

Instituto de Matemática e Estatística
Universidade de São Paulo

Usando Redes Aleatórias na Análise de Mobilidade
Monografia Sobre a Publicação

Samuel Praça de Paula

MAC5743 - Computação Móvel
Professor Alfredo Goldman vel Lejbman

São Paulo, junho de 2013

Sumário

1	Introdução	1
2	Cenário e abordagens existentes	1
3	Proposta	2
3.1	Definições	2
3.2	Os cenários estudados	3
3.3	Redes Complexas Baseadas em Decisão	4
3.4	Gerando um grafo aleatório	5
3.5	O modelo	5
4	Aplicações	8
4.1	Classificação	8
4.2	Predição	10
4.3	Disseminação de dados	11
5	Trabalhos futuros	12
6	Conclusão	12

1 Introdução

Atualmente, a existência de uma concentração de pessoas em determinada área tem se tornado cada vez mais próxima de significar uma concentração de dispositivos móveis comparável. Em muitos locais com grande movimentação de pessoas, como prédios comerciais, universidades, ou até regiões inteiras de um centro urbano, é possível observar a aproximação simultânea de um grande número de dispositivos capazes de se comunicar entre si.

Isso permite cogitar-se uma grande variedade de usos dessa rede móvel que surge espontaneamente através da dinâmica social humana. Uma das ideias mais naturais é estabelecer algum tipo de comunicação entre os agentes da rede que não dependa de uma infra-estrutura como uma antena, roteadores conectados a internet cabeada, etc.

E, para explorar a rede móvel com fins desse tipo, é de grande interesse conseguir descobrir o *comportamento* da rede. Esta é a premissa do artigo “Usando Redes Móveis na Análise de Mobilidade”, por Melo et al. [3], que é o ponto central desta monografia.

Nesse artigo, os autores afirmam tratar-se de um problema importante e que já foi estudado no passado. No entanto, apontam para o fato de que, embora os trabalhos existentes tenham apresentado resultados interessantes na compreensão desse tipo de fenômeno, os modelos propostos por eles falham em descrever precisamente o comportamento das redes [3, 10].

Os autores sugerem que isto se dá pelo fato de que os trabalhos existentes descrevem apenas os aspectos sociais da formação destas redes. Trata-se de fatores como amizades, relações de trabalho, etc., que têm grande influência sobre as interações que acontecem, mas, defendem Melo e os demais autores, não são o único fator. A estratégia que eles propõem é olhar também para os aspectos aleatórios que influenciam na formação da rede.

Desta forma, o trabalho propõe uma técnica para estudar as redes móveis que consiste em modelá-las como grafos que variam com o tempo. A partir destes grafos, são gerados grafos aleatórios correspondentes. Tais grafos aleatórios também evoluem no tempo, de uma maneira que tenta capturar as mesmas características da rede real. Ao fim do período a cujos dados temos acesso, podemos comparar o grafo real com o aleatório para tentar entender o quanto do comportamento da rede é aleatório. Mais precisamente, tenta-se usar esta comparação para separar as interações da rede entre sociais, isto é, determinadas pelos fatores “clássicos” conhecidos, e aleatórias, isto é, ocorridas por motivos imprevisíveis, não sociais, etc.

Nesta monografia, apresentaremos a técnica proposta pelo artigo central, bem como os resultados apresentados por seus autores. Nosso objetivo será explicar o conteúdo do artigo, ao mesmo tempo em que oferecemos nossa visão crítica sobre tal trabalho. Note-se que as figuras aqui presentes são originárias do artigo, e não de nossa autoria, exceto onde indicarmos o contrário.

2 Cenário e abordagens existentes

Numa situação em que desejamos prever o comportamento de uma rede, é necessário capturar de alguma forma quais são as motivações por trás das ações dos agentes. Para tanto, faz-se necessário analisar os *aspectos sociais* da rede. Isso significa que deseja-se encontrar propriedades nas interações entre as entidades que permitam o entendimento de quais são suas motivações para ações futuras. O ponto central nas abordagens existentes é interpretar os agentes como seres humanos, com rotinas, amizades, relações de trabalho, crenças e motivações pessoais, etc. Isto é, explorar algum tipo de regularidade no comportamento dos agentes.

Um exemplo dado são os resultados de Crandall et al. [2], que sugerem que a ocorrência de interações

está relacionada à similaridade entre os agentes envolvidos. Os autores desse artigo afirmam que isso se dá por dois processos: as entidades evoluem de modo a se tornarem mais parecidas com aquelas com as quais se relacionam – o que é chamado de “influência social”; ou se relacionam com aquelas com as quais já são parecidas – o que é chamado de “seleção”. Esta similaridade se dá em termos do comportamento na rede, isto é, quantidade de interações, entidades com as quais são estabelecidas relações, etc.

Outro estudo [1] relacionado ao comportamento em redes sociais investiga o mecanismo de ingresso de um indivíduo numa comunidade (subconjunto dos nós altamente conectado). Verifica-se que a probabilidade de um agente entrar numa comunidade é influenciada pela quantidade de vizinhos seus que já dela participam, mas, mais ainda, também pela forma como esses contatos (amigos) estão conectados entre si.

Além desse tipo de pesquisa, também há artigos que pesquisam como obter alguma vantagem dos laços sociais entre os agentes. Há estudos [6, 8], por exemplo, que mostram como usar métricas derivadas de dados sobre comunicações entre usuários para fazer decisões de encaminhamento de mensagens mais eficientes e que limitam a sobrecarga da comunicação.

Como apontado por Melo et al. [3], todos os estudos citados procuram explorar exclusivamente “a regularidade e os repetidos padrões espaço-temporais” que se esperam do comportamento humano. Assim, tendem a ignorar a existência de interações aleatórias ou não-sociais entre os usuários das redes móveis.

Thakur et al. [10] verificaram que as modelagens existentes frequentemente falham em prever corretamente o comportamento da rede. Embora levem em conta aspectos importantes do comportamento social observados em redes reais, os modelos acabam evoluindo para redes muito diferentes das verdadeiras. Um exemplo de anomalia é o surgimento de uma única grande comunidade, quando se sabe que o comportamento real das redes sociais é o desenvolvimento de um grande número de comunidades menores [10].

Os autores do artigo aqui estudado afirmam acreditar que levar em conta a existência de fatores aleatórios e não sociais pode ser o elemento que falta para atingir-se uma predição mais acertada. Grosso modo, a técnica proposta consiste em modelar a evolução da rede como se esta se desse apenas por fatores aleatórios e, então, de alguma forma comparar a rede aleatória resultante com a real. Com tal comparação se tentará entender quais interações são aleatórias e quais são sociais, e então prever-se quais comunicações devem voltar a se realizar no futuro. A seguir discutiremos o método, para mais adiante abordarmos os resultados apresentados.

