Usando Redes Aleatórias na Análise de Mobilidade

Pedro O.S. Vaz de Melo[1,2], Aline C. Viana[2], Marco Fiore[2,3], Katia Jaffrès-Runser[4], Frédéric Le Mouël[3], Antonio A.F. Loureiro[1]

[1]Universidade Federal de Minas Gerais, Brasil

[2]INRIA, França

[3]INSA Lyon, França

[4]University of Toulouse, França

{olmo,loureiro}@dcc.ufmg.br

aline.viana@inria.fr

{marco.fiore,frederic.le-mouel}@insa-lyon.fr

kjr@n7.fr

Abstract. *The constant advancement of information systems has allowed more data to be generated and stored from the most diverse situations. It is fascinating that, behind these records, we see the reflection of the environment itself, since every record represents a decision made by some entity. In this work, we modeled real-world scenarios of mobility from using temporal complex networks. The analysis assumes that these systems are composed of entities able to interact in a rational manner, reflecting their interests and activity dynamic. In this direction, we propose a technique for analyzing mobility scenarios from random graphs. This technique examines how the real system would evolve if the agents' decisions were random, and from there, you can check, for example, which edges are random and which are derived from social relationships, such as friendship or professional.*

Resumo. *O avanço constante de sistemas de informação tem permitido que mais dados sejam gerados e armazenados a partir das mais diversas situações. É fascinante que, por trás de cada registro, seja possível ver o reflexo do ambiente em si, ou seja, alguma decisão tomada por alguma entidade. Neste trabalho, são estudados cenários reais de mobilidade a partir de uma modelagem usando redes complexas temporais. A análise parte do pressuposto que esses sistemas são compostos de entidades capazes de interagir entre si de uma maneira racional, refletindo seus interesses e dinâmica de atividade. Nessa direção, é proposta uma técnica para analisar cenários de mobilidade a partir de grafos aleatórios. Essa técnica verifica como o sistema real evoluiria caso as decisões dos seus agentes fossem aleatórias e, a partir dela, pode-se verificar, por exemplo, quais arestas são aleatórias e quais são provenientes de relações sociais, tais como relações de amizade ou profissionais.*

1. Introdução

Trabalhos recentes sobre redes móveis sem fio têm analisado dados de mobilidade de indivíduos a partir de medições reais em cidades, como táxis em São Fran-

cisco (EUA), ou grandes campus universitários, tais como USC, MIT, Dartmouth e UF. Em [Thakur et al. 2010], os autores observaram distribuições de similaridade das populações de usuários móveis dessas bases de dados e verificaram que os atuais modelos de mobilidade falham em capturar o comportamento dos indivíduos de forma precisa. Numerosos protocolos baseados em aspectos sociais da mobilidade também têm lidado com o problema de previsão de futuros contatos em redes intermitentes. Essas obras são baseadas no fato de que o comportamento humano tende a ter um padrão espaço-temporal, repetindo locais e horários do dia e mostrando algum grau de regularidade [Gonzalez et al. 2008]. Assim, pode-se avaliar, por exemplo, o potencial dos nós da rede para atuarem como disseminadores ou roteadores de mensagens a partir de métricas de redes complexas, tal como a centralidade [E. M. Daly and M. Haahr 2009, Barbera et al. 2011, Kochem Vendramin et al. 2011].

O foco deste trabalho é estudar cenários reais de mobilidade a partir de uma modelagem usando redes complexas temporais. A análise parte do pressuposto que esses sistemas são compostos de entidades capazes de interagir entre si de uma maneira racional, refletindo seus interesses e dinâmica de atividade. Chamamos esses sistemas de *Redes Complexas baseadas em Decisão* (RCBDs), essas entidades de nós (ou agentes) e as interações entre os nós, de arestas. A principal característica de uma RCBD é que ela evolui de acordo com as decisões tomadas pelos agentes desses sistemas, as quais são guiadas principalmente por suas motivações pessoais [Backstrom et al. 2006a, Kumar et al. 2006, Leskovec et al. 2008, Mislove et al. 2007]. Além disso, RCBDs são caracterizadas por terem um grande número de vértices e arestas que apresentam um padrão como, por exemplo, comunidades ou vértices altamente conectados, chamados *hubs* [Albert et al. 1999]. Enquanto em uma rede simples, com no máximo centenas de nós, o olho humano é um instrumento de poder considerável, em uma rede complexa, esta abordagem é inútil. Assim, para estudar, analisar e caracterizar as redes complexas, métodos estatísticos e algoritmos eficientes são necessários.

Este trabalho propõe uma técnica para analisar cenários de mobilidade a partir de grafos aleatórios. Resumidamente essa técnica verifica como o sistema real evoluiria caso as decisões dos seus agentes fossem aleatórias. Mais especificamente, usa-se um algoritmo que, a partir de uma rede real, constrói uma versão aleatória contendo o mesmo número de nós, arestas e distribuição empírica do grau dos nós. Esse algoritmo é estendido para grafos temporais e, a partir dele, pode-se verificar, por exemplo, quais arestas são aleatórias e quais são provenientes de relações sociais, tais como relações de amizade ou profissionais. Essa classificação pode ser usada em diferentes aplicações. Por exemplo, o administrador de uma empresa de dispositivos móveis pode usar este método para detectar as relações sociais de seus clientes e, portanto, usar esse conhecimento para disseminar dados entre os clientes sem a utilização da infraestrutura da rede [Daly and Haahr 2007, Hossmann et al. 2009, Katsaros et al. 2010]. Ao saber como classificar um encontro como aleatório ou social, reduz-se o grau máximo de um nó a um número viável e escalável, reduzindo o espaço de armazenamento.

