura, bem como números de chamadas, *tweets* e usuários. Note que o volume de *tweets* é similar em todas as coleções, enquanto o volume de chamadas é muito maior nas coleções de Recife e Belo Horizonte, sendo comparável ao volume de *tweets* coletados no mesmo período. Note também a variação na cobertura de usuários. Para Belo Horizonte (02/03/13), o número de usuários é duas vezes maior na coleção de chamadas. Já para as coleções do Rio de Janeiro, o número de usuários é muito maior na coleção de *tweets*. Estas diferenças podem ser explicadas por aspectos socioculturais e eventos específicos que ocorreram em cada cidade nos períodos monitorados.

As Figuras 1 e 2 mostram os números de chamadas, *tweets* e usuários ao longo do tempo nas coleções do Rio de Janeiro (29/06/2014). Esta figura ilustra que o volume de dados disponível varia bastante segundo o horário de coleta, o que é esperado também para os padrões de movimentação das pessoas. Este resultado motiva a aplicação de modelos de predição específicos para diferentes períodos, conforme descrito na próxima seção.

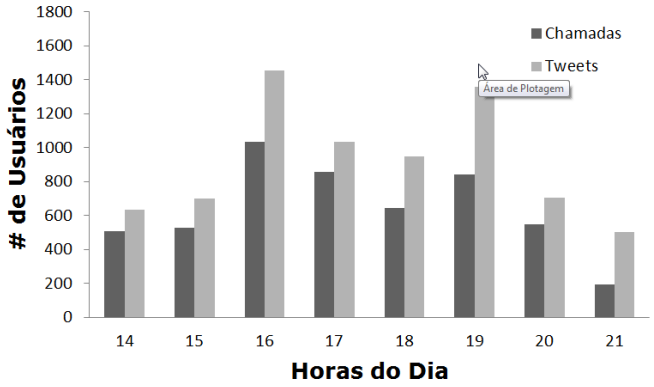

Figura 2: Quantidade de Usuários por Hora - Rio de Janeiro 29/06/14

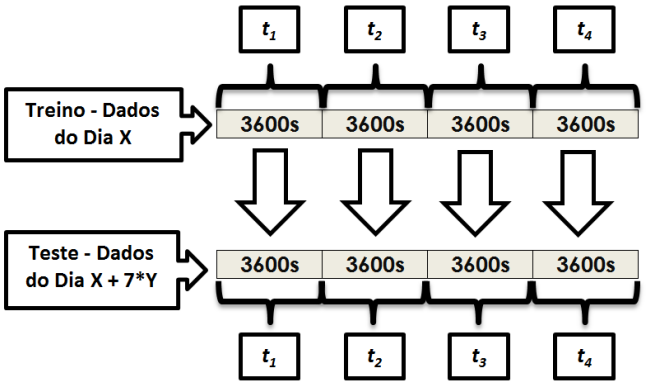

Figura 3: Separação dos dados em Treino e Teste - Utilizando Dois Dias

B. Avaliação

Cada modelo de previsão foi avaliado em termos da taxa de acerto[5] das previsões em janelas de tempo Δt iguais a 5 minutos. Para tal, cada coleção foi subdividida em intervalos de uma hora. Dadas as variações nos volumes de dados observadas (Figura 1), optou-se por desenvolver um modelo de predição para cada hora de forma a melhor capturar os padrões de locomoção em diferentes períodos. Em seguida, os dados de cada subcoleção foram divididos entre treino $\mathcal{D}$ e teste $\mathcal{T}$, conforme discutido na Seção III-A. Duas estratégias foram adotadas para fazer esta divisão, como descrito a seguir.

Para as coleções do Rio de Janeiro, como ambas cobrem o mesmo período no mesmo dia de semanas diferentes, a coleção do dia 29/06/14 foi utilizada como treino para aprender os modelos de cada hora. Estes modelos foram avaliados nos períodos correspondentes da base do dia 13/07/2014 (teste). Esta estratégia é mostrada na Figura 3. Para as demais coleções, como não temos acesso a dados de múltiplos dias cobrindo o mesmo período, foi adotada uma estratégia diferente. Cada hora no intervalo de tempo coberto pela coleção foi dividida em dois períodos de meia hora. O primeiro período foi utilizado para aprender o modelo (treino), e o segundo para avaliá-lo (teste). Esta estratégia é ilustrada na Figura 4.