3 Proposta

Melo et al. [3] propõem estudar-se a influência de elementos aleatórios na formação da rede. Para isso, modelam a rede como um *grafo*. A partir do grafo da rede, gera-se um grafo aleatório com características semelhantes, porém com conexões ocorrendo ao acaso, e não em decorrência de qualquer mecanismo social de escolha. Aplicam-se, então, métricas sobre os grafos real e aleatório (gerado). A partir das métricas procura-se determinar algo sobre a natureza das interações e do grafo em si – por exemplo, classificando-as entre sociais e aleatórias. Antes de maiores esclarecimentos sobre o processo, vamos prosseguir a uma formalização dos conceitos envolvidos.

3.1 Definições

Um *grafo* G é um objeto matemático consistindo de um par de conjuntos $G = (V, E)$. Os elementos do conjunto V são chamados de *vértices* ou *nós*, e aqui representam os agentes da rede. A princípio V

é qualquer conjunto. Podemos considerar que é um conjunto de rótulos. Geralmente nos referimos ao *tamanho* do grafo como o número de vértices.

Cada elemento em E é um par de vértices, e é chamado de *aresta*. Representa, aqui, uma interação entre os agentes correspondentes a esses nós. Nesse contexto, todas as arestas são *não direcionadas*, isto é, simétricas: se o nó v se comunica com w , então o contrário também ocorre. Isso, como veremos, é bastante razoável dado o que iremos definir como comunicação entre dois nós.

Um exemplo bastante simples de grafo é dado pela figura 3.1, que representa o grafo $G = (V, E)$ dado por $V = \{a, b, c, d, e, f\}$, $E = \{ab, ac, bc, de, df\}$. Aqui, vw é uma forma mais curta de escrever o par $\{v, w\}$. Como $\{v, w\} = \{w, v\}$, também vw é o mesmo que wv . Tal figura não é proveniente do artigo aqui discutido.

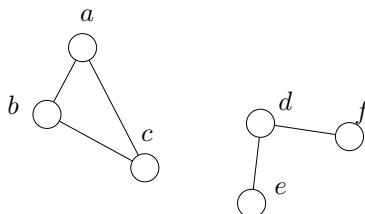


Figura 3.1: Exemplo simples de grafo.

Se v e w são nós e a aresta $\{v, w\}$ existe, dizemos que v e w são *vizinhos*, ou que estão *conectados*. Um grafo em que todos os vértices estão, dois a dois, conectados, é dito *completo*. Um grafo sem qualquer aresta é dito *vazio*.

A *densidade* de um grafo é a quantidade de arestas existentes, em comparação ao número máximo de arestas possíveis. Note que, no caso de arestas não direcionadas, este número máximo é $n(n-1)/2$, se n é o número de nós. Quanto mais arestas o grafo possui, mais denso é. No entanto, não há uma distinção clara entre o que caracteriza um grafo denso ou não denso – isto é, esparsos. Esses termos são mais úteis quando pensamos em comparar dois grafos do mesmo tamanho.

Outro termo que não possui uma definição rígida é o de *aglomerado* (ou *cluster*). Um aglomerado é um conjunto de vértices com grande número de arestas entre si. Um aglomerado no grafo corresponderá à nossa noção de comunidade na rede social. Essa noção ficará melhor formalizada através de uma das métricas consideradas.

Finalmente, iremos nos referir ao *grau* de um vértice, que é seu número de vizinhos. Ou, formalmente, no grafo $G = (V, E)$, dado um vértice $v \in V$, seu grau em G é

$$d_G(v) = |\{w \in V : vw \in E\}|.$$

Um vértice com grau muito alto corresponde a um agente que se comunica com uma grande quantidade de pares na rede. É o que por vezes chamamos de *hub*, e tende a ocorrer em redes sociais [3].

3.2 Os cenários estudados

Os dados usados nesse artigo são provenientes de três casos reais de redes móveis. Dois deles são de *campi* universitários, a saber, das universidades de Dartmouth [4] e da Carolina do Sul (USC) [5].

Nas redes das universidades, considera-se que há comunicação entre duas unidades móveis se, em algum momento, ambas se conectam ao mesmo ponto de acesso (AP). Conhecemos o tempo de duração dessas interações com precisão de segundos.

O terceiro cenário é uma rede proveniente de estudos sobre a mobilidade de táxis na cidade americana de São Francisco [9]. Neste caso, considera-se que dois agentes estão em contato se a distância entre os dois for inferior a 250 metros, o alcance máximo de uma rede padrão IEEE 802.11 [3].

Os dados de ambas as universidades foram coletados ao longo de 8 semanas, contra 4 semanas no caso da rede dos táxis. Veremos adiante que esse período menor de coleta de dados não representará um problema para a compreensão da natureza desta.

A tabela 3.1 mostra o número total de nós dessas redes, bem como o número total de arestas observadas.

Tabela 3.1:

Rede	Número de nós	Período	Número de arestas
Dartmouth [4]	1156	8 semanas	175722
USC [5]	4558	8 semanas	1287992
Táxis [9]	551	4 semanas	142332

3.3 Redes Complexas Baseadas em Decisão

No artigo, os autores chamam sistemas como as redes móveis, em que as entidades são “capazes de interagir entre si de uma maneira racional, refletindo seus interesses e dinâmica de atividade”, de *Redes Complexas Baseadas em Decisão* (RCBDs). Entende-se tais redes como grafos. Aqui, “complexas” se refere principalmente ao fato de pensar-se em redes com um grande número de nós e arestas – ou seja, o tamanho e densidade do grafo.

O que caracteriza uma RCBD, na concepção dos autores, é que as arestas normalmente se realizam como consequência de decisões semi-rationais. Tal definição é explicada da seguinte maneira: o “normalmente” se refere ao fato de que o sistema está sujeito a eventos aleatórios; “semi-rationais” alude à existência de decisões irracionais.

O exemplo dado é o dos colegas de trabalho Silva e Moreira, que trabalham no mesmo escritório no período das 8:00 às 18:00. Levando em conta apenas as decisões racionais, e supondo que a decisão racional aqui é chegar no trabalho no horário certo, esperamos que Silva e Moreira estabeleçam comunicação diária a partir das 8:00.