Este trabalho está organizado da seguinte maneira. As seção 2 descreve os trabalhos relacionados. A seção 3 apresenta como é feita a modelagem de cenários de mobilidade em redes complexas temporais e também como se constrói a rede aleatória correspondente. A seção 4 discute três possíveis aplicações para a modelagem proposta neste

trabalho: classificação de relacionamentos, predição de futuros encontros e disseminação de dados. Finalmente, a seção 5 apresentada as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

A análise dos aspectos sociais de sistemas complexos é fundamental para a compreensão das motivações por trás das ações de suas entidades. Ao revelar as razões por trás das decisões, é possível separar os eventos aleatórios daqueles movidos por relações sociais. Em [Crandall et al. 2008], por exemplo, foi mostrado que a ocorrência de arestas na rede está relacionada com as semelhanças entre os nós. Em outro exemplo interessante, [Backstrom et al. 2006b] mostra que a probabilidade de um indivíduo a entrar em uma comunidade é influenciada não só pelo número de amigos que ele tem na comunidade, mas principalmente pela forma como os amigos estão conectados entre si.

Os laços sociais entre os usuários têm sido amplamente explorado em redes móveis oportunistas de modo a favorecer os serviços de rede. Os problemas considerados vão desde o encaminhamento de mensagens *multi-hop* [Hui et al. 2008, Mtibaa et al. 2010, Mary Schurgot and Cristina Comaniciu and Katia Jaffrès-Runser 2012] e *multicasting*, a segurança da rede [Conti et al. 2010]. Em [Hui et al. 2008, Mtibaa et al. 2010], por exemplo, os autores utilizam métricas sociais derivadas de comunicações entre os usuários para melhorar a eficiência das decisões de encaminhamento oportunistas e limitando, simultaneamente, a sobrecarga da comunicação. Todas as obras acima tentar explorar a regularidade e os repetidos padrões espaço-temporais esperado no comportamento humano e tendem a ignorar as ligações aleatórias ou não-sociais entre os usuários móveis. Em vez disso, focamos na análise simultânea das relações aleatórias e social entre os indivíduos.

Até o alcance do nosso conhecimento, [Miklas et al. 2007] e [Zyba et al. 2011] são os trabalhos mais estreitamente relacionados com o apresentado aqui. [Zyba et al. 2011] distingue usuários sociais e itinerantes de acordo com seu comportamento de mobilidade: regularidade ou duração das visitas em uma determinada área. [Miklas et al. 2007] classifica as arestas entre amigos e desconhecidos, sendo que encontros frequentes entre pares de nós caracterizam interações de amizade. Os autores definiram empiricamente que indivíduos que se encontraram 10 dias ou mais dos 101 dias são amigos, outros são estranhos. Neste trabalho usamos redes aleatórias para definir o limite que separa relações sociais de aleatórias.

3. Modelagem

3.1. Fundamentos

Dentre as principais características verificadas em redes sociais reais, destaca-se a presença de comunidades, que são grupos de indivíduos fortemente conectados entre si porque compartilham os mesmos interesses ou dinâmicas de atividades [Backstrom et al. 2006a, Kumar et al. 2006]. A formação dessas comunidades só é possível porque a criação das arestas reflete as decisões sociais dos indivíduos, que geralmente são regulares e tendem a se repetir. Por outro lado, em uma rede aleatória as arestas são criadas independentemente dos atributos de cada nó, ou seja, um nó i tem a mesma probabilidade p de se conectar a qualquer outro nó j da rede.

Assim, uma métrica interessante para diferenciar uma rede aleatória de uma rede social é o coeficiente de aglomeração da rede. O coeficiente de aglomeração cc_i de um nó i caracteriza a densidade de conexões próximas a i. Mais especificamente, ele mede a probabilidade de dois vizinhos do nó i serem conectados entre si [Newman 2003]. Ele é calculado dividindo o número total de arestas que existe entre os vizinhos do nó i pelo número de arestas possível entre eles. O coeficiente de agrupamento da rede é a média $cc_i, \forall i \in V$. Por causa da existência de comunidades, as redes sociais têm um coeficiente de aglomeração alto, enquanto uma rede aleatória tem um coeficiente de aglomeração baixo, uma vez que as arestas são uniformemente distribuídas na rede [Watts and Strogatz 1998]. No restante deste trabalho, esse conceito é usado para diferenciar redes aleatórias de redes sociais.

3.2. Redes Complexas baseadas em Decisão

A principal característica das RCBDs é que as interações entre as suas entidades são, geralmente, consequência de decisões semi-racionais. Escreve-se "normalmente" e decisões "semi-racionais" porque qualquer sistema está sujeito a eventos aleatórios e escolhas irracionais. No entanto, uma vez que a maioria das interações ainda decorrem de decisões conscientes feitas por suas entidades, a evolução de RCBDs é significativamente diferente da evolução de redes aleatórias como, por exemplo, redes de Erdös and Rényi [Erdös and Rényi 1960]. Assim, enquanto em uma RCBD as arestas são geradas a partir de decisões semi-racionais, que tendem a ser regulares e a se repetir, em uma rede aleatória as arestas são geradas independentemente dos atributos dos nós, ou seja, a probabilidade de dois nós se conectarem é constante.

Considere, por exemplo, uma rede de pessoas e suas rotinas de mobilidade. Uma interação entre duas pessoas ocorre se elas se encontram. Se Silva e Moreira trabalham no mesmo escritório e suas horas de trabalho são de 8:00 às 18:00, pode-se prever facilmente que uma interação entre Silva e Moreira irá ocorrer em torno de 8:00 durante os dias da semana. Entretanto, se o proprietário da empresa decide alterar o horário de trabalho para de 12:00 às 20:00, então é quase certo que essa interação também se moverá diariamente para 12:00. Tudo isso é baseado no fato de que acreditamos fortemente que tanto Silva quanto Moreira decidirão ir ao trabalho todos os dias pontualmente, já que este é, provavelmente, a decisão racional para se tomar.

Por outro lado, a maioria dos cenários está sujeita a eventos aleatórios que podem desviar o comportamento esperado dos agentes. No exemplo anterior, Silva pode esquecer da mudança de horário da empresa e chegar ao trabalho no mesmo horário como antes, ou seja, às 8:00. Além disso, ele poderia ficar preso no trânsito e se atrasar. O fato é que, apesar de decisões racionais serem regulares, as decisões aleatórias podem muitas vezes ocorrer também.