Para cada coleção, cada modelo de mobilidade foi avaliado usando, como conjuntos de treino e teste, somente as chamadas, somente os *tweets* e tanto chamadas quanto *tweets*. O objetivo do último cenário é avaliar o desempenho dos modelos quando configurados com dados heterogêneos. A melhor forma de combinar os dados das duas fontes não é óbvia já que modelos diferentes podem ter desempenhos relativos diferentes, dependendo do dado de entrada. Assim, nós consideramos e avaliamos duas estratégias:

- Associação de *tweets* a chamadas: cada *tweet* é associado à antena mais próxima de sua localização.
- Associação de chamadas a *tweets*: cada antena da coleção de chamadas é considerado um ponto na região de simulação.

Um aspecto importante da avaliação é a definição das regiões R de cada modelo. Para as chamadas, o Leap Graph e o SMOOTH consideram a localização de cada antena que aparece no conjunto $\mathcal{D}$ como uma região r_i. Porém no caso do SMOOTH, novas regiões podem surgir durante o treino, de acordo com a movimentação dos usuários. Para os *tweets*,

[5]A taxa de acerto é calculada como a fração de tuplas $< u_i, r_i, t >$ para as quais a previsão foi correta durante a fase de teste

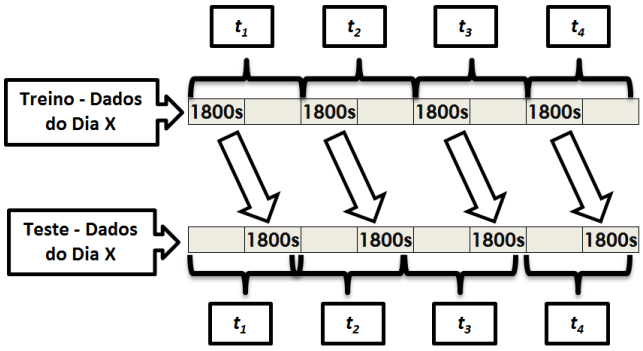

Figura 4: Separação dos dados em Treino e Teste - Utilizando Um Dia

o SMOOTH considera inicialmente cada localização distinta associada a um *tweet* em $\mathcal{D}$ como uma região[6], e novas regiões podem surgir durante o treino. Para o Leap Graph, os *tweets* foram agrupados em regiões circulares (com raio d) considerando suas localizações, e estas regiões foram inseridas em R. Para os dados combinados, foram adotadas as mesmas abordagens dependendo da estratégia de combinação feita. Já o MobDatU sempre divide a área total de cada cidade em regiões quadrangulares não sobrepostas (Seção III-C), independentemente do tipo de dado.

Em todos os casos, consideramos a distância d que define cada região como 500 metros. Este valor foi escolhido para que os resultados dos vários cenários pudessem ser comparados[7], uma vez que, para os dados de telefonia móvel, a posição geográfica informada tem granularidade dada pelo raio de cobertura de uma antena, a saber 500 metros .

V. Experimentos

Nossos resultados são sumarizados na Tabela II que apresenta, para cada cenário e para cada coleção de dados, os valores de taxa de acerto mínimo, máximo e médio considerando todos os intervalos de tempo $\Delta t = 5$ minutos em todas as horas do período de cada coleção. No geral, nota-se que cada modelo de referência funciona melhor se configurado para as fontes de dados homogêneas no qual ele foi avaliado anteriormente: o Leap Graph consegue taxa de acertos melhores quando utiliza somente chamadas, e o SMOOTH consegue resultados melhores quando utiliza somente *tweets*.