No entanto, um dia Moreira pode ficar preso no trânsito e chegar atrasado. Esse evento aleatório muda o comportamento da rede. Ainda, pode ser que o chefe deles mude o turno de Moreira para das 10:00 às 20:00. Moreira, mesmo tomando uma decisão racional, terá mudado de comportamento na rede. Ainda, pode ser que Moreira, no primeiro dia após a decisão do chefe, acabe esquecendo e chegue às 8:00, o que é uma decisão irracional.

Os autores formalizam esses diferentes aspectos da seguinte maneira. Cada agente pode tomar uma decisão *social* D_s ou uma decisão *aleatória* D_a . Ele tem probabilidade p_s de executar D_s , e probabilidade $p_a = 1 - p_s$ de executar D_a . De maneira informal, se $p_s \gg p_r$, a rede evolui para uma rede aleatória bem estruturada, com fenômenos típicos como *hubs* e comunidades. Se, por outro lado, $p_r \gg p_s$, a rede tende a desenvolver características de uma rede aleatória.

É para atender a tais fatores que os autores introduzem a técnica de comparar o grafo real com um aleatório para entender a influência dos aspectos sociais e não sociais/aleatórios.

3.4 Gerando um grafo aleatório

Para implementar o que acabamos de descrever, os autores desejam ter uma rotina que, dado um grafo G , construa um grafo aleatório G^R com características semelhantes. Essas características são:

1. O mesmo número de nós. Ou simplesmente o mesmo conjunto de vértices: $V^R = V$.
2. O mesmo número de arestas.
3. A mesma “distribuição empírica dos graus”. O significado disto não é explicado no artigo, mas, a julgar pelo algoritmo, supomos que signifique que, para cada vértice do grafo real, seu correspondente no grafo aleatório deve ter o mesmo grau.

Interpretamos esta lista como a intenção que, de alguma forma, a única diferença entre os grafos aleatório e original seja a escolha de *quais* arestas adicionadas. Não se altera o número de agentes nem quantas conexões cada agente faz, mas apenas o critério de escolha das arestas realizadas. No grafo aleatório, o critério é simplesmente um sorteio (supostamente uniforme).

O algoritmo proposto é o seguinte. Para cada vértice v do grafo, adiciona-se a uma urna uma cópia de v para cada vizinho que v possui. Isto é, na urna, cada vértice aparece tantas vezes quanto for seu grau $d_G(v)$.

A seguir, sorteiam-se pares de vértices da urna. Como comentamos, supomos que este sorteio seja uniforme, mas isto não é explicitamente informado no texto. A cada sorteio, se os vértices sorteados são diferentes e ainda não há conexão entre eles no grafo aleatório, eles passam a ser vizinhos em G^R . Caso contrário, são devolvidos à urna.

O processo de adição de novas arestas para quando a urna fica vazia ou quando ocorrem 1000 sorteios seguidos sem uma nova adição de aresta. O algoritmo é formalmente descrito abaixo, de maneira levemente diferente da descrição original (não só devido à eliminação de alguns erros tipográficos neste ponto do artigo).

Os autores afirmam, mas não dão justificativa nem intuição do porquê, que ao fim do processo há uma diferença de cerca de 0,001% entre o número de arestas de G e G^R . Também não se comenta a satisfação da exigência sobre a “distribuição empírica dos graus”, embora se perceba a intenção de atendê-la ao adicionar $d_G(v)$ cópias de cada vértice v à urna.

3.5 O modelo

A primeira observação a ser feita é que a tradução dos nossos cenários reais para a estrutura matemática que descrevemos anteriormente (o grafo) não é totalmente imediata. Como já dissemos, os nós representam os agentes e as arestas, as interações (comunicações) entre eles. No entanto, é importante notar que nossas redes evoluem com o tempo, o que um grafo por si só não captura.

Portanto, precisamos usar um grafo que também se modifica com o tempo. Iremos chamá-lo de G_t , com o parâmetro t representando o instante observado. O que consideraremos um instante dependerá da aplicação. Considere, por ora, que um instante é um dia, de modo que o grafo G_1 é aquele representando todas as interações registradas até o fim do primeiro dia, G_2 contém todas as arestas realizadas até o fim do segundo dia, e assim por diante. Podemos chamar de G_0 o grafo inicial, isto é, sem nenhuma interação – um grafo vazio. Temos na figura 3.2 um exemplo de evolução de grafo (não proveniente do artigo de que estamos tratando).

Algoritmo 3.1: Gerador de grafos aleatórios

Entrada: Grafo $G = (V, E)$ que o aleatório pretende refletir

Saída: Grafo G^R gerado

```
1  $G^R \leftarrow$  grafo vazio com  $n = |V|$  nós;
2  $\text{urna} \leftarrow$  urna vazia;
3 para cada  $v_i \in V$  faça
4   para  $j \leftarrow 1$  até  $d_G(v_i)$  faça
5      $\text{urna.adiciona}(i)$ ;
6   fim
7 fim
8  $\text{falhas} \leftarrow 0$ ;
9 repita
10   $i \leftarrow \text{urna.removeAleatorio}()$ ;
11   $j \leftarrow \text{urna.removeAleatorio}()$ ;
12  se  $i \neq j$  e  $\neg G.\text{aresta}(i, j)$  então
13     $G^R.\text{conecta}(i, j)$ ;
14     $\text{falhas} \leftarrow 0$ ;
15  fim
16  senão
17     $\text{urna.adiciona}(i)$ ;
18     $\text{urna.adiciona}(j)$ ;
19     $\text{falhas} \leftarrow \text{falhas} + 1$ ;
20  fim
21 até  $\text{urna.vazia}()$  ou  $\text{falhas} \geq 1000$  ;
```

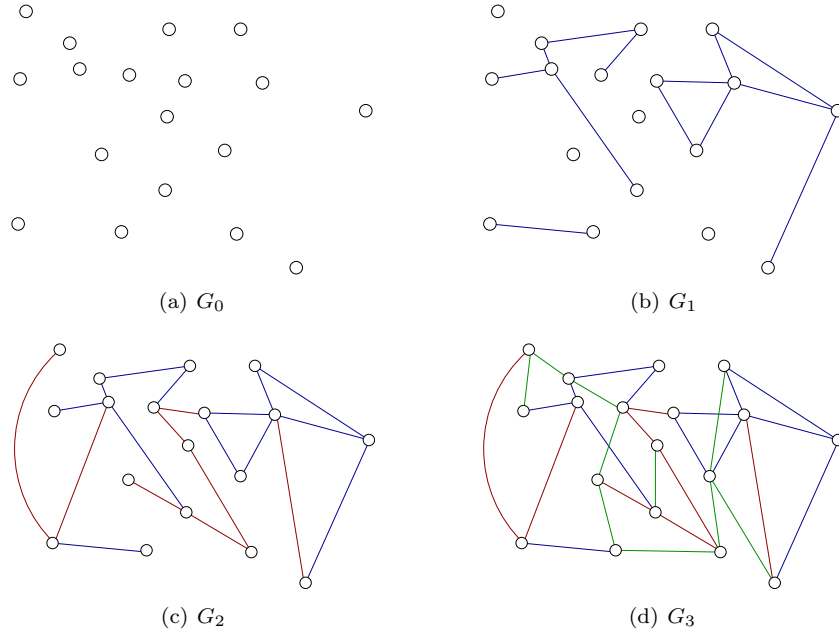


Figura 3.2: A evolução de um grafo G_t , para $0 \leq t \leq 3$.