Formalmente, um agente pode executar uma decisão **social** D_s ou uma decisão **aleatória** D_a. Ele tem uma probabilidade p_s da execução de uma decisão social D_s e uma probabilidade $p_r = 1 - p_s$ da execução de uma decisão aleatória D_a. Intuitivamente, quando $p_s \gg p_r$, a rede normalmente evolui para uma rede social bem estruturada, com a presença de todas as características comuns de uma rede social vistas na seção 2, tais como a presença de comunidades e *hubs*. Por outro lado, quando $p_r \gg p_s$, a rede normalmente desenvolve características de uma rede aleatória, como as redes de Erdös e Rényi [Erdös and Rényi 1960].

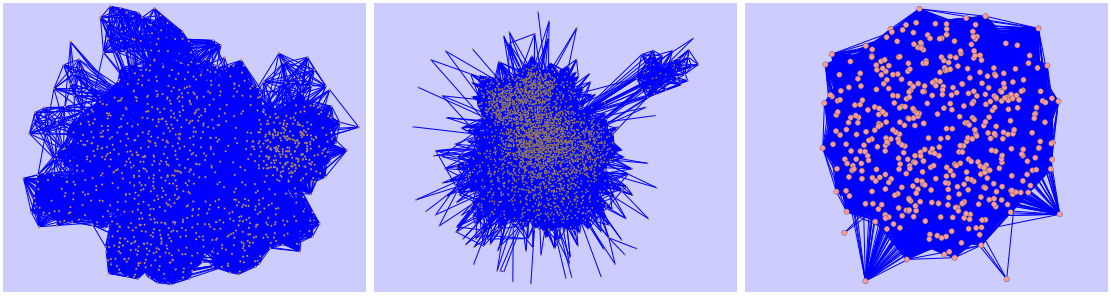
Este trabalho analisa três conjuntos de dados públicos de mobilidade. O conjunto de dados da Universidade de Dartmouth [Henderson et al. 2004] (rede Dartmouth), que contém registros de mobilidade de mais de mil pessoas dentro do campus de Dartmouth, durante mais de oito semanas. O conjunto de dados do campus da USC [jen Hsu and Helmy 2005] (rede USC), que também contém registros de mobilidade em um cenário de campus universitário, com mais de quatro mil indivíduos durante mais de oito semanas. Finalmente, o conjunto de dados de mobilidade de táxis em São Francisco (EUA) [Rojas et al. 2005] (rede Táxi), que contém registros da mobilidade de 551 táxis dentro da cidade, durante quatro semanas. Em todos os casos analisamos as ocorrências de contato entre dois indivíduos, onde para cada evento é registrado o inicio e a duração do contato, todos na precisão de segundos. Nos registros de Dartmouth e USC, dois indivíduos estão em contato se eles estão usando o mesmo ponto de acesso sem fio(*wireless access point*) para se conectar à rede do campus. Nos registros de táxi, dois indivíduos estão em contato se a distância for inferior a 250 metros, que é o alcance máximo de uma rede padrão IEEE 802.11 [Piorkowski et al. 2009]. Mais detalhes dos conjuntos de dados que utilizamos neste trabalho são descritos na tabela 3.2.

Conjunto de Dados	Referência	Local	Tamanho	Número de Agentes
Dartmouth	[Henderson et al. 2004]	Campus universitário	700MB	1156
USC	[jen Hsu and Helmy 2005]	Campus universitário	160MB	4558
Taxi	[Rojas et al. 2005]	Cidade	161MB	551

Table 1. Conjunto de dados usados neste trabalho.

3.3. Grafo de Mobilidade

Neste trabalho, um cenário de mobilidade é modelado a partir de um grafo $G_t(V_t, E_t)$, em que o conjunto de vértices V_t contém os indivíduos que estão nos registros da base de dados antes do tempo t e o conjunto de arestas E_t representa todos os encontros que aconteceram antes do tempo t. Assim, esse grafo evolui ao longo do tempo e considera tanto os encontros rotineiros quanto os encontros aleatórios entre dois indivíduos. A figura 1 mostra retratos dessas redes em um dado intervalo de tempo. Observe como é difícil comparar as redes usando somente a visualização.



(a) Dartmouth, 1 semana (b) USC, 1 semana (c) Taxi, 100 minutos

Figure 1. Retratos das redes em um dado intervalo de tempo.

O peso $w_t(i, j)$ de uma aresta (i, j) é modelado como a persistência dessa aresta desde o tempo inicial 1 até o tempo corrente t [Hidalgo and Rodriguez-Sickert 2008]. Nos conjuntos de dados de Dartmouth e da USC, um intervalo de tempo é o período de 24 horas. No conjunto de dados dos táxis, um intervalo de tempo é de quatro horas. Por

exemplo, no conjunto de dados de Dartmouth, considerando que os nós i e j se reuniram em 15 dias dos últimos 30 dias, então a persistência da aresta $w_{30days}(i, j) = 0.5$, ou seja, em 50% dos últimos 30 dias, os nós i e j se encontraram. A grande vantagem de modelar o peso da aresta como a persistência em vez de, por exemplo, a duração agregada do contato é que, com a persistência, é possível detectar relações regulares e de rotina entre dois indivíduos.

O primeiro passo para analisar os padrões de mobilidade do grafo $G_t(V_t, E_t)$ é construir a versão aleatória $G_t^R(V_t, E_t^R)$ do mesmo. Essa versão aleatória deve conter as mesmas características topológicas do grafo real, ou seja, o mesmo número de nós, arestas e distribuição empírica do grau. Dessa maneira, a única diferença entre G_t e G_t^R está nas conexões entre os nós. Enquanto em G_t os nós se conectam "semi-racionalmente", em G_t^R a conexão é feita de forma puramente aleatória. Isso permite verificar com exatidão a extensão da aleatoriedade na mobilidade dos indivíduos nos cenários analisados.