Por exemplo, considerando a coleção de dados do Rio de Janeiro, nota-se que, quando os dados de chamadas são usados como entrada, o SMOOTH tem um desempenho médio degradado de 53% em relação ao Leap Graph. Já quando são usados os *tweets*, o Leap Graph possui um desempenho de 14% pior que o SMOOTH. Quanto aos cenários com dados heterogêneos, a taxa de acerto de cada modelo foi um pouco menor em relação ao seu melhor cenário. Ou seja, os modelos de referência se comportam um pouco pior em cenários com dados heterogêneos.

Em contrapartida, o MobDatU tem um desempenho comparável aos melhores resultados dos modelos de referência em todos os cenários. Para o cenário de chamadas, ele tem uma taxa de acerto ligeiramente maior que a do Leap Graph e muito superior (109%) à do SMOOTH. Já para o cenário de *tweets*, os resultados do Leap Graph são piores em 19% quando comparados aos do MobDatU, enquanto o SMOOTH consegue resultados bem próximos aos do nosso modelo. Comparando os dois cenários com dados homogêneos, nota-se que o MobDatU tem uma taxa de acerto bem maior no cenário de chamadas. A razão para este resultado é um maior número de regiões distintas presentes somente no conjunto de teste para os dados de *tweets*. Regiões que não aparecem no conjunto de treino terão popularidade e taxas de transição nulas no modelo e nunca serão previstas. Logo, regiões que só aparecem no teste necessariamente levam a previsões erradas do modelo[8].

Em relação ao cenário com dados heterogêneos, assim como os modelos de referência, o MobDatU apresenta um desempenho um pouco pior em relação aos cenários de dados homogêneos, mas ainda superior, em média, ao dos modelos de referência. Quanto às estratégias de combinação de dados, observa-se que cada modelo de referência teve melhor desempenho quando a estratégia adotada faz a associação tendo como alvo o tipo de dado para o qual o modelo tem melhor precisão (chamadas para o Leap Graph e para o MobDatU, *tweets* para o SMOOTH).

Os resultados obtidos para as outras coleções de dados apresentam um comportamento similar aos dados do Rio de Janeiro. Uma exceção ocorre na coleção de Belo Horizonte (11/09/2013), na qual o MobDatU teve resultados melhores nos cenários com dados heterogêneos. Vale ressaltar que o uso de dados heterogêneos viabiliza uma realização de

[6]Como descrito na Seção III-B1, cada região é definida no SMOOTH por um ponto, e a associação de um usuário a uma região é feita verificando um raio de cobertura d a partir do *usuário*.

[7]Note que as áreas das regiões do SMOOTH e do Leap Graph (πd^2) são maiores que as áreas das regiões definidas pelo MobDatU (d^2). Logo a nossa avaliação favorece os modelos de referência, uma vez que as predições são mais difíceis para regiões com áreas menores.

[8]Note que este maior número de regiões novas no teste foi observada em todas coleções, a despeito das diferenças de volumes de dados de chamadas e *tweets* discutidos na Seção IV-A.