Compare, agora com a figura 3.3, em que temos G_t para os grafos das universidades após uma semana. O grande número de arestas nos impede de conseguir observar qualquer estrutura sem o auxílio de um processo automatizado.

Note que, da maneira como definimos G_t , o conjunto de arestas nunca diminui. Isto é, $E_{t-1} \subseteq E_t$

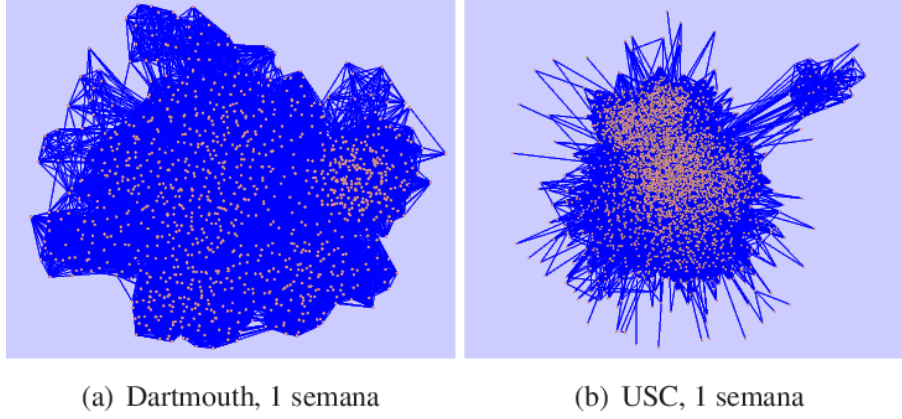


Figura 3.3: Os grafos das universidades.

para todo $t > 0$. No entanto, queremos, de alguma maneira, conseguir capturar o que acontece a cada instante, e não apenas o acúmulo de todas as interações.

Para gerar o grafo aleatório G_t^R correspondente ao nosso G_t , com as propriedades desejadas, usando a rotina já proposta, teremos que separar G_t da seguinte maneira. Chamaremos de \mathcal{G}_t o grafo que contém as arestas realizadas do instante $t - 1$ até o instante t . Desta forma, temos $G_1 = \mathcal{G}_1$, $G_2 = \mathcal{G}_1 \cup \mathcal{G}_2$, e assim por diante.

Para cada \mathcal{G}_t , geramos o grafo aleatório correspondente \mathcal{G}_t^R usando o algoritmo já apresentado. Como $G_t = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \dots \cup \mathcal{G}_t$, o grafo aleatório correspondente é $G_t^R = \mathcal{G}_1^R \cup \dots \cup \mathcal{G}_t^R$. A figura 3.4 mostra um exemplo desse conceito, baseado no mesmo grafo da figura 3.2. Novamente, a figura não faz parte do material original que o nosso texto discute.

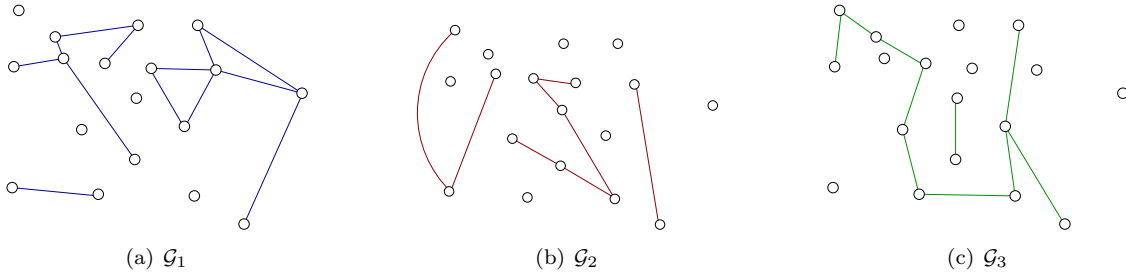


Figura 3.4: Os \mathcal{G} tais que $G_3 = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$.

Para aplicar nosso modelo, precisamos primeiro definir duas métricas que serão centrais na classificação de arestas e na predição de comportamento da rede.

- A **persistência** de uma aresta ij para o período t , denotada por $w_t(i, j) = w_t(j, i)$ (seu *peso*), é a **proporção de instantes em que ij ocorre**. Isto é,

$$w_t(i, j) = \frac{|\{i \leq t: ij \in E(\mathcal{G}_i)\}|}{t}.$$

Isto exige que definamos o que é um “instante”. Para as redes dos *campi* universitários, será um dia. Para a rede dos táxis, 4 horas. Assim, por exemplo, se na rede de Dartmouth uma aresta ocorre em 15 dentre 30 dias, sua persistência nesse período é de 0,5.

- O **coeficiente de aglomeração** (ou de *clustering*) de um vértice i no instante t é o número de arestas, em G_t , que conectam os vizinhos de i entre si, dividido pelo número máximo possível de

tais arestas. Isto é, o coeficiente de aglomeração é

$$cc_i = \frac{\text{número de arestas entre vizinhos de } i}{\text{número de arestas possíveis entre vizinhos de } i}.$$

A interpretação dessa métrica é que ela captura o quanto da vizinhança de i se parece com uma comunidade/aglomerado.

O coeficiente de aglomeração **do grafo** é definido como a média dentre os coeficientes de aglomeração de todos os seus vértices.

A intuição é que arestas sociais devem ter persistência alta, pois ocorrem consistentemente, e arestas aleatórias devem ter persistência baixa, pois não têm razão particular para se repetirem muitas vezes. De semelhante modo, esperamos que grafos que correspondem a redes sociais apresentem coeficiente de aglomeração alto em comparação com grafos aleatórios.