3.4. Geração do Grafo Aleatório

Para construir G^R, é proposto o uso de um algoritmo de urna tão comum na geração de estruturas aleatórias [Johnson and Kotz 1977]. O primeiro passo do algoritmo é colocar na urna, para cada nó n_i, d_i "bolas" marcadas com o identificador i do nó, sendo d_i o grau do nó n_i. Depois disso, retira-se aleatoriamente da urna duas bolas b_i e b_j que estão marcadas com os identificadores i e j dos nós n_i e n_j. Se $i \neq j$ e não há uma aresta (i, j) em G^R, então os nós n_i e n_j são conectados em G^R. Esse passo é repetido até que (i) a urna esteja vazia ou que (ii) não seja mais possível conectar os nós que estão na urna. Quando não é mais possível conectar os nós que estão na urna, há uma diferença $\epsilon \approx 0.001\%$ entre o número de arestas de G e de G_R, o que consideramos insignificante. Uma análise mais detalhada sobre ϵ deixamos como trabalho futuro. Todo o procedimento de geração de grafos aleatórios é descrito no Algoritmo 1.

Algorithm 1 : Gerando um Grafo Aleatório G^R a partir de G

```
1:
2: procedimento GERAGRAFOALEATORIO(G)
3:     para todos nós n_i ∈ G faça
4:         para j = 1 → d_i faça
5:             urna.adiciona(j)
6:         fim para
7:     fim para
8:     tentativa ← 0;
9:     enquanto !urna.vazia() e tentativa < 1000 faça
10:        i ← urna.removeAleatorio();
11:        j ← urna.removeAleatorio();
12:        se i ≠ j and !G.aresta(i,j) então
13:            G^R.conecta(i,j);
14:            tentativa ← 0;
15:        senão
16:            urna.adiciona(i);
17:            urna.adiciona(j);
18:            tentativa ← tentativa +1;
19:        fim se
20:    fim enquanto
21:    ε ←urna.tamanho() /2 × G.numArestas()
22: fim procedimento
```

A partir desse algoritmo pode-se gerar um grafo aleatório G^R a partir de qualquer grafo G. No entanto, neste trabalho são analisados grafos temporais do tipo G_t que

evoluem à medida que encontros entre indivíduos ocorrem. Como nenhuma aresta é removida de G_t e novos encontros são sempre agregados, então $|E_{t+1}| > |E_t|$. Assim, para construir um grafo aleatório temporal G_t^R, decompomos G_t em t grafos de eventos $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_t$, em que cada grafo de evento $\mathcal{G}_t$ contém somente os eventos que ocorreram entre os tempos $t-1$ e t. Assim, $G_t = \{\mathcal{G}_1 \cup \mathcal{G}_2 \cup ... \cup \mathcal{G}_t\}$.

Para gerar o grafo temporal aleatório G_t^R com as devidas persistências aleatórias nas arestas, executa-se o Algoritmo 1 em cada grafo de evento $\mathcal{G}_t$, criando assim o correspondente grafo aleatório de evento $\mathcal{G}_t^R$. Assim, o grafo G_t^R nada mais é que a união dos grafos aleatórios de eventos , $G_t^R = \{\mathcal{G}_1^R \cup \mathcal{G}_2^R \cup ... \cup \mathcal{G}_t^R\}$. O peso, ou persistência, aleatório de uma aresta $(i,j) \in E_t^R$ é calculado dividindo o número de vezes que (i,j) apareceu nos grafos aleatórios de eventos $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_t$ por t. A comparação da persistência real dos encontros com a persistência aleatória permite, por exemplo, verificar a probabilidade de um encontro entre dois nós i e j ocorrer x vezes aleatoriamente em um determinado cenário de mobilidade.

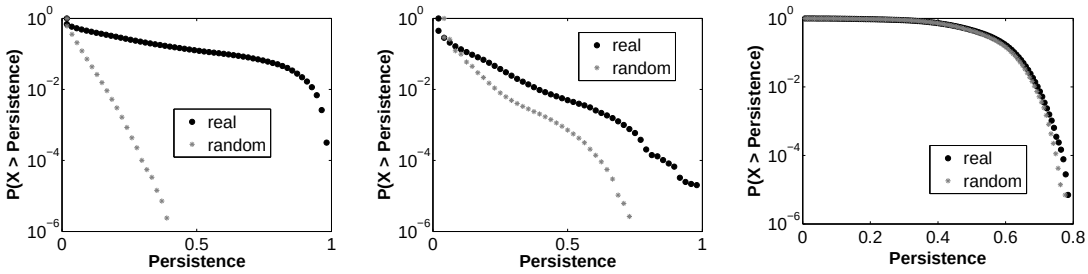
4. Aplicações

4.1. Classificação

Através da comparação de G_t com G_t^R é possível quantificar a probabilidade de uma determinada persistência de uma aresta $w(i,j)$ ser consequência da aleatoriedade de eventos ou de uma relação social real. Como os conjuntos de dados de Dartmouth e USC têm oito semanas de registros de mobilidade, usaremos as primeiras sete semanas (base de treino) para classificar as arestas e a última semana (base de teste) para verificar a eficiência da classificação. Na base de táxis, que contém quatro semanas de dados, é mantida a mesma base de testes de uma semana, ficando três semanas para a base de treino. No final da base de treino, há 175722 arestas na rede de Dartmouth, 1287992 arestas na rede USC rede e 142332 arestas da rede Táxi.

A figura 2 mostra a distribuição cumulativa complementar da persistência das arestas tanto para a rede real G_t quanto para a sua correspondente aleatória G_t^R, em que t varia até o fim da base de treino. Pode-se observar na figura 2-a que, para o conjunto de dados Dartmouth, as probabilidades de ter valores elevados de persistência é ordens de magnitude menor para a rede aleatória em comparação com a rede real. De fato, enquanto para a rede de Dartmouth a persistência das arestas é quase uniformemente distribuída, para a sua rede aleatória correspondente não existem arestas com persistência superior a $0,4$. Por outro lado, observa-se na figura 2-b que, para o conjunto de dados USC há uma diferença significativa apenas para altos valores. Isso indica que no cenário USC a mobilidade é superior à do cenário Dartmouth, favorecendo constantes encontros aleatórios. Finalmente, para o conjunto de dados Táxi, as distribuições de persistência das arestas são praticamente as mesmas. Nesse cenário, uma vez que a mobilidade é significativamente alta, mesmo aleatoriamente pode-se ter uma persistência de 100%.