Tabela II: Taxa de Acerto dos Modelos de Previsão de Mobilidade

Rio de Janeiro								
Modelo	Chamadas				Tweets			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	18%	55%	36,75%	0,13	34%	73%	52,12%	0,138
Leap Graph	68%	88%	78,13%	0,072	21%	60%	44,75%	0,129
MobDatU	70%	87%	79,88%	0,063	37%	77%	55%	0,143
Modelo	*Tweets* a Chamadas				Chamadas a *Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	45%	52%	48%	0,26	38%	77%	51,5%	0,13
Leap Graph	65%	84%	73,75%	0,062	25%	58%	41,63%	0,12
MobDatU	67%	85%	74,38%	0,059	40%	79%	53,12%	0,129
Belo Horizonte 02/03/2013								
Modelo	Chamadas				*Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	20%	48%	35,37%	0,104	34%	70%	51,25%	0,125
Leap Graph	68%	80%	73%	0,044	21%	50%	40,75%	0,097
MobDatU	67%	82%	75,12%	0,047	37%	77%	54,5%	0,144
Modelo	*Tweets* a Chamadas				Chamadas a *Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	20%	35%	28,37%	0,053	38%	73%	50,85%	0,117
Leap Graph	62%	77%	67,75%	0,054	30%	50%	36,75%	0,07
MobDatU	58%	78%	66,87%	0,071	37%	77%	52,37%	0,131
Belo Horizonte 11/09/2013								
Modelo	Chamadas				*Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	18%	40%	27,25%	0,08	38%	77%	53,25%	0,130
Leap Graph	67%	75%	72,25%	0,025	22%	50%	36,5%	0,092
MobDatU	70%	81%	73,75%	0,043	43%	78%	55,13%	0,128
Modelo	*Tweets* a Chamadas				Chamadas a *Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	18%	32%	25,75%	0,051	42%	78%	57,62	0,138
Leap Graph	67%	86%	76,13%	0,063	22%	67%	41,75%	0,152
MobDatU	67%	87%	75%	0,067	42%	80%	57,87	0,14
Recife								
Modelo	Chamadas				*Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	20%	35%	27,25%	0,053	42%	78%	43,12	0,139
Leap Graph	67%	86%	75,75	0,062	25%	67%	43,12%	0,139
MobDatU	67%	87%	75%	0,067	42%	80%	57,87%	0,140
Modelo	*Tweets* a Chamadas				Chamadas a *Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	19%	34%	26,13%	0,061	42%	77%	53,38%	0,121
Leap Graph	69%	75%	72,12%	0,023	22%	45%	35,13%	0,075
MobDatU	70%	77%	73%	0,031	43%	78%	54,63%	0,124
Fortaleza								
Modelo	Chamadas				*Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	18%	35%	26,25%	0,062	40%	77%	56,63%	0,138
Leap Graph	65%	84%	73,75%	0,063	25%	67%	42,75%	0,140
MobDatU	67%	85%	74,37%	0,059	42%	80%	57,87%	0,14
Modelo	*Tweets* a Chamadas				Chamadas a *Tweets*			
	Mínimo	Máximo	Média	Desvio Padrão	Mínimo	Máximo	Média	Desvio Padrão
SMOOTH	22%	37%	28,75%	0,053	42%	77%	52,12%	0,118
Leap Graph	62%	75%	66,87%	0,039	30%	50%	37,63%	0,078
MobDatU	56%	78%	67,37%	0,072	38%	79%	52,25%	0,129

previsões para um número de usuários (isto é cobertura de usuários) potencialmente muito maior[9], o que pode compensar eventuais perdas nas taxas de acerto quando comparadas às obtidas com dados homogêneos.

As Figuras 5, 6 e 7 mostram as taxas de acerto de cada modelo ao longo do período coberto pela coleção do Rio

[9]Já que as coleções de chamadas e *tweets* não apresentam nenhum parâmetro que indicasse a existência de um mesmo usuário em ambas as fontes, foi considerado que os usuários dos dados de chamadas eram diferentes dos de *tweets* conforme descrito na Seção IV-A.

