Extração e Análise de Dados Como Suporte a Estratégias de Comunicação D2D Cientes do Humano

Rafael L. Costa[1,4], **Aline Viana**[3], **Artur Ziviani**[2], **Leobino N. Sampaio**[1]

[1]Universidade Federal da Bahia (UFBA)

[2]Laboratório Nacional de Computação Científica (LNCC)

[3]Inria Saclay – Île-De-France

[4]École Polytechnique – Université Paris-Saclay

{rlimacosta,leobino}@ufba.br, aline.viana@inria.fr,ziviani@lncc.br

Abstract. *Real Datasets can reveal user characteristics such as mobility, social interactions, and others that will support Future Mobile Networks in routines prediction and resource management. This work introduces a framework with practices for user-data extraction and manipulation, and proposes human-aware metrics to support a novel opportunistic communication strategy. The experience is reported through a study case with MACACO Dataset and results from trace and metrics analysis, showing the importance of human-behavior based decision factors for future networking solutions.*

Resumo. *Datasets Reais podem revelar características dos usuários como mobilidade, interações sociais e outras que darão suporte a Redes Móveis do Futuro na predição de rotinas e gerenciamento de recursos. Este trabalho apresenta um framework com práticas para manipulação e extração de dados de usuários e propõe métricas ciente do humano como suporte a uma nova estratégia de comunicação oportunística. A experiência é relatada através de um estudo de caso do MACACO Dataset e resultados da análise do trace e das métricas, expondo a importância de fatores de decisão baseados em comportamento humano para soluções de rede do futuro.*

1. Introdução

A intermitência na conectividade, a forte dependência dos padrões de mobilidade dos seus usuários e dispositivos com baixo poder computacional são características que fazem com que a pesquisa em redes móveis seja bastante desafiadora. Estudos sobre a efetividade e desempenho de soluções de redes para cenários de mobilidade, em geral, são apoiados por experimentos baseados em simulação que utilizam modelos de mobilidade para a geração de *traces* sintéticos. Mais recentemente, pesquisadores têm envidado esforços na utilização de *datasets* reais (também referidos como traces) a fim de obter resultados mais robustos e realísticos.

O uso de *datasets* reais tornou-se ainda mais relevante para estudos de arquiteturas das redes móveis do futuro, que serão cada vez mais centradas nos requisitos comportamentais dos usuários [Costa et al. 2018]. Ou seja, arquiteturas em que as características e hábitos do ser humano por trás de um dispositivo de comunicação

são explorados de forma a melhor servi-lo. Por tais motivos, o estudo do comportamento humano [Thilakarathna et al. 2017] é essencial para o avanço de pesquisas na área. Um indivíduo possui características como mobilidade, traços de personalidade e sócio-econômicos, interações sociais, caráter, humor, perfil de tráfego e contexto que podem ser estudadas para oferecer um serviço de rede mais adequado [Oliveira et al. 2017]. Historicamente, diversos trabalhos examinaram características do usuário como a mobilidade [Lau et al. 2017, Xia et al. 2018]. No entanto, pouco ainda foi feito no sentido de aproximar métricas de avaliação e outros fatores de aspectos inerentes ao comportamento humano. A utilização dessas informações na predição e identificação de rotinas também carece de mais iniciativas.

Diante da necessidade de aproximar mais as redes e os seus usuários, o estudo de soluções cientes do humano através de *datasets* reais torna-se essencial. Além da disponibilidade desses *traces* reais, diversos desafios precisam ser tratados, tais como lacunas temporais e dados inconsistentes causados por erros na coleta. Critérios relativos à aquisição, armazenamento, processamento, modelagem, extração de conhecimento, análise, validação e privacidade dos dados precisam ser levados em conta, até que os dados brutos se tornem úteis para pesquisa. Com isso, esse trabalho busca responder "Como podemos extrair conhecimento da mobilidade humana de datasets reais para dar suporte ao desenvolvimento de estratégias de comunicação oportunística?".

Introduzimos um framework para extração e manipulação de dados de usuários com etapas, técnicas no solucionamento de problemas e exemplos de fontes de dados. Através de conhecimento extraído do MACACO (*Dataset* Europeu privado) [1], propomos métricas cientes do humano como suporte a uma nova estratégia de comunicação oportunística via D2D. O MACACO foi obtido por crowdsensing-móvel e contém dados de mobilidade, traços de personalidade e outros. Apresentamos um estudo de caso da manipulação do trace com a metodologia utilizada desde a seleção de usuários, inferência de locais de trabalho e casa, filtragens de erros e preenchimento de lacunas. Essa metodologia irá servir de guia para futuros trabalhos com traces reais, já que nem sempre essas práticas são adotadas para garantir resultados confiáveis. Por fim, trazemos resultados da manipulação do dataset e da análise das métricas.

O restante do trabalho está organizado da seguinte forma: na Seção 2 constam os trabalhos relacionados; na Seção 3 é detalhado o framework de extração de dados do contexto humano; na Seção 4 são propostas as métricas e divisão temporal ciente do humano; na Seção 5 estão o estudo de caso com o MACACO, análise dos resultados da sua manipulação e das métricas, e discussão de uma estratégia de comunicação oportunística; na Seção 6, conclusão, objetivos futuros e oportunidades de pesquisa.

2. Trabalhos relacionados

Ao longo dos anos, pesquisas estudaram a mobilidade através de *traces* reais e também modelos sintéticos, buscando emular a movimentação de nós na rede, seguindo determinados padrões (p.ex., humanos ou veículos). Apesar da grande disponibilidade de modelos sintéticos, na maior parte do tempo eles falham ao tentar refletir a mobilidade observada na vida real [Batabyal and Bhaumik 2015]. Com isso, a coleta e análise de *datasets* reais ganhou notoriedade por refletir naturalmente a mobilidade dos indivíduos monitorados.