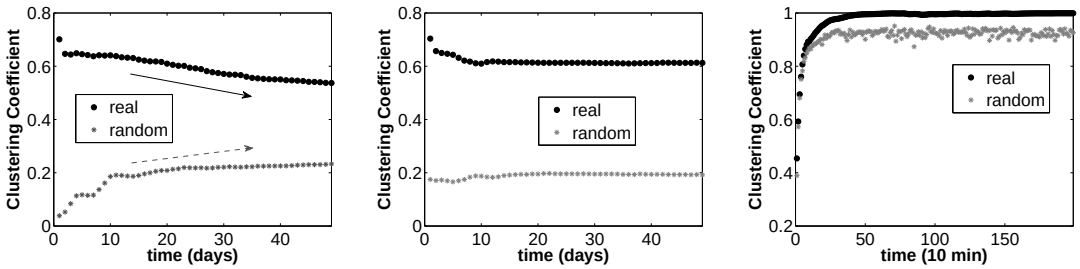
Como mencionado na Seção 3.1, o coeficiente de agrupamento de uma rede é uma métrica interessante para verificar o quão social ou aleatória é a rede. Portanto, a figura 3 mostra o comportamento do coeficiente de agrupamento para as três redes analisadas G_t e seus correspondentes aleatórios G_t^R ao longo do tempo. A figura 3-a indica que, para o conjunto de dados Dartmouth, nos primeiros dias o coeficiente de aglomeração de G e G^R é diferente em ordens de magnitude. No entanto, ao longo



(a) Dartmouth (b) USC (c) Taxi

Figure 2. Distribuição da persistência de aresta (peso da aresta) para as três redes analisadas Gs e suas correspondentes aleatórias G^Rs.

do tempo os seus valores se tornam mais próximos, já que mais encontros aleatórios ocorrem. Por outro lado, como vemos na figura 3-b, os coeficientes de aglomeração de G e G^R para o conjunto de dados USC são quase constantes ao longo do tempo. No entanto, a diferença entre eles não é tão significativa quanto na rede Dartmouth, possuindo a mesma ordem de magnitude. Finalmente, os coeficientes de agrupamento das redes de táxi são praticamente os mesmos, como observa-se na figura 3-c. Isso, juntamente com o resultado mostrado na figura 2-c, indica que G e G^R são muito semelhantes para o conjunto de dados Táxi, ou seja, nesse caso $p_r \gg p_s$. Isso faz sentido já que as decisões tomadas pelos táxis dependem da decisão do indivíduo que é levado por eles e, uma vez que táxis coletam indivíduos ao acaso na rua, $p_r \gg p_s$. Por isso, a partir de agora só serão analisados os conjuntos de dados de Dartmouth e USC.



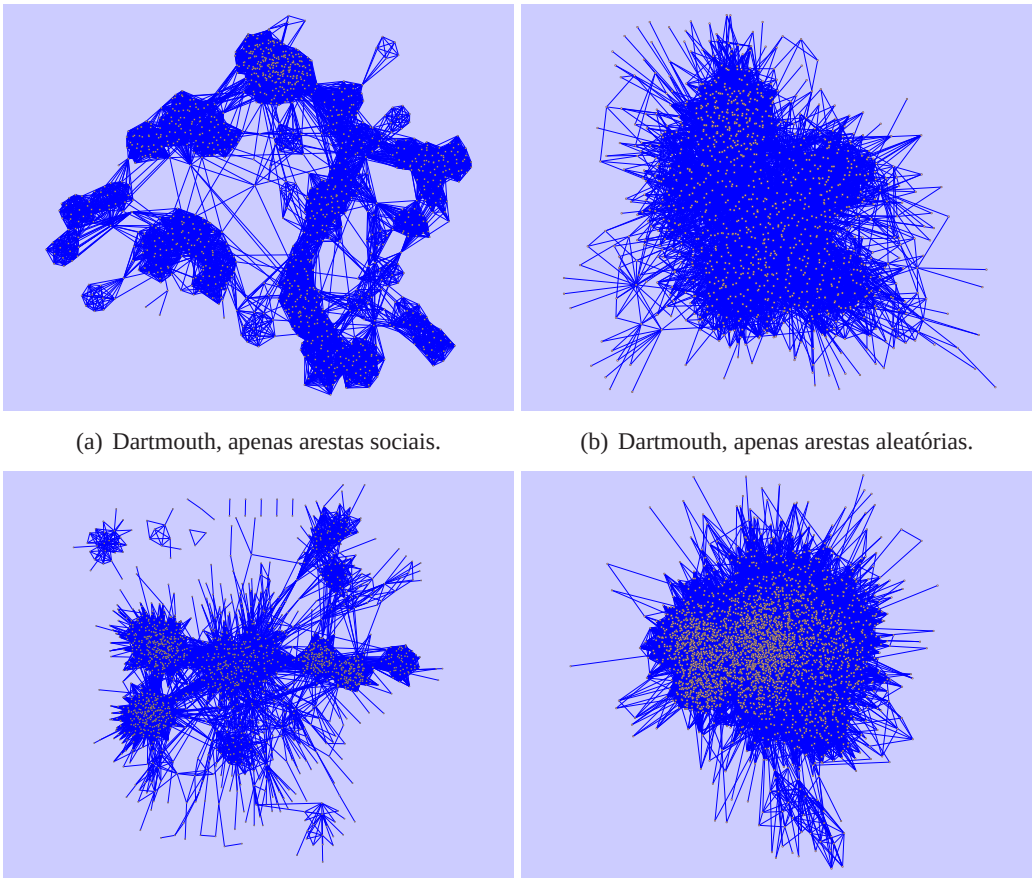
(a) Dartmouth (b) USC (c) Taxi

Figure 3. Evolução do coeficiente de aglomeração das três redes analisadas Gs e suas correspondentes aleatórias G^Rs.