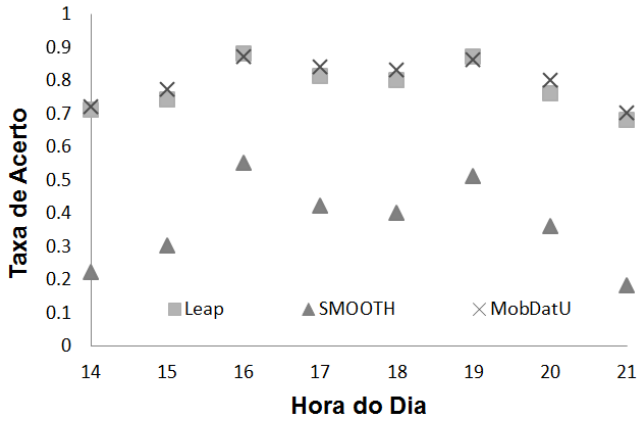

Figura 5: Taxa de Acerto para Cenários 1, 2 e 3 de Chamadas - Rio de Janeiro

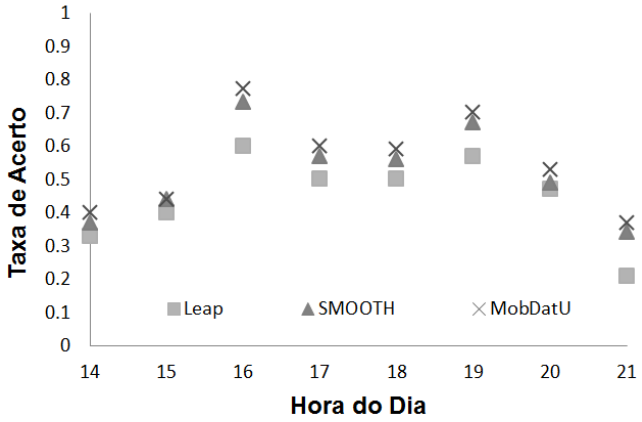

Figura 6: Taxa de Acerto para Cenários 1, 2 e 3 de *Tweets*- Rio de Janeiro

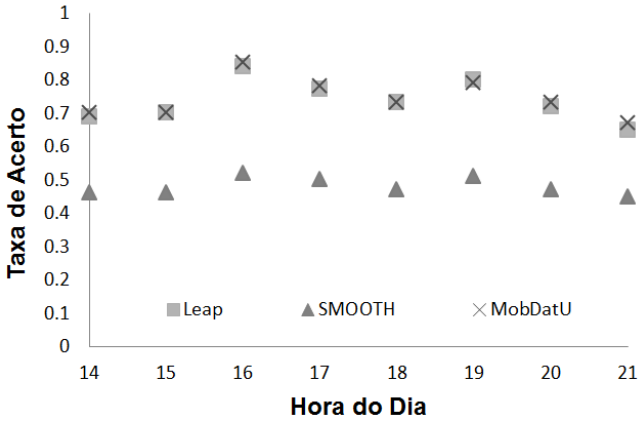

Figura 7: Taxa de Acerto para Cenários 1, 2 e 3 de *Tweets* para Chamadas - Rio de Janeiro

de Janeiro para os cenários com dados homogêneos e para um cenário com dado heterogêneo. Comparando esta figura com a Figura 1, nota-se que as maiores taxas de acertos foram obtidas em períodos com maior o volume de dados, o que era esperado. A figura também mostra que os ganhos do MobDatU sobre o melhor modelo de referência podem ser maiores que os mostrados na Tabela II. Por exemplo, na Figura 6, para o intervalo de 16 horas, a taxa de acerto do MobDatU supera em 6% a do SMOOTH (melhor modelo de referência), enquanto na Figura 5, o MobDatU supera o Leap Graph em 5% para o período de 20 horas. Resultados semelhantes foram observados também para as outras coleções.

Em suma, o MobDatU mostrou se adequar bem a dados distintos, conseguindo desempenho comparável (e por vezes superior) ao do melhor modelo de referência, em todos os cenários com dados homogêneos e heterogêneos. Em

contrapartida, tanto SMOOTH quanto Leap Graph podem ter grande perda de desempenho, dependendo do tipo de dado de entrada. A superioridade do nosso modelo pode ser explicada por ele considerar tanto a popularidade das regiões quanto a frequência de transições entre as regiões. Ambos os aspectos são fatores importantes que influenciam como as pessoas se movimentam em uma cidade. Além disto, como estes aspectos podem ser capturados tanto pelas chamadas telefônicas realizadas quanto pelos *tweets* compartilhados, o MobDatU mostrou um bom funcionamento para ambos os tipos de dados.

VI. Conclusão e Trabalhos Futuros

Neste artigo, foi proposto um novo modelo de previsão de mobilidade humana, o MobDatU. O modelo proposto, assim como dois modelos estado-da-arte, foram avaliados em diversos cenários com dados reais homogêneos e heterogêneos. Os experimentos indicaram que nenhum dos dois modelos de referência é superior em todos os cenários investigados, o que demonstra a sensibilidade dos mesmos aos dados disponíveis. Já o MobDatU se adequou bem a ambos tipos de dados, sendo pelo menos comparável (e por vezes superior) ao melhor modelo de referência em todos os cenários. Também foi mostrado que o volume de dados, tanto de chamadas telefônicas quanto de *tweets*, varia bastante durante o dia, o que afeta a precisão de todos os modelos. Logo, a diferença do volume de dados por período de tempo é algo importante a se considerar quando for realizar a previsão da mobilidade humana.