[1]MACACO Dataset. Disponível em: http://macaco.inria.fr/

Apesar da quantidade considerável de trabalhos baseados em *datasets*, muitos não detalham desafios encontrados e lições aprendidas na sua manipulação, incluindo processamento, filtragem, formatação, limpeza, modelagem, enriquecimento dos dados, análise, dentre outros. O uso desses datasets sem a aplicação dessas etapas necessárias pode levar a resultados tendenciosos ou sem confiabilidade.

Dentre os *traces* populares com dados reais de usuários estão o Dartmouth, que engloba logs SNMP e tcpdump de um conjunto de cerca de 450 APs cobrindo a conexão/desconexão de clientes num Campus por um período de 5 anos; o MIT Reality Mining [Eagle and Pentland 2006], que armazena atividade de 100 smartphones relativas à proximidade de dispositivos bluetooth, torres de celular, utilização de aplicações, entre outras; e o NCCU Dataset [Tsai and Chan 2015], que conta com movimentos reais coletados através de uma aplicação instalada nos dispositivos móveis de 115 usuários ao longo de duas semanas, num Campus em Shanghai. Diferentes trabalhos extraíram de *datasets*, métricas relacionadas com o desempenho de redes móveis, tais como o tempo de duração de contato, tempo inter-contato, *Centrality* e *Radius of Gyration* [Batabyal and Bhaumik 2015].

Em [Lohan and e Silva 2017], os autores analisaram quatro *traces* com informações de geolocalização em relação a parâmetros como velocidade, aceleração, tempo de pausa e mudanças de direção na mobilidade. Já em [Shah et al. 2017], é proposto um mecanismo de reconhecimento de atividades (p. ex: identificação de pontos de parada) baseado em *traces* de mobilidade. Desafios relacionados à parte de mineração de dados dos *datasets* são mostrados. Em [Domingues et al. 2018], os autores chamam atenção sobre a importância de características espaciais e sociais extraídas de datasets reais. Um algoritmo de encaminhamento oportunístico é proposto e validado através de um dataset real e um sintético, mostrando diferenças de desempenho entre as fontes de dados. Em [Kondor et al. 2018], os autores verificam a possibilidade de cruzar dados de usuários de *traces* diferentes e anonimizados para identificá-los. O estudo é baseado em um *dataset* de telefonia móvel e um de sistema de transporte, ambos com milhões de usuários. Através dos *traces* são extraídas propriedades que causam impacto na identificação cruzada de usuários e um dos resultados mostra que a frequência na coleta de dados do *trace* aumenta a precisão das técnicas propostas.

3. Manipulação e extração de conhecimento de *datasets*

A "transformação" de dados brutos do contexto humano em conhecimento útil no âmbito de soluções de rede, apresenta desafios. Nesta seção, apresentamos o *framework* descrito na Figura 1 para extração e manipulação de dados do contexto humano, contemplando gerenciamento, análise e privacidade.

3.1. Gerenciamento de dados

O gerenciamento consiste nas etapas de (i) Aquisição, (ii) Armazenamento, Processamento e Enriquecimento e (iii) Modelagem, descritas a seguir:

- **Aquisição**: A análise do comportamento humano requer disponibilidade de dados, muitas vezes, de fontes diversas. Entre as opções, a coleta através de medições de infra-estruturas de rede sem fio, APIs específicas de serviços ou redes sociais (p.ex., Foursquare) e os dispositivos móveis. Esses representam uma das principais fontes, dada a sua ubiquidade e diversidade de sensores equipados.

Figura 1. Framewok para extração de dados brutos do contexto humano no auxílio a soluções de rede do futuro.

- **Armazenamento, Processamento e Enriquecimento**: A grande quantidade de dados que pode ser gerada requer plataformas de armazenamento seguras, escalonáveis e tolerantes a falhas que possibilitem solicitações paralelas e em tempo real. Ainda no processamento, a associação e integração de dados também podem ser necessárias: diversas fontes de dados que reagrupam diferentes tipos de dados são exploradas simultaneamente para extrair informações úteis. Além disso, é preciso aplicar técnicas de limpeza, enriquecimento por normalização, detecção de entradas falsas, interpolação geográfica e temporal, entre outras, para compensar lacunas e reduzir a inconsistência dos dados brutos [Montjoye et al. 2013]. Finalmente, a redução dimensional de dados multidimensionais também pode ser necessária antes da análise. Para isso, a seleção de recursos úteis é eficaz.
- **Modelagem**: Dados devem ser modelados em um formato que permita a extração de informações espaço-temporais e a relação entre diferentes componentes. Os grafos têm sido o formato mais utilizado para modelar comportamentos espaço-temporais de pessoas com ambientes ou laços sociais. Nesse contexto, o vértice em um grafo pode representar usuários em uma rede ou seus locais visitados [Nunes et al. 2018], enquanto as arestas conectam vértices quando um encontro acontece, ou quando locais são visitados sequencialmente por um indivíduo. Além dos grafos, Trajetórias espaço-temporais de pontos cronologicamente ordenados são bastante usadas.

3.2. Análise de dados

A análise de dados inclui extração e análise de conhecimento, bem como validação de dados, conforme discutido a seguir.

- **Extração e análise de conhecimento**: Dentre os tipos de extração de conhecimento, podemos citar: detecção e modelagem de padrões, correlação e causalidade entre entidades envolvidas, detecção de perfil comportamental, classificação

ou agrupamento de dados, e detecção de alterações ou irregularidades nos dados. A extração de conhecimento também pode ser feita através da atividade do usuário em suas redes sociais, já que suas publicações podem conter dados valiosos. Técnicas de visualização, aprendizado de máquina, inteligência artificial, métodos de interação humano-computador, modelagem de séries temporais, métricas complexas de rede, estatística e análise empírica são exemplos úteis nesse contexto.

- **Validação**: A validação consiste em verificar a corretude e a utilidade dos dados, fornecendo garantias de adequação, precisão e/ou consistência dos mesmos. Na análise estatística, a validação cruzada é uma técnica usada para avaliar como os resultados serão generalizados para um conjunto de dados independente. Outra técnica comum é o cruzamento de dados usados (geralmente incompletos ou reduzidos) com o que é chamado de dados *ground truth* (geralmente dados oficiais ou completos).