Para as redes não-aleatórias, ou seja, Dartmouth e USC, pode-se usar as informações observadas na figura 2 para quantificar a probabilidade p_w de existir uma aresta com persistência superior a w e que seja consequência de encontros aleatórios. Com isso, pode-se definir um limite T_w para a persistência de uma aresta (i, j) de modo que, se $w(i, j) < T_w$, então essa aresta é classificada como *aleatória*, caso contrário, como *social*. Ao observar os valores de (x, y) da curva aleatória na figura 2, pode-se definir um valor de T_w que torna a probabilidade p_w de erroneamente classificar uma aresta aleatória como social suficientemente pequena e aceitável para uma determinada aplicação. Por exemplo, se definirmos $T_w = 0,4$ para a rede de Dartmouth, então a probabilidade de erroneamente classificar uma aresta aleatória como social é ≈ 0. É importante ressaltar que esse método é capaz de estimar a taxa de erros do tipo falso positivo para a classificação

de arestas como sociais (p_w), mas como a base de dados usada neste trabalho não possui a informação sobre quais relacionamentos são sociais, não é possível estimar a taxa de erro do tipo falso negativo.

Uma vez que pode-se classificar arestas como sociais e aleatórias, pode-se verificar como as redes evoluiriam caso fossem constituídas somente por um desses dois tipos de arestas. Assim como foi feito na figura 1, a figura 4 ilustra retratos das redes de Dartmouth e USC após uma semana de interações, mas considerando apenas as arestas sociais (primeira coluna) ou apenas as arestas aleatórias (segunda coluna). Para a rede de Dartmouth $T_w = 0,2$ e, para a rede USC, $T_w = 0,1$.[1] Pode-se observar que, enquanto as redes geradas usando apenas arestas aleatórias são muito semelhantes a uma rede aleatória gerada, por exemplo, pelo modelo de Erdös e Rényi, as redes geradas usando apenas arestas sociais são mais similares com as redes sociais vistas na literatura, com a presença clara de comunidades disjuntas. Note também como elas são visualmente diferentes das redes completas vistas na figura 1.



(a) Dartmouth, apenas arestas sociais. (b) Dartmouth, apenas arestas aleatórias.

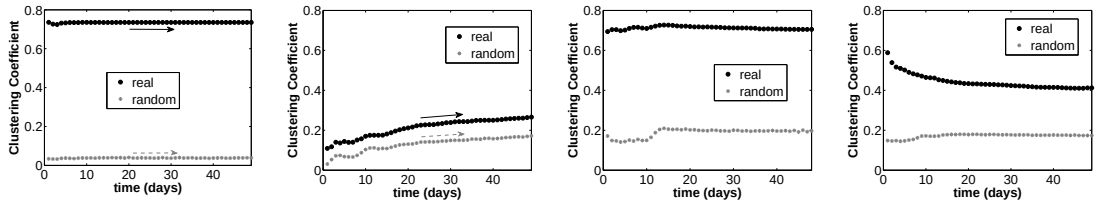
(c) USC, apenas arestas sociais. (d) USC, apenas arestas aleatórias.

Figure 4. Retratos das redes de Dartmouth e USC depois de 1 semana de interações, considerando apenas arestas sociais ou aleatórias.

É importante também analisar o comportamento do coeficiente de agrupamento das redes quando elas contêm somente arestas de um determinado tipo, uma vez que

[1]Esses valores foram definidos manualmente, visando garantir tanto uma baixa taxa de falso positivos quanto uma quantidade significativa de arestas sociais.

elas indicam precisamente se uma rede é classificada como social ou aleatória. A maior diferença entre os valores é verificada na rede de Dartmouth, como pode ser observado nas figuras 5-a e 5-b. Quando são consideradas apenas arestas sociais, a diferença entre os coeficientes de agrupamento é constante e é maior que uma ordem de magnitude. Por outro lado, quando são consideradas apenas as arestas aleatórias, os coeficientes de agrupamento são semelhantes e evoluem juntos. Para a rede da USC, como pode ser observado nas figuras 5-a e 5-b, a diferença entre os coeficientes de agrupamento aumentou quando se considera apenas as arestas sociais em comparação com o cenário quando a rede contém todas as arestas. Por outro lado, quando se considera apenas as arestas aleatórias, a diferença diminuiu ao longo do tempo, com os valores se tornando semelhantes.



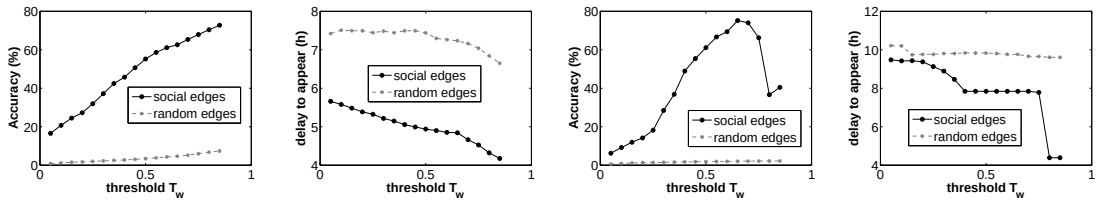
(a) Dartmouth, apenas arestas sociais. (b) Dartmouth, apenas arestas aleatórias. (c) USC, apenas arestas sociais. (d) USC, apenas arestas aleatórias.

Figure 5. Evolução do coeficiente de aglomeração das redes de Dartmouth e USC Gs e suas correspondentes redes aleatórias G^Rs quando são consideradas apenas arestas sociais ou aleatórias.

4.2. Predição

Dado que as relações sociais são regulares e se repetem com o tempo, é interessante verificar se as arestas que classificamos como sociais na base de treino aparecem também na base de teste. As figuras 6-a e 6-c mostram a porcentagem de arestas que foram classificadas como sociais e aleatórias no nosso conjunto de treino para um dado valor de T_w que também aparece no nosso conjunto de teste. Pode-se observar que à medida que o parâmetro limite T_w aumenta, também aumenta a probabilidade de uma aresta classificada como social aparecer também no conjunto de teste. Isso é explicado pelo fato de que à medida que T_w aumenta, a persistência mínima de arestas sociais também aumenta, tornando-as mais propensas a aparecer no futuro. No entanto, a taxa de aparecimento de arestas sociais diminui para a rede USC quando $T_w > 0,65$, pois há poucas arestas na rede com persistência superior a este valor ($\approx 0.1\%$, ver figura 2), o que torna a análise tendenciosa para essas arestas. Por outro lado, para a rede de Dartmouth, quando T_w aumenta, mais arestas sociais são erroneamente classificados como aleatórias, o que aumenta a probabilidade de uma aresta classificada como aleatória aparecer no futuro.