4 Aplicações

4.1 Classificação

Para a classificação, foram considerados os dados de todas as redes até a penúltima semana, deixando-se a última semana como a base de dados usada para testar a capacidade de predição da técnica.

Primeiro, observemos a distribuição cumulativa complementar da persistência das arestas, na figura 4.1 ao fim do período considerado.

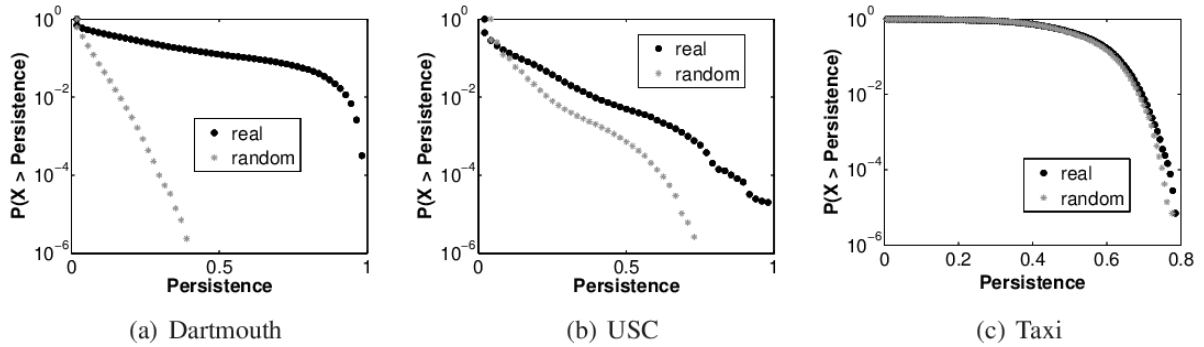


Figura 4.1: Comparação entre persistências nos grafos reais e aleatórios correspondentes

Note que, para a rede de Dartmouth, o número de arestas do grafo aleatório com alta persistência é ordens de magnitude menor que no grafo real. No caso da USC, essa distinção é um pouco mais discreta, mas ainda existe. Já no caso dos táxis, o grafo aleatório tem comportamento praticamente coincidente ao da rede real.

Os autores sugerem que uma maneira de classificar as arestas entre sociais e aleatórias seria escolher um valor de persistência “limiar” T_w tal que, a aresta é considerada social, se tem persistência superior a T_w , e aleatória, caso contrário. Nesse caso, medir o comportamento das arestas no grafo aleatório permite fazer uma boa escolha de T_w . Na figura 4.2 observa-se o resultado da classificação da rede de Dartmouth para $T_w = 0,2$. Observe que o grafo da figura (a) apresenta estruturas que esperamos encontrar numa rede social, como aglomerados e *hubs*.

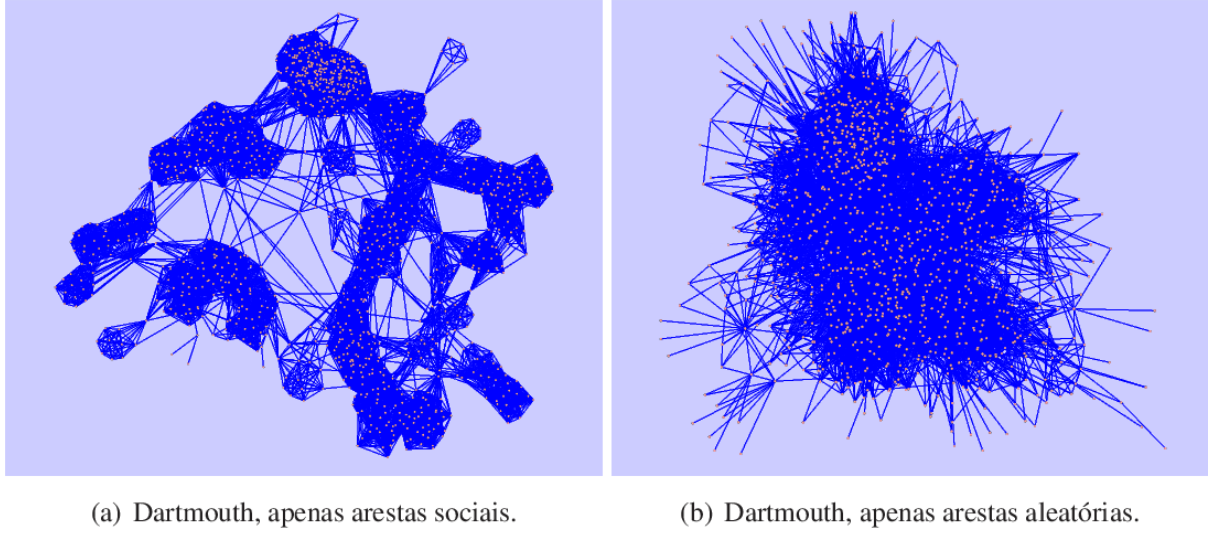


Figura 4.2: Retrato da rede de Dartmouth depois de uma semana de interações, considerando apenas arestas sociais ou aleatórias.

No caso de Dartmouth, parece mais fácil eleger um T_w . Para os táxis, é impossível fazer uma escolha que faça qualquer distinção, o que sugere que praticamente a totalidade das arestas são aleatórias. Isso faz sentido se observarmos que trata-se, de fato, de uma RCBG com $p_r \gg p_s$, uma vez que os taxistas têm suas ações definidas pela vontade dos clientes, que é um fator aleatório do ponto de vista desses agentes.

A figura 4.3 mostra a evolução dos coeficientes de aglomeração ao longo do tempo, que parece corroborar o que já se havia afirmado: a rede dos táxis tem comportamento quase idêntico ao do grafo aleatório, enquanto nos demais casos há uma distinção relativamente clara.