3.3. Privacidade dos dados

A privacidade dos dados do usuário deve ser garantida para oferecer suporte a aplicativos e inovação, sem prejudicar os direitos e a segurança individuais. No contexto de segurança e privacidade, aspectos, como mecanismos de autenticação e autorização, anonimização, e esquemas de incentivo e reputação foram investigados. Como a Privacidade de Dados é uma questão muito importante para a evolução das redes ciente do humano, defendemos que técnicas apropriadas devem ser usadas de acordo com o tipo de dados a serem armazenados e analisados. A seguir, mostramos métricas extraídas da mobilidade de usuários e uma divisão temporal ciente do humano.

4. Métricas para aplicações ciente do humano

Trabalhos anteriores analisaram *datasets* e extraíram métricas de mobilidade, aplicando esse conhecimento, dentre outras soluções, em estratégias de comunicação oportunística. No entanto, até então, pouco foi feito no sentido de identificar rotinas, prever deslocamentos e encontros, e sobretudo, entender a influência de aspectos do comportamento humano na mobilidade do usuário [Domingues et al. 2018]. A seguir, apresentamos métricas e noções extraídas do MACACO *Dataset*, propostas no sentido de aproximar naturalmente soluções de rede do comportamento humano e suas rotinas.

4.1. Identificando a rotina da mobilidade humana

De acordo com trabalhos na literatura como [de Melo et al. 2015], a regularidade existente na mobilidade humana contribui com a baixa entropia desses movimentos. Com isso, aspectos como regularidade espacial, temporal e social (encontros) se tornaram importantes no estudo da mobilidade. Por conta das nossas rotinas, os encontros, pontos de interesse e deslocamentos podem ser previstos[2], fatores que podem alavancar gerência de recursos de redes e tipos de comunicação (p.ex. dispositivo-a-dispositivo). Contudo, a depender do período do dia, nosso perfil de mobilidade geralmente muda (p.ex. é restrito quando estamos em casa, e mais móvel ao se deslocar de casa/trabalho). Sendo assim, propomos aqui uma observação temporal mais granular. Essa ideia se opõe a iniciativas

[2]Facebook Filed A Patent To Calculate Your Future Location. https://www.buzzfeednews.com/article/nicolenguyen/facebook-location-data-prediction-patent

anteriores em que métricas de mobilidade são calculadas em janelas constantes (p.ex., 24 horas, ou as 6 horas anteriores observadas) [Domingues et al. 2018].

Na Tabela 1 apresentamos nossa abordagem temporal onde o dia é dividido em 6 períodos não-uniformes, com durações diferentes. Ressaltamos que essa foi uma divisão adotada no MACACO Dataset para refletir diferentes perfis de mobilidade da população ao longo do dia. Defendemos que a temporalidade deve ser adaptada à população estudada para trazer resultados mais precisos nas pesquisas. Justificamos a nossa proposta, relacionando os períodos com os horários que a população efetua deslocamentos e/ou apresenta confinamentos mais longos. O período "EM", por exemplo, busca contabilizar o deslocamento de casa ao trabalho de grande parte dos indivíduos , enquanto no "M" o confinamento no trabalho (também no período "A") e deslocamentos menores durante o horário de almoço. Com essa divisão temporal, a intenção é extrair dos *traces* de mobilidade, métricas e fatores de decisão cada vez mais precisos e próximos das atividades reais e das rotinas dos humanos nos devidos períodos do dia. A seguir, apresentamos métricas extraídas do MACACO Dataset, mas que podem ser obtidas ou utilizadas em outros traces de mobilidade.

Tabela 1. Divisão Temporal Diária Proposta

#	Período	Intervalo de tempo
EM	Manhã Cedo	06:00:00 - 09:59:59
M	Manhã	10:00:00 - 13:59:59
A	Tarde	14:00:00 - 17:59:59
EE	Anoitecer	18:00:00 - 20:59:59
E	Noite	21:00:00 - 23:59:59
N	Madrugada	00:00:00 - 05:59:59

4.2. Grau de Centralidade como consciência social

O *Grau de Centralidade* (do inglês *Centrality Degree - CD*) mede os vínculos sociais de um usuário (humano), ou seja, seus encontros. Um usuário com maior grau é mais "popular", ou seja, encontra mais pessoas e, portanto, tem um potencial maior de entregar mensagens em uma estratégia de comunicação D2D oportunística. Calculamos coeficientes das métricas por período do dia e aprendemos com a mobilidade do usuário durante uma semana k para aplicar o conhecimento na semana $k+1$. Portanto, a equação 1, representa o *CD* médio de um nó u em um período $p \in (EM, M, A, EE, E, N)$ como $(\Delta_{CD_p(u)})$. O somatório de $i = 1$ a d representa os dias consecutivos da semana k e d o número de dias anteriores considerados (5, excluindo finais de semana). Já n é o número de nós na rede e $e_{(u,v)}$ é um índice do valor 1 se houver uma aresta e entre os nós u e v no período p. Considerando nossa rede como um grafo de contato dinâmico $G_t = (V, E_t)$, em que V é o conjunto de usuários (nós móveis) e E_t é o conjunto de arestas (contatos) detectados, onde $t \in (1, 2, ..., a)$ e $a \leq 432.000$ segundos (duração em segundos dos dias úteis da semana k). Existe uma aresta $e \in E_t$ entre dois nós, se esses estiverem no intervalo de comunicação definido no momento t, ou seja, eles estão em contato.

$$\Delta_{CD_p(u)} = \frac{\sum_{i=1}^{d} CD_p^i(u)}{d}, onde\ CD_p(u) = \frac{\sum_{v=1}^{n} e_{(u,v)}}{n-1} \quad (1)$$