Além disso, na Figura 6-b e 6-d, é mostrado o atraso médio que as arestas classificadas como sociais e como aleatórios levam para aparecer no conjunto de teste. Como esperado, as arestas sociais geralmente aparecem antes que as arestas aleatórias. Uma vez que as arestas sociais representam encontros regulares e de rotina, espera-se que esse encontro ocorra muitas vezes durante a semana, enquanto encontros aleatórios devem ser distribuídos uniformemente ao longo do tempo. Mais uma vez, à medida que o parâmetro limite T_w aumenta, mais arestas sociais são erroneamente classificados como aleatórias, fazendo com que o atraso médio das arestas aleatórias diminua na rede de Dartmouth.



(a) Dartmouth, % of edges that appear. (b) Dartmouth, median delay. (c) USC, % of edges that appear. (d) USC, median delay.

Figure 6. O impacto do parâmetro limite T_w. (coluna da esquerda) O percentual de arestas que foram classificadas como sociais e como aleatórias que aparecem também na base de testes (coluna da direita). A mediana do atraso (em horas) para as arestas aparecerem.

Dataset	T_w	Social edges	Random edges
Dartmouth	0.2	57378 (48%)	118344 (52%)
USC	0.1	210790 (20%)	1077202 (80%)

Table 2. Parâmetros da disseminação de dados.

4.3. Disseminação de Dados

Dadas todas as evidências de que nossa classificação foi eficaz, é possível monitorar as redes a partir da sua rede social, que desconsidera as conexões aleatórias. O administrador de uma operadora de telefonia móvel, por exemplo, pode usar esse método para detectar as relações sociais de seus clientes e, portanto, usar esse conhecimento para disseminar uma informação prestada na rede sem a utilização da infraestrutura da rede, como proposto em [Daly and Haahr 2007, Hossmann et al. 2009, Katsaros et al. 2010].

Nessa direção, observa-se, então, como a informação se dissemina na rede quando (i) todas as arestas estão disponíveis, (ii) quando apenas arestas sociais estão disponíveis, e (iii) quando apenas arestas aleatórias estão disponíveis na rede. Mais uma vez, para os pontos (ii) e (iii), para a rede de Dartmouth, $T_w = 0,2$, e para a rede USC, $T_w = 0,1$. O número de arestas classificadas como sociais e como aleatórias por conjunto de dados é mostrada na Tabela 4.3.

Uma inundação simples do tipo *unicast* é realizada na rede usando fontes diferentes e, em seguida, observa-se o tempo de atraso para que os nós sejam atingidos pela disseminação de dados. Mais especificamente, é escolhido um nó aleatório n_s para ser o nó de origem, ou seja, o nó que vai começar a divulgar a informação. Depois, define-se o tempo inicial t_0 como o momento do primeiro contato entre n_s e o primeiro nó que vai receber a informação n_i. O atraso l_i do nó i é definido como 0, pois é o primeiro contato. Após a simulação, a mediana dos atrasos $\hat{l}$ é calculada para todos os atrasos l_j para todos os nós n_j que foram atingidos pela inundação.

A figura 7 mostra a média das medianas dos atrasos $\hat{l}$ e seus respectivos intervalos de confiança de 95%. Em primeiro lugar, observe como são diferentes os impactos das arestas sociais e aleatórias em ambas as redes. Enquanto na rede Dartmouth as arestas sociais, que representam 48% das arestas, são capazes de propagar informação tão bem quanto quando se tem todas as arestas, na rede USC as arestas sociais, que representam 20% das arestas, têm um desempenho significativamente pior. Isso acontece devido às características da distribuição da persistência das arestas de cada rede. Enquanto na rede Dartmouth a persistência das arestas é quase uniformemente distribuída, na rede

USC a distribuição é quase exponencial, com as arestas de baixa persistência (ou arestas aleatórias) dominando a distribuição.

Outro fato interessante é que, na rede USC, o desempenho da disseminação quando se usa apenas arestas aleatórias é melhor que quando se usa todas as arestas. Isso acontece devido à forma que se calcula t_0, que é calculado apenas quando o primeiro contato ocorre. Assim, quando o primeiro contato é um contato social, a difusão se espalha apenas na comunidade em que os nós estão inseridos (por exemplo, um laboratório ou uma sala de aula). É sabido que as pessoas ficam uma quantidade significativa de tempo nesses ambientes antes de sair, o que aumenta o atraso. Por outro lado, quando o primeiro contato é um contato aleatório, isso provavelmente significa que ambos os nós estão em áreas de transição ou áreas povoadas, como restaurantes ou arenas esportivas, o que leva a uma rápida disseminação entre os indivíduos de diferentes comunidades.

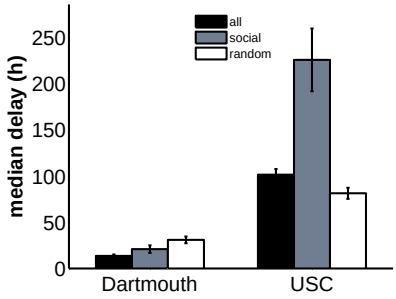


Figure 7. Comportamento da disseminação de dados quando se usa (1) todas as arestas, (2) apenas arestas sociais e (3) apenas arestas aleatórias.

Essas observações são fundamentais quando se quer divulgar uma informação prestada em uma rede sem o uso de um controle central, tais como a infraestrutura 3G das operadoras móveis. O projeto de uma solução de disseminação deve ser diferente e dependente do cenário analisado. Na rede de Dartmouth, por exemplo, pode-se usar somente arestas sociais para disseminar informações, economizando memória (apenas 48% das arestas são sociais quando $T_w = 0,2$) e recursos de processamento (o cálculo de caminhos em grafos menores/esparsos é mais rápido). Por outro lado, uma vez que foi observado que a mobilidade na rede USC é alta e arestas aleatórias divulgam informações mais rapidamente que as arestas sociais, então pode-se projetar uma solução de disseminação completamente *ad hoc*, com nós propagando informações de forma aleatória em toda a rede, sem usar recursos de memória ou processamento.