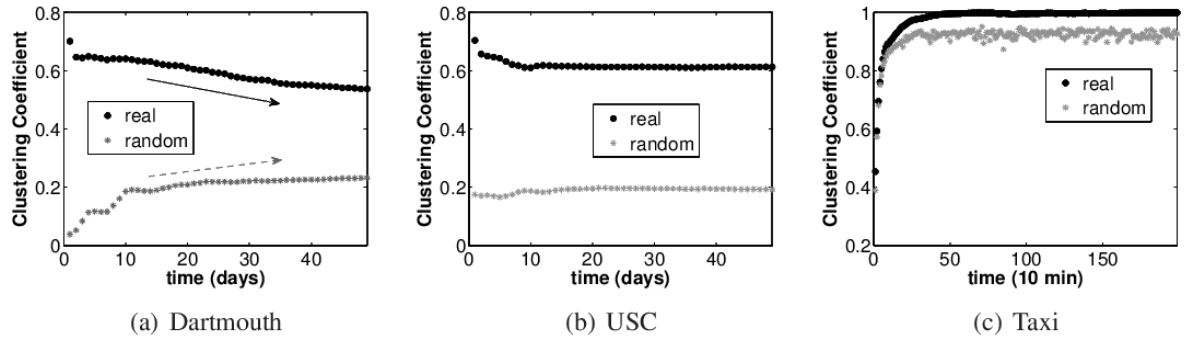


Figura 4.3: Evolução dos coeficientes de aglomeração nos grafos reais e aleatórios correspondentes.

Os autores apontam com particular ênfase para essa correspondência entre o comportamento da rede dos táxis e a expectativa de que trata-se, realmente, de uma rede predominantemente aleatória. Acreditamos que isso faz sentido, mas que é preciso tomar cuidado. Talvez não se trate de um resultado tão expressivo quanto parece à primeira vista. É preciso atentar para o fato de que o grafo dos táxis é próximo do grafo completo, em número de arestas. Assim, independentemente do comportamento dos agentes, o grafo aleatório tende a ser parecido com o real de alguma maneira. Discutimos isso brevemente na conclusão deste trabalho.

Dado o caráter considerado aleatório da rede dos táxis, os autores do trabalho não mais a consideram a partir desse ponto.

Um resultado interessante da classificação é o coeficiente de aglomeração quando consideram-se apenas arestas sociais ou apenas arestas aleatórias. A figura 4.4 mostra essa comparação, com $T_w = 0,2$ para Dartmouth e $0,1$ para a USC. Note que a classificação parece ser eficaz, com as arestas classificadas como sociais se comportando de modo muito diferente do grafo aleatório, enquanto que as arestas classificadas como aleatórias se aproximam deste. Isso é mais claro em Dartmouth, em que a distinção, como já vimos, é mais fácil.

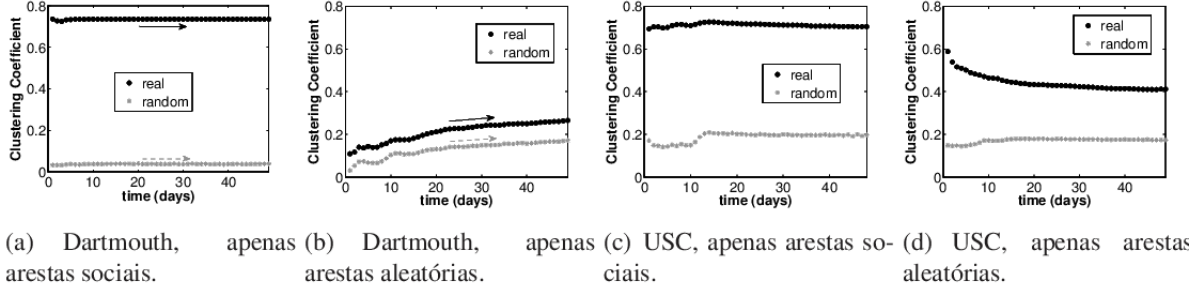


Figura 4.4: Evolução do coeficiente de aglomeração quando consideradas apenas arestas de determinada classificação, em comparação com os grafos aleatórios correspondentes.

Para esses valores de T_w , 48% das arestas de Dartmouth são classificadas como sociais, contra 20% na USC.

4.2 Predição

Os autores também medem a capacidade de predição que esse método de classificação de arestas possui. O objetivo aqui é verificar quantas das arestas de cada categoria voltam a aparecer na última semana, cujos dados não foram usados na etapa de classificação.

Espera-se que muitas arestas sociais voltem a ocorrer, enquanto poucas aleatórias apareçam de novo, pois não têm motivo particular para se realizarem novamente.

Mede-se também a demora mediana para que as arestas voltem a aparecer, métrica novamente dividida entre as categorias. Nesse caso, espera-se que as arestas sociais voltem a ocorrer mais rapidamente do que as aleatórias, pois são fruto de comportamentos regulares.

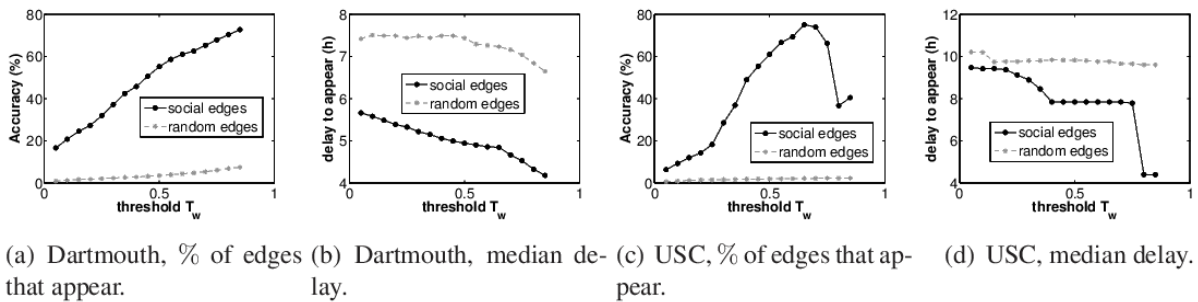


Figura 4.5: A capacidade de predição da técnica em função do T_w escolhido. Aqui estão representadas as frequências de reaparição de arestas por categoria, bem como a demora para reaparição para cada categoria.

Os resultados estão todos na figura 4.5, aparecendo como função do T_w . Observa-se uma anomalia na rede USC quando o T_w é grande demais, que os autores creditam ao fato de que nesse extremo o número

de arestas consideradas sociais é muito baixo para esse grafo, causando viés. À parte disso, os resultados sugerem uma boa capacidade de predição de interações futuras a partir da classificação realizada.

4.3 Disseminação de dados

Os autores defendem que a classificação de arestas tem aplicações relacionadas à comunicação não infra-estruturada, como as citadas na Seção 2. Argumenta-se que o conhecimento de quais arestas são sociais e quais são aleatórias permite uma exploração mais eficaz da estrutura da rede. Em particular, pode-se “descartar” as arestas aleatórias e apenas monitorar e utilizar as sociais. Trata-se de uma ideia já existente [7], e que se torna mais palpável com a existência da técnica proposta pelo artigo que discutimos aqui.