4.3. Radius of Gyration como área de cobertura

O *Radius of Gyration - RG* quantifica a mobilidade de um indivíduo em relação a um centro de massa, calculado a partir de seus movimentos. Em nossa estratégia, essa métrica é usada para selecionar usuários que fizeram mais deslocamentos dentro de uma célula da rede (detalhes na próxima seção). Aqui também fazemos o aprendizado por período $p \in (EM, M, A, EE, E, N)$ de uma semana k para aplicar o conhecimento na semana $k+1$. Portanto, na equação 2 calculamos o ***RG*** médio de um nó u em um período p como $\Delta_{RG_p}(u)$. Em $i = 1$ a d, os dias da semana k, d representa o número de dias anteriores considerados (5), N é o número de posições (coordenadas) registradas, l_j é um local no índice j e l_{cm} é o centro de massa. Para cada usuário $u \in V$, consideramos que existe um conjunto $L_p = (l_1, l_2, ..., l_n)$ de locais registrados no período p da semana k. Cada local $l = (x, y)$, onde x, y são coordenadas associadas a um instante temporal. Com essa abordagem buscamos uma métrica mais precisa, refletindo os movimentos em cada período do dia.

$$\Delta_{RG_p}(u) = \frac{\sum_{i=1}^{d} RG_p^i(u)}{d},\ onde\ RG_p(u) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(l_j - l_{cm})^2}\ e\ l_{cm} = \frac{1}{N}\sum_{j=1}^{N} l_j \tag{2}$$

4.4. Sojourn Time

Com o *Sojourn Time - ST*, quantificamos a permanência de um usuário em uma célula da rede. Na nossa estratégia, aplicamos essa métrica para identificar nós que permanecem na mesma célula de potenciais interessados num conteúdo (mais detalhes na Seção seguinte). Consideramos um espaço geográfico dividido em diferentes células de uma operadora. Investigamos a mobilidade de cada usuário $u \in V$ numa semana k, calculando o seu tempo de permanência por período p em cada célula $c \in (c_1, c_2, ..., c_n)$ para aplicar na semana $k+1$. Portanto, na equação 3, $\Delta_{ST_p^c}(u)$ é o ***ST*** médio em minutos de u em c durante p. Continuando, $i = 1$ a d representam os dias da semana k e d é igual ao número de dias anteriores considerados (5). A duração (em minutos) é obtida dos registros de data e hora associados a pelo menos dois pares consecutivos de coordenadas $(x, y) \in L_p$; e $(x, y) \in c$ (domínio geográfico de c). Já $\Delta_t^p(u)_c$ é o tempo total t no período p que o nó u permaneceu na célula c.

$$\Delta_{ST_p^c}(u) = \frac{\sum_{i=1}^{d} ST_p^c(u)_i}{d},\ onde\ ST_p^c(u) = \Delta_t^p(u)_c \tag{3}$$

4.5. Proximidade de Destino por ciência geográfica

Dadas área de cobertura e localização de uma célula, o quão próximo dela um usuário costuma chegar? Essa é a pergunta que tentamos responder por meio dessa nova métrica de ciência geográfica. Assim como anteriormente, examinamos a mobilidade do usuário por período p em uma semana k para aplicar o conhecimento na semana seguinte (k+1). O objetivo é encontrar usuários que possam se aproximar da borda ou alcançar uma próxima célula. Com isso, na Equação 4 calculamos a *Maximum Proximity - MP* média $\Delta_{MP_p^c(u)}$, ou seja um coeficiente de proximidade de um usuário u em relação a uma célula $c \in$

$(c_1, c_2, ..., c_n)$ durante um período p. Os dias da semana k variam de $i = 1$ a d, que é igual ao número de dias anteriores considerados (5). A distância entre a célula e o nó u é calculada através da tradicional fórmula de haversine. Com isso, em min obtemos a menor distância entre u e c, considerando todos pares de coordenadas de u disponíveis no período p. Ressaltamos que caso essa métrica seja aplicada a um dataset com coordenadas em outros formatos (ex: plano cartesiano), a fórmula de cálculo de distância adequada deverá ser usada.

$$\Delta_{MP_p^c(u)} = \frac{\sum_{i=1}^{d} MP_p^c(u)_i}{d},\ onde\ MP_p^c(u) = min|haversine(u, c)| \tag{4}$$

4.6. Ciência de Direção Geográfica

Essa é uma nova métrica relativa à mobilidade instantânea do usuário. Nela, extraímos a direção do deslocamento recente de um nó u para verificar se o movimento é direcionado a uma determinada célula c da rede. Para isso, calculamos a direção geodésica entre pares de coordenadas, considerando os últimos 30 minutos de deslocamento. O resultado é a moda (a direção mais comum em um conjunto de direções). Essa é a única métrica sob demanda da estratégia e que é calculada apenas se em um encontro entre dois nós $u, v \in V$, o algoritmo precise testar sua condição. Cada nó verifica sua direção recente em relação a uma determinada célula c e compara seus resultado. A seguir, apresentamos o estudo de caso do MACACO, além de resultados da análise do dataset, das métricas e da divisão temporal.

5. Experimentos, Avaliação e Estratégia

5.1. Estudo de caso: Dataset MACACO

O MACACO é um *Dataset* privado e possui um banco de dados de informações como coordenadas de localização, redes Wi-Fi e dispositivos *bluetooth* no alcance, traços de personalidade, dentre outros, de cerca de 190 usuários de diferentes países, ao longo de períodos distintos. A aquisição de dados se deu através de um aplicativo instalado em smartphones, que armazenava informações dos usuários num servidor de banco de dados.

Nas fases de processamento e enriquecimento, o primeiro desafio foi selecionar um subconjunto de usuários com características espaciais em comum, já que nosso intuito era analisar suas rotinas e dinâmica de contatos. Desta forma, identificamos uma população de 62 usuários da UFMG e PUC-MG. Essa seleção se deu com o cruzamento de rastros de mobilidade dos usuários com coordenadas geográficas da cidade de Belo Horizonte e Região Metropolitana. Inferimos uma "HOME LOCATION", selecionando usuários que permaneceram na cidade durante a madrugada, de *02:00* às *06:00*, ou seja, quando maior parte das pessoas encontram-se em casa. Ainda cruzamos dados das *10:00* às *21:00* para identificar presença no Campus da UFMG ou PUC-MG da manhã até a noite. Eliminamos 28 usuários por possuírem poucos dados (esparsos), ou informações inconsistentes, por exemplo, medições de GPS zeradas, atrasos de coleta e excesso de entradas duplicadas.