5. Conclusões e Trabalhos Futuros

Neste trabalho foi feita uma análise dos aspectos sociais de três cenários de mobilidade reais: estudantes nos campus universitários de Dartmouth e USC e táxis na cidade de São Francisco, EUA. Esses cenários foram modelados como grafos temporais em que os indivíduos são os nós e os encontros entre indivíduos são as arestas. A partir da comparação desses grafos com grafos aleatórios que possuem as mesmas características foi possível identificar quais encontros são provenientes de relações sociais e quais são provenientes de encontros aleatórios. Enquanto os cenários dos campi universitários possuem fortes características de uma rede social, o cenário dos táxis é muito semelhante a uma rede aleatória.

Como trabalho futuro, planeja-se estudar novos cenários de mobilidade real, além dos cenários gerados por modelos que existem na literatura. Além disso, acredita-se que o uso de outras métricas de redes complexas, como, por exemplo, métricas de centralidade, pode melhorar significativamente o desempenho do classificador de relacionamentos. Por fim, pretende-se avaliar mais detalhadamente o algoritmo de geração de grafos aleatórios proposto neste trabalho a fim de que se possa entender os seus limites assintóticos de erro.

References

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the World Wide Web. *Nature*, 401:130–131.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006a). Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD*, pages 44–54, New York, NY, USA. ACM.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006b). Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA. ACM.

Barbera, M., Stefa, J., Viana, A., de Amorim, M., and Boc, M. (2011). VIP delegation: Enabling VIPs to offload data in wireless social mobile networks. In *Proc. of IEEE DCOSS*, pages 1 –8.

Conti, M., Di Pietro, R., Gabrielli, A., Mancini, L. V., and Mei, A. (2010). The Smallville Effect: Social Ties Make Mobile Networks More Secure Against the Node Capture Attack. In *ACM MobiWac*.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., and Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of 14th ACM SIGKDD*, pages 160–168, New York, NY, USA. ACM.

Daly, E. M. and Haahr, M. (2007). Social network analysis for routing in disconnected delay-tolerant manets. In *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '07, pages 32–40, New York, NY, USA. ACM.

E. M. Daly and M. Haahr (2009). Social Network Analysis for Information Flow in Disconnected Delay-Tolerant MANETs. *IEEE Transactions on Mobile Computing*, 8(5):606–621.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 7:17.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453:779–782.

Henderson, T., Kotz, D., and Abyzov, I. (2004). The changing usage of a mature campus-wide wireless network. In *Proceedings of the 10th annual international conference on Mobile computing and networking*, MobiCom '04, pages 187–201, New York, NY, USA. ACM.

Hidalgo, C. A. and Rodriguez-Sickert, C. (2008). The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017 – 3024.

Hossmann, T., Legendre, F., and Spyropoulos, T. (2009). From contacts to graphs: pitfalls in using complex network analysis for dtn routing. In *Proceedings of the 28th IEEE international conference on Computer Communications Workshops*, INFOCOM'09, pages 260–265, Piscataway, NJ, USA. IEEE Press.

Hui, P., Crowcroft, J., and Yoneki, E. (2008). Bubble rap: social-based forwarding in delay tolerant networks. In *ACM MobiHoc*.

jen Hsu, W. and Helmy, A. (2005). Impact: Investigation of mobile-user patterns across university campuses using wlan trace analysis. *CoRR*, abs/cs/0508009.

Johnson, N. and Kotz, S. (1977). *Urn models and their application: an approach to modern discrete probability theory*. WILEY SERIES in PROBABILITY and STATISTICS: APPLIED PROBABILITY and STATIST ICS SECTION Series. Wiley.

Katsaros, D., Dimokas, N., and Tassiulas, L. (2010). Social network analysis concepts in the design of wireless ad hoc network protocols. *Netwrk. Mag. of Global Internetwkg.*, 24:23–29.

Kochem Vendramin, A. C., Munaretto, A., Regattieri Delgado, M., and Carneiro Viana, A. (2011). GrAnt: Inferring Best Forwarders from Complex Networks' Dynamics through a Greedy Ant Colony Optimization. Rapport de recherche RR-7694, INRIA.

Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD*, pages 611–617, New York, NY, USA. ACM.

Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD*, pages 462–470, New York, NY, USA. ACM.

Mary Schurgot and Cristina Comaniciu and Katia Jaffrès-Runser (2012). Beyond Traditional DTN Routing: Social Networks for Opportunistic Communication. *accepted for publication in IEEE Communications Magazine*.

Miklas, A. G., Gollu, K. K., Chan, K. K. W., Saroiu, S., Gummadi, K. P., and De Lara, E. (2007). Exploiting social interactions in mobile systems. In *Proceedings of the UbiComp '07*, pages 409–428, Berlin, Heidelberg. Springer-Verlag.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA. ACM.

Mtibaa, A., May, M., Diot, C., and Ammar, M. (2010). Peoplerank: Social opportunistic forwarding. In *IEEE Infocom*.

Newman, M. (2003). The structure and function of complex networks.

Piorkowski, M., Djukic, N. S., and Grossglauser, M. (2009). A Parsimonious Model of Mobile Partitioned Networks with Clustering. In *The First International Conference on COMmunication Systems and NETworkS (COMSNETS)*, Bangalore, India.

Rojas, A., Branch, P., and Armitage, G. (2005). Experimental validation of the random waypoint mobility model through a real world mobility trace for large geographical areas. In *Proceedings of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, MSWiM '05, pages 174–177, New York, NY, USA. ACM.

Thakur, G. S., Helmy, A., and Hsu, W.-J. (2010). Similarity analysis and modeling in mobile societies: the missing link. In *Proc. ACM CHANTS*, pages 13–20.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, 393:440–442.

Zyba, G., Voelker, G. M., Ioannidis, S., and Diot, C. (2011). Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd. In *Proceedings of IEEE INFOCOM 2011*, pages 1179–1187. IEEE.