Para ilustrar essa família de aplicações, por assim dizer, os autores estudam o comportamento das redes de Dartmouth e da USC para um teste de disseminação de informação através de uma inundação simples.

Primeiro, é escolhido ao acaso um nó inicial n_s , que começa a disseminar a informação. Depois, define-se como tempo t_0 o instante do primeiro contato de n_s com o primeiro nó que irá receber a informação, n_i . O atraso ℓ_i desse nó n_i é definido como 0. Após esta interação, as demais têm os atrasos calculados com relação a t_0 .

Ao fim da simulação, a mediana $\hat{\ell}$ dos atrasos ℓ_j , para os j atingidos pela inundação, é calculada. Calcula-se $\hat{\ell}$ para o caso de (i) todas as arestas estarem disponíveis, (ii) apenas as arestas sociais poderem ser utilizadas e (iii) apenas as arestas aleatórias serem consideradas. Para a classificação, usa-se $T_w = 0,2$, para Dartmouth, e $T_w = 0,1$, para a USC. Os resultados estão representados na figura 4.6, com os respectivos intervalos de confiança de 95%.

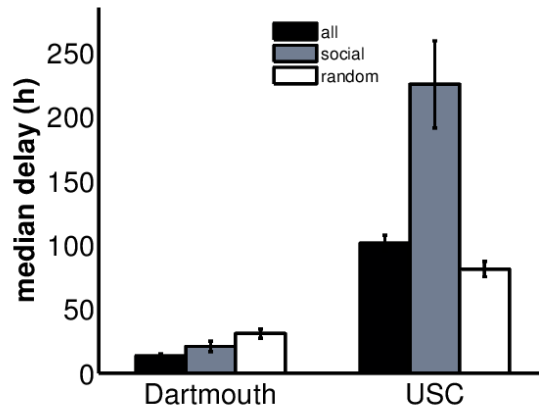


Figura 4.6: Atraso mediano para a disseminação dos dados quando usadas (i) todas as arestas, (ii) arestas sociais apenas, (iii) arestas aleatórias apenas.

Esses resultados sugerem que, no caso de Dartmouth, usar apenas arestas sociais é muito pouco menos eficaz do que usar todas as arestas, e é mais vantajoso do que usar apenas as aleatórias. Desta forma, ao projetar um protocolo de comunicação nessa rede, poderíamos considerar apenas as arestas sociais, que, para $T_w = 0,2$, são menos da metade (48%). Assim, atinge-se uma eficiência de comunicação muito próxima da melhor possível gastando muito menos recursos computacionais – memória para guardar as arestas, processamento para encontrar caminhos mínimos, etc.

Para o caso da USC, parece ocorrer o contrário: seria mais vantajoso usar apenas arestas aleatórias.

Atinge-se a melhor eficiência sem guardar qualquer informação sobre a rede, podendo-se, portanto, adotar uma estratégia completamente *ad hoc* nesse caso.

No entanto, os resultados para a USC sugerem uma distorção na métrica. Chama atenção o fato de que parece ser mais eficaz a comunicação usando apenas arestas aleatórias do que usando todas as arestas. Isto não faz sentido, uma vez que uma disseminação não deveria se tornar menos eficiente com a estrita adição de arestas disponíveis. Além disso, também causa estranhamento o fato do uso de apenas arestas sociais parecer tão desvantajoso.

Os autores creditam esse fato à definição do tempo t_0 como o momento da primeira interação, em vez do momento de início da disseminação. Isso faz com que um cenário com t_0 muito alto – ou seja, grande demora para a primeira interação – tenha uma vantagem sobre as outras no cálculo dos demais atrasos. No caso, a rede com todas as arestas terá t_0 menor que o t_0 da rede com arestas aleatórias apenas. Além disso, como nas interações sociais o primeiro encontro se dá rapidamente, seu t_0 também tende a ser pequeno, prejudicando seu desempenho em comparação com a comunicação através de arestas aleatórias.

Isso não nos parece razoável se o interesse é estudar a velocidade de comunicação em cada caso, pois, para efeitos práticos, o atraso se dá assim que a informação começa a ser divulgada por n_s . Sentimos falta, aqui, de resultados com a mediana calculada de maneira mais razoável. Seria interessante, também, medir o número de nós atingidos para cada caso.

5 Trabalhos futuros

Os autores citam como trabalhos futuros:

- Estudar novos cenários de mobilidade real e cenários gerados por modelos existentes na literatura. Consideramos importante, para validar a técnica, que se avalie seu desempenho em outros casos. Quanto à sua aplicação aos modelos existentes, especulamos que a intenção dos autores seja mais dizer algo sobre os modelos do que sobre a técnica.
- Usar outras métricas de redes complexas na comparação entre os grafos aleatório e real. Um item que nos parece muito razoável para enriquecer a técnica e avaliar melhor seu desempenho. Com novas métricas, pode ser que seja aperfeiçoada a capacidade de classificação do método aqui descrito. Outra possibilidade é que se encontrem métricas em que a comparação com o grafo aleatório não nos ajude a diferenciar as arestas sociais das aleatórias, o que por si só pode nos dizer algo sobre quais elementos de uma rede móvel são inerentemente aleatórios.
- Olhar mais cuidadosamente para o algoritmo de geração de grafos aleatórios proposto para entender seus “limites assintóticos de erro”. Não está claro, para nós, o que isso quer dizer. Talvez os autores se refiram à diferença no número de arestas, que alegam ser da ordem de 0,001%. Acreditamos que, de fato, vários aspectos do algoritmo deveriam ser ao menos cuidadosamente revisados.

6 Conclusão

O artigo se propõe a analisar as redes móveis de modo a capturar com mais precisão o comportamento de seus agentes. Especialmente, propõe uma forma de classificar as interações entre eles. A classificação sugerida divide as interações entre sociais, quando se dão em função da ação social do agente como indivíduo em seu meio – rotina, amigos, etc. –, e aleatórias, quando ocorrem por outros motivos. O

trabalho defende que essas informações são úteis para se estudar o comportamento da rede em cenários de colaboração na rede, como comunicação não infra-estruturada.

A publicação conta com uma estimativa da taxa de acerto dessa classificação (em função do parâmetro de classificação). Nos testes realizados, dentre as arestas (interações) consideradas “sociais”, para um dado parâmetro, após sete semanas de evolução da rede, mais de 60% voltam a ocorrer na semana seguinte. Em contrapartida, para o mesmo parâmetro, observa-se frequência de reaparição inferior a 5% entre as arestas tidas como “aleatórias”. Além disso, dentre as arestas que reaparecem na semana seguinte, as tidas como sociais demoram até cerca de 50% menos tempo para terem a nova ocorrência registrada.