O segundo desafio foi agrupar dados dos usuários, já que o período de coleta em que cada um utilizou o aplicativo é único. O *Dataset* foi coletado entre 2015 e 2018.

Dados anteriores a 22/05/2015 foram descartados, já que houve alteração no intervalo de requisição de 1 para 5 minutos nessa data. No intuito de selecionar 4 semanas de dados (20 dias úteis), filtramos os "melhores" dias úteis de cada usuário, excluindo feriados. Essa técnica foi utilizada para poder comparar atividades por dias da semana, já que a população selecionada compartilha um contexto social (frequenta uma das Universidades).

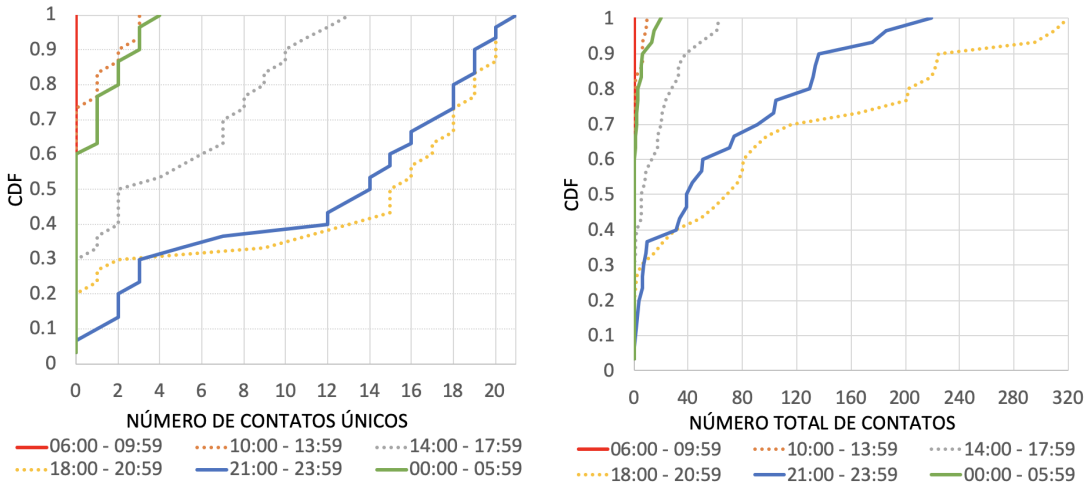
Na sequência, descartamos entradas inconsistentes geradas por cache do SO ou atraso de coleta. Feito isso, preenchemos eventuais lacunas. No período de 02:00 às 06:00, aplicamos a "HOME LOCATION" como a localização mais frequente durante os 20 dias no horário. As demais lacunas ao longo do dia foram preenchidas através de interpolação linear geodésica. Dados dois pares de coordenadas geográficas associados cada um à um instante temporal, foram calculados recursivamente os pontos intermediários (latitude e longitude) a serem preenchidos no *trace*. Finalmente, na modelagem de dados, desenvolvemos um *script* para identificar e quantificar a duração de contatos entre os usuários. O *script* recebe os dados de localização dos usuários e detecta proximidade geodésica através de um cálculo de distância. Consideramos o alcance de até 30m (alcance médio do Wi-Fi Direct). Como não há sincronização dos relógios para cada medição de cada nó, aplicamos janelas deslizantes de 5 minutos. Por exemplo, a janela inicial é de *00:00:00* a *00:05:00*. Caso dois nós u e v estivessem no alcance dentro desse instante temporal e no mesmo dia, o contato era informado. Se na próxima janela eles se mantiveram no alcance, a duração do contato também era incrementada. A saída do script foi modelada como um grafo de contatos. Nas sub-seções a seguir, realizamos extração e análise de conhecimento dos dados.

5.2. Resultados obtidos e análise

Na Figura 2(a), temos uma CDF (Função Distribuição Acumulada) com número de contatos únicos. O período "EM" possui zero contatos, portanto, não foi plotado em análises posteriores, enquanto os períodos "M" e "N" apresentam uma ocorrência de contatos muito baixa. A partir desse gráfico, vemos os períodos "EE" e "E" com cerca de 70% dos usuários com pelo menos 5 contatos únicos. Em comparação com a figura 2(b) em que plotamos a CDF para o número total de contatos, vemos coeficientes maiores, o que é justificado pelas rotinas dos usuários (neste caso, repetindo os links sociais). O período "EE" tem mais de 20% de usuários com 200 a 320 contatos.

Nas Figuras 3(a) e 3(b), avaliamos o tempo em que os usuários estiveram em contato. No primeiro gráfico, vemos cerca de 40% do trace com menos de 5% do tempo em contato. Os períodos de baixa incidência de contatos (mencionados anteriormente) geram grande impacto nesse gráfico, por isso avaliamos essa realidade por períodos na figura 3 (b). Neste gráfico, encontramos porcentagens mais altas nos períodos "EE" e "E". Nas figuras 4(a) e 4(b), avaliamos a duração média e máxima dos contatos. A duração mínima dos contatos não foi avaliada devido à incerteza resultante da falta de sincronização dos relógios. Há heterogeneidade de informações em cada período, o que reforça nossa intuição de que os seres humanos se comportam de maneira diferente de acordo com a hora do dia e o contexto e isso impacta diretamente o coeficiente das métricas.