Como trabalho futuro pretende-se investigar novas estratégias de combinação de dados heterogêneos que permitam a identificação (ou pelo menos a inferência da presença) de um mesmo usuário em múltiplas coleções. Este é um grande desafio técnico uma vez que essas coleções tipicamente são obtidas de forma independente. Pretende-se também adaptar o MobDatU para prever o volume de pessoas que estarão em uma determinada região em um certo período. Tais previsões são úteis para suportar decisões de gerenciamento e planejamento urbano.

Agradecimentos

Os autores agradecem o apoio do INCTWeb (MCT/CNPq 573871/2008-6), CNPq, CAPES, FAPEMIG e FAPERJ.

Referências

[1] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, "Uncovering individual and collective human dynamics from mobile phone records," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 22, p. 224015, 2008. [Online]. Available: http://stacks.iop.org/1751-8121/41/i=22/a=224015

[2] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012.

[3] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010. [Online]. Available: http://www.sciencemag.org/content/327/5968/1018.abstract

[4] A. Munjal, T. Camp, and W. C. Navidi, "Smooth: A simple way to model human mobility," in *Proc. 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2011.

[5] W. Dong, N. Duffield, Z. Ge, S. Lee, and J. Pang, "Modeling cellular user mobility using a leap graph," in *Proc. 14th International Conference on Passive and Active Measurement*, 2013.

[6] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: Self-similar least-action human walk," *IEEE/ACM Transactions on Networking*, vol. 20, no. 2, pp. 515–529, April 2012.

[7] M. Allamanis, S. Scellato, and C. Mascolo, "Evolution of a location-based online social network: Analysis and models," in *Proceedings ACM Conference on Internet Measurement*, 2012.

[8] N. Bui, N. Bui, F. Michelinakis, F. Michelinakis, and J. Widmer, "A model for throughput prediction for mobile users," in *Proceedings 20th European Wireless Conference;*, May 2014.

[9] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated gps traces," *ACM Transactions on Intelligent Systems Technology*, vol. 2, no. 1, 2011.

[10] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: A new mobility model for human walks," in *Proc. INFOCOM*, 2009.

[11] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the levy-walk nature of human mobility," in *Proc. INFOCOM*, 2008.

[12] H. Jeung, Q. Liu, H. T. Shen, and X. Zhou, "A hybrid prediction model for moving objects," in *Proc. IEEE 24th International Conference on Data Engineering*, 2008.

[13] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "A random walk around the city: New venue recommendation in location-based social networks," in *Proc. International Conference on Social Computing*, 2012.

[14] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM International Conference on Knowledge Discovery and Data Mining*, 2011.

[15] M. Musolesi and C. Mascolo, "Designing mobility models based on social network theory," *SIGMOBILE Mobile Computing Communications Review*, vol. 11, no. 3, 2007.

[16] T. Nguyen and B. K. Szymanski, "Using location-based social networks to validate human mobility and relationships models," in *Proc. International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[17] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature Publishing Group*, vol. 453, 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06958

[18] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 484–21 489, 2009. [Online]. Available: http://www.pnas.org/content/106/51/21484.abstract

[19] F. H. Z. Xavier, L. M. Silveira, J. M. d. Almeida, A. Ziviani, C. H. S. Malab, and H. T. Marques-Neto, "Analyzing the workload dynamics of a mobile phone network in large scale events," in *Proc. First Workshop on Urban Networking*, 2012.

[20] Twitter, "Twitter streaming api (dev.twitter.com/streaming/overview)," 2013. [Online]. Available: https://dev.twitter.com/docs/streaming-apis