A eficácia da técnica foi evidente em Dartmouth, cuja rede parece ter mais características sociais do que aleatórias. Ainda assim, na Universidade da Carolina do Sul, que apresenta maior mobilidade e número mais expressivo de arestas aleatórias, a capacidade de predição foi semelhante, comparando-se a taxa de reincidência das arestas conforme a classificação.

Além disso, o método proposto parece ser interessante, também, para classificar se o grafo em si é mais social ou aleatório. Na rede dos táxis, a comparação com o grafo aleatório sugere $p_r \gg p_s$ ao mostrar que a prevalência das arestas e o coeficiente de aglomeração do grafo real se comporta de maneira muito similar ao grafo aleatório.

Nas demais redes, houve um certo equilíbrio: ao que parece, nenhuma das características se sobrepõe tão nitidamente à outra. E, nesses casos, a técnica pareceu permitir estimar o quanto as características sociais se expressam, em comparação com os fatores aleatórios.

O trabalho, parece-nos, poderia melhorar em alguns pontos, a saber:

1. Definir exatamente as propriedades desejadas no grafo aleatório G^R gerado a partir de G . O que entendemos por “mesma distribuição empírica dos graus” é baseado na interpretação do algoritmo em si, mas o significado real nunca é devidamente explicitado. A falta de clareza poderia ter sido resolvida caso houvesse uma prova ou pelo menos argumentação informal de que o algoritmo atende a essa propriedade.
2. Não há uma prova, ou sequer argumentação sucinta, sobre as propriedades alegadas do algoritmo que gera o grafo aleatório. Não há, inclusive, uma boa descrição em prosa, o que seria útil, dado que a descrição formal contém erros tipográficos que, somando-se à falta de clareza em relação aos objetivos, torna difícil o entendimento da técnica de geração do grafo.

Também não se discute uma validação da técnica de geração do grafo do ponto de vista estatístico. Não somos capazes de avaliar tal aspecto objetivamente, e talvez seja difícil provar algo explicitamente sobre a distribuição de probabilidade do grafo resultante. Além disso, o estudo mais aprofundado da técnica de geração do grafo aleatório é citado como trabalho a ser realizado futuramente, mas acreditamos que seria importante o artigo contar com alguma discussão nesse sentido.

3. Parece-nos que a comparação entre o grafo real e o grafo aleatório é bem menos expressivo no caso da rede dos táxis do que nos outros. Isto pela seguinte observação: o grafo dos táxis, ao fim do processo, tem 142332 arestas. Dado que há 551 vértices, o número máximo de arestas possíveis é $\frac{551 \cdot 550}{2} = 151525$. Ou seja, o grafo dos táxis tem quase 94% de todas as arestas possíveis. Em comparação, nos grafos de Dartmouth e da Carolina do Sul, tal índice é inferior a 27% e 13%, respectivamente.

Observe que, para um grafo com 100% das arestas possíveis (o *grafo completo*), o grafo aleatório correspondente seria exatamente igual. Isto nos dá uma perspectiva de qual o comportamento assintótico da técnica de comparação conforme o grafo estudado se torna mais denso.

Portanto, no caso dos táxis, o fato de que um grafo aleatório e o grafo real tiveram comportamentos semelhantes não parece ser um indício tão óbvio da eficácia da técnica.

Talvez fosse interessante testar o comportamento da técnica num caso em que a rede tivesse comportamento esperado predominantemente aleatório, mas com um índice de mobilidade inferior (ou seja, um grafo não tão denso). Entendemos que há limitações quanto aos dados disponíveis, mas poderia haver ao menos uma menção a esse fato.

4. A maneira como foi medido o atraso no teste de inundação nos pareceu pouco razoável do ponto de vista da aplicação – o atraso, para efeitos práticos de comunicação, começa desde que o nó emissor começa a transmitir a informação –, admitidamente introduz viés no resultado, e sua escolha não é justificada. Portanto, embora este teste esteja bem alinhado com a proposta de aplicação da técnica, é difícil interpretar seus resultados de maneira objetiva.

Entretanto, se estes são pontos em que o artigo poderia se aperfeiçoar, não significa que a técnica proposta seja inválida. Ao contrário, trata-se de uma abordagem bastante razoável, que **ataca diretamente o problema apontado**, isto é, a deficiência nas técnicas existentes, e vários dos resultados sugerem sua validade.

Como, além disso, trata-se de uma técnica para estudar fenômenos de interesse prático (o que é evidenciado pelo número de trabalhos afins), esta publicação resulta interessante e potencialmente bastante relevante.

Referências

- [1] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 44–54, New York, NY, USA, 2006. ACM.
- [2] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 160–168, New York, NY, USA, 2008. ACM.
- [3] Pedro O. S. Vaz de Melo, Aline Carneiro Viana, Marco Fiore, Katia Jaffrès-Runser, Frédéric Le Mouël, and Antonio A. F. Loureiro. Usando Redes Aleatórias na Análise de Mobilidade. In *30th Brazilian Symposium on Computer Networks and Distributed Systems (SBRC)*, April 2012.
- [4] Tristan Henderson, David Kotz, and Ilya Abyzov. 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, 2004. ACM.
- [5] W Hsu and A Helmy. Impact: Investigation of mobile-user patterns across university campuses using wlan trace analysis. Technical Report cs.NI/0508009, Aug 2005.
- [6] Pan Hui, Jon Crowcroft, and Eiko Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. In *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '08, pages 241–250, New York, NY, USA, 2008. ACM.
- [7] Dimitrios Katsaros, Nikos Dimokas, and Leandros Tassiulas. Social network analysis concepts in the design of wireless ad hoc network protocols. *Netw. Mag. of Global Internetwk.*, 24(6):23–29, November 2010.
- [8] Abderrahmen Mtibaa, Martin May, Christophe Diot, and Mostafa Ammar. Peoplerank: social opportunistic forwarding. In *Proceedings of the 29th conference on Information communications, INFOCOM'10*, pages 111–115, Piscataway, NJ, USA, 2010. IEEE Press.
- [9] Andres Rojas, Philip Branch, and Grenville Armitage. 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, 2005. ACM.
- [10] Gautam S Thakur, Ahmed Helmy, and Wei-Jen Hsu. Similarity analysis and modeling in mobile societies: the missing link. In *Proceedings of the 5th ACM workshop on Challenged networks*, pages 13–20. ACM, 2010.