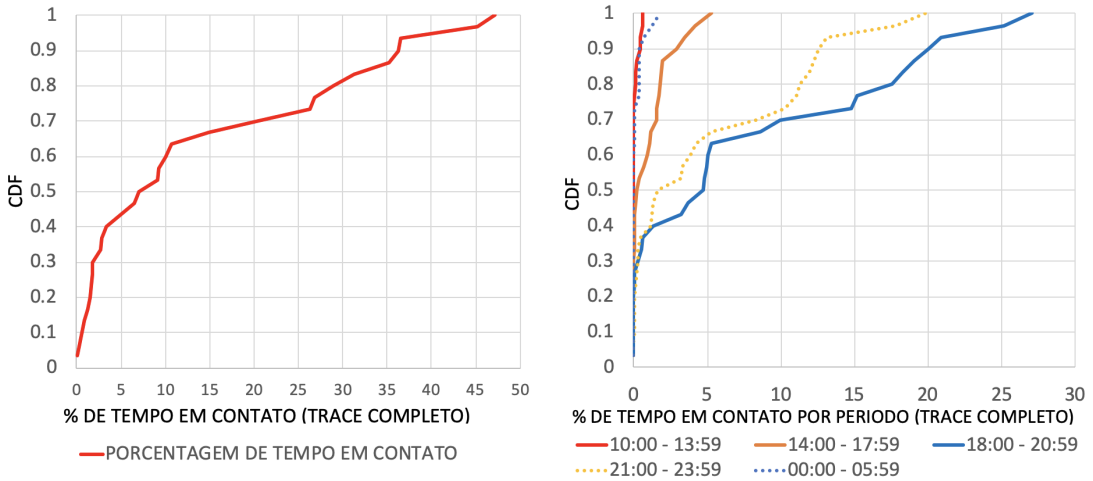
Na Figura 4(a), observamos que quase 80% dos usuários não entram em contato com ninguém no período "M". Existem apenas cerca de 10% de contatos com duração



(a) CDF do número de contatos únicos.

(b) CDF do número total de contatos.

Figura 2. Análise de Contatos (únicos e totais) no MACACO.

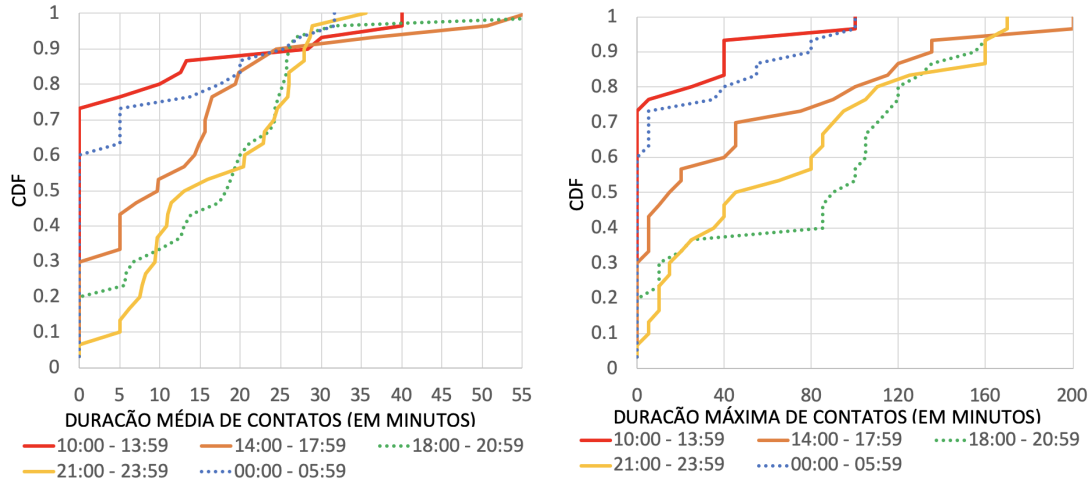


(a) CDF da porcentagem de tempo que os usuários permanecem em contato.

(b) CDF da porcentagem de tempo que os usuários permanecem em contato por período.

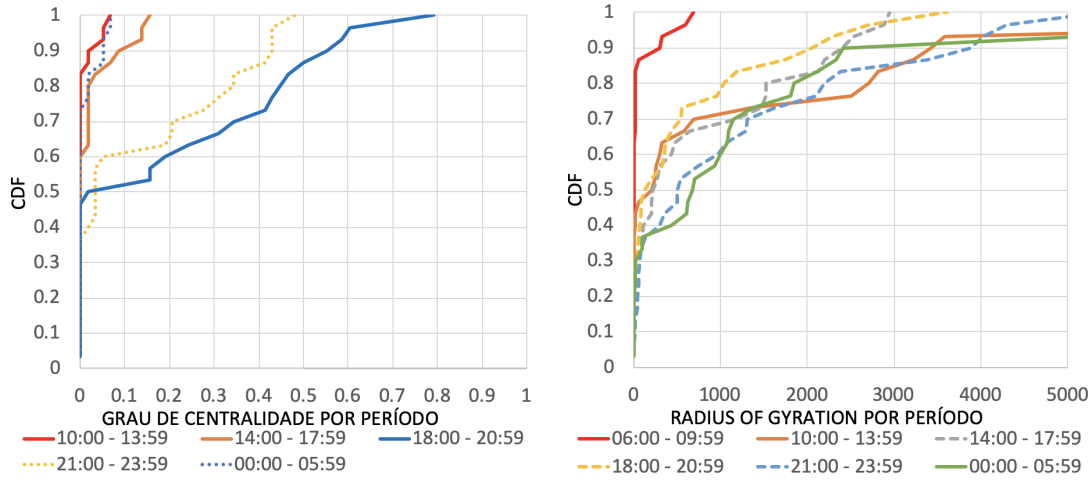
Figura 3. Análise de Contatos (porcentagem de tempo) no MACACO.

mais longa (15 a 40 minutos). No período "A", a duração aumenta. Cerca de 70% dos usuários entram em contato. Esse e o período "EE" podem ser interessantes para estudar as métricas relacionadas à ciência geográfica, pois as pessoas realizam deslocamentos maiores ao voltar para casa ou ir à a Universidade, por exemplo. No período "EE", cerca de 80% dos usuários entram em contato, com cerca de 50% de contatos com pelo menos 20 minutos de duração. O mesmo pode ser aplicado ao próximo período ("E"), com maior ocorrência de contatos e duração de 5 a 200 minutos (figura 4(b)). No período "N", obviamente, há menor incidência de contatos e curta duração, já que a maioria das pessoas tendem a estar em casa e não compartilham um contexto social neste momento. Seria diferente num trace em que os usuários moram num campus, por exemplo. Isso reforça a nossa ideia que precisamos conhecer os humanos por trás dos dispositivos para propor soluções adequadas.



(a) CDF do tempo de duração médio de contatos. (b) CDF do tempo de duração máximo de contatos.

Figura 4. Análise de Contatos (tempo médio e máximo) no MACACO.



(a) CDF do Grau de Centralidade por período. (b) CDF do Radius of Gyration por período.

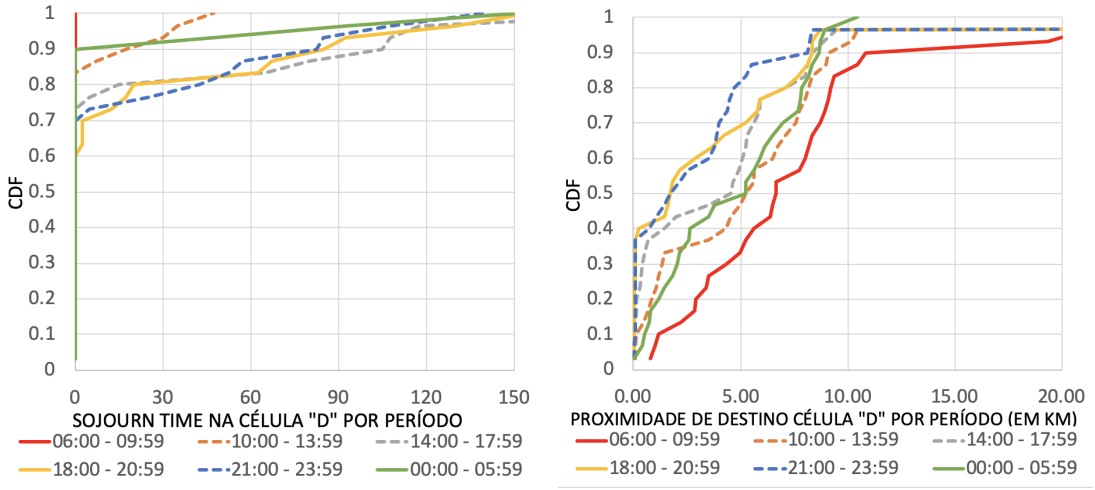
Figura 5. Análise do Grau de Centralidade e Radius of Gyration no MACACO.

Após examinar os contatos, foi realizada uma análise espacial. Devido ao contexto social compartilhado, a maior parte dos contatos acontece no campus. Assim, decidimos considerar apenas essa área geográfica para construir uma divisão de células. Não utilizamos células originais de operadoras locais pois essas são muito grandes em relação à área do campus. Isso invalidaria nossas métricas que dependem da ciência geográfica e das noções de células. Após algumas experiências com tamanhos de células, a área foi dividida em 9 células (cada uma com 113m x 113m).

Além da análise do *dataset*, aplicamos as fórmulas das métricas propostas no MACACO e avaliamos a distribuição delas. Na Figura 5(a), vemos como a "popularidade" do usuário está relacionada ao tempo. Na Figura 5(b), a distribuição do radius of gyration (em metros) apresenta deslocamentos maiores nos períodos da manhã e da tarde. Assim como na análise de contatos, verificamos que os coeficientes diferem por período,

seguindo nossa intuição.

A seguir, nas Figuras 6(a) e 6(b) plotamos a CDF das métricas *Sojourn Time* e *Proximidade de Destino*, onde usamos aleatoriamente a célula "D" (variando de "A" a "I"). No primeiro, vemos mais permanência na célula "D", nos períodos "A","EE" e "E". No segundo gráfico, pode-se observar que em períodos como "EE" e "E", cerca de 40% dos usuários estão presentes ou muito perto da célula. A proximidade pode trazer informações interessantes quando pretendemos alcançar uma célula vizinha. O reconhecimento da direção geográfica não foi plotado, pois se refere à mobilidade instantânea e deve ser calculado a partir dos últimos 30 minutos de atividade do usuário, caso seja necessário de acordo com a estratégia.



(a) CDF do Sojourn Time por período.

(b) CDF da Proximidade de Destino por período.

Figura 6. Análise do Sojourn Time e Proximidade de Destino no MACACO.

5.3. Uma Estratégia de comunicação oportunística via D2D

Com base nas métricas propostas e avaliadas, passamos à uma estratégia de encaminhamento oportunístico para tomar decisões mais adaptativas aos usuários. O intuito é fazer o descarregamento de dados de uma rede celular através de nós que oportunisticamente levam conteúdo até consumidores interessados. Nós pressupomos duas etapas nesse contexto. A primeira consiste em uma política de escolha de nós disseminadores onde o conteúdo a ser descarregado oportunisticamente é replicado. No segundo estágio (descrito nessa seção), propomos uma estratégia de encaminhamento baseada nas métricas anteriormente descritas. O objetivo dessa seção é mostrar como podemos formatar decisões guiadas por fatores que podem ir além de métricas tradicionais de mobilidade. Sendo assim, descrevemos ideias e intuições da estratégia a ser implementada. Dentre os trabalhos futuros, avaliaremos a mesma em relação a métricas de desempenho e a outras propostas com intuito similar.

A estratégia de encaminhamento é executada em cada nó u que carrega um conteúdo c, quando u encontra um nó v num instante t. Caso c seja encaminhado com êxito a v com base em uma condição do algoritmo da estratégia, v também irá passar a executá-la ao encontrar outros nós, até o tempo-de-vida (TTL) de c. Ao solicitar

um conteúdo c, um determinado consumidor informa sua célula de rede atual. Assumimos que cada nó armazena os coeficientes das métricas em uma tabela local dividida por período p, e u envia c a v apenas se uma das condições do algoritmo for atendida. Caso contrário, u aguarda o próximo encontro. Num encontro entre u e v, o algoritmo primeiro testa se v já possui o conteúdo. Se sim, aguarda o próximo encontro. Caso contrário, ele verifica se v é o destino final, e encaminha c se for verdadeiro. Caso não seja, a estratégia verifica se v está na mesma célula do nó que requisitou c. Nesse caso, o grau de centralidade é comparado, ou seja, se $\Delta_{CD_p}(v) > \Delta_{CD_p}(u)$. Se verdadeiro, significa que v encontrou historicamente mais nós nesse período p, e c é encaminhado para v, dada sua maior capacidade de disseminação local. Caso contrário, são testados o Radius of Gyration e Sojourn Time. Se $\Delta_{RG_p}(v) > \Delta_{RG_p}(u)$ e $\Delta_{ST_p^c}(v) > \Delta_{ST_p^c}(u)$ significa que v não é mais "popular" (seu grau de centralidade é menor). Por outro lado, ele tem o potencial de cobrir uma área maior naquela célula (seu Radius of Gyration é maior), além de tender a permanecer mais tempo na célula que u. Se verdadeiro, o conteúdo c é encaminhado; senão, o nó u aguarda o próximo encontro. Caso o nó consumidor tenha informado localização em uma célula diferente de v, é testada a Direção Geográfica. Caso v esteja se movendo em direção à célula do consumidor, c é encaminhado. Senão, é verificada a proximidade de destino. Caso $\Delta_{MP_p^c(v)} < \Delta_{MP_p^c(u)}$, o conteúdo é encaminhado, já que há uma chance maior de v encontrar nós na borda da célula ou dele visitar a célula de destino do consumidor durante o período p. Com essa estratégia, esperamos atingir de uma forma natural o destino do conteúdo, baseando as decisões em aspectos do comportamento humano.

6. Conclusão e trabalhos futuros

A disponibilidade de *traces* com dados reais de usuários e a disseminação de técnicas de manipulação deverão alavancar soluções de redes móveis e aproximá-las dos seus usuários humanos. Nesse trabalho, propusemos um framework para extração de dados do contexto humano e apresentamos um estudo de caso da manipulação do MACACO *Dataset*, demonstrando como algumas métricas já conhecidas podem ser abordadas diferentemente e outras novas podem ser extraídas através do conhecimento e manipulação dos dados. Trouxemos resultados da manipulação do Dataset com análise das métricas e da nova resolução temporal proposta. Esse conhecimento será aplicado numa estratégia de comunicação oportunística.

Dentre os trabalhos futuros, está a implementação da estratégia no simulador e comparação com outras propostas similares. Além disso, se faz necessária uma análise das métricas em outros datasets. Avaliaremos nossa estratégia de comunicação oportunística em termos de taxa de entrega (porcentagem de mensagens entregues com êxito), número de transmissões (para medir a sobrecarga da rede) e tempo de entrega.

Agradecimentos

Agradecemos o apoio parcial da CAPES, Inria, CNPq, FAPERJ e FAPESP.

Referências

Batabyal, S. and Bhaumik, P. (2015). Mobility models, traces and impact of mobility on opportunistic routing algorithms: A survey. *IEEE Communications Surveys Tutorials*, 17(3):1679–1707.

Costa, R., Sampaio, L., Ziviani, A., and Viana, A. (2018). Humanos no ciclo de comunicação: facilitadores das redes de próxima geração. In *Livro de Minicursos do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) 2018*, Campos do Jordão, SP.

de Melo, P. O. V., Viana, A. C., Fiore, M., Jaffrès-Runser, K., Mouël, F. L., Loureiro, A. A., Addepalli, L., and Guangshuo, C. (2015). Recast: Telling apart social and random relationships in dynamic networks. *Performance Evaluation*, 87:19 – 36. Special Issue: Recent Advances in Modeling and Performance Evaluation in Wireless and Mobile Systems.

Domingues, A. C. S. A., Silva, F. A., and Loureiro, A. A. F. (2018). Space and time matter: An analysis about route selection in mobility traces. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00958–00963.

Eagle, N. and Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268.

Kondor, D., Hashemian, B., de Montjoye, Y., and Ratti, C. (2018). Towards matching user mobility traces in large-scale datasets. *IEEE Transactions on Big Data*, pages 1–1.

Lau, C. P., Alabbasi, A., and Shihada, B. (2017). On the analysis of human mobility model for content broadcasting in 5g networks. In *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–7.

Lohan, E. S. and e Silva, P. F. (2017). User traces analysis based on crowdsourced data. In *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 1303–1308.

Montjoye, Y.-A., Hidalgo, C., Verleysen, M., and Blondel, V. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376.

Nunes, I. O., Celes, C., Nunes, I., Vaz de Melo, P. O. S., and Loureiro, A. A. F. (2018). Combining spatial and social awareness in d2d opportunistic routing. *IEEE Communications Magazine*, 56(1):128–135.

Oliveira, E. M. R., Viana, A., Naveen, K. P., and Sarraute, C. (2017). Mobile data traffic modeling: Revealing temporal facets. *Computer Networks*, 112:176–193.

Shah, A., Belyaev, P., Ferrer, B. R., Mohammed, W. M., and Lastra, J. L. M. (2017). Processing mobility traces for activity recognition in smart cities. In *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, pages 8654–8661.

Thilakarathna, K., Viana, A. C., Seneviratne, A., and Petander, H. (2017). Design and analysis of an efficient friend-to-friend content dissemination system. *IEEE Transactions on Mobile Computing*, 16(3):702–715.

Tsai, T. and Chan, H. (2015). Nccu trace: social-network-aware mobility trace. *IEEE Communications Magazine*, 53(10):144–149.

Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., and Liu, C. (2018). Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE Communications Magazine*, 56(3):142–149.