

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

José Mauro Ribeiro

**Caracterização e análise de redes sociais a
partir de dados de telefonia móvel**

São João del-Rei

2021

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

José Mauro Ribeiro

**Caracterização e análise de redes sociais a partir de
dados de telefonia móvel**

Dissertação apresentada como requisito para
obtenção do título de mestre em Ciências no
Curso de Mestrado do Programa de Pós Gra-
duação em Ciência da Computação da UFSJ.

Orientador: Vinícius da Fonseca Vieira

Universidade Federal de São João del-Rei – UFSJ

Mestrado em Ciência da Computação

São João del-Rei

2021

Ficha catalográfica elaborada pela Divisão de Biblioteca (DIBIB)
e Núcleo de Tecnologia da Informação (NTINF) da UFSJ,
com os dados fornecidos pelo(a) autor(a)

R83C Ribeiro, José Mauro.
 Caracterização e análise de redes sociais a partir
de dados de telefonia móvel / José Mauro Ribeiro ;
orientador Vinícius da Fonseca Vieira. -- São João
del-Rei, 2021.
 63 p.

 Dissertação (Mestrado - Ciência da Computação) --
Universidade Federal de São João del-Rei, 2021.

 1. Redes complexas. 2. Redes sociais. 3. Dados de
telefonia. I. da Fonseca Vieira, Vinícius, orient.
II. Título.

José Mauro Ribeiro

Caracterização e análise de redes sociais a partir de dados de telefonia móvel

Dissertação apresentada como requisito para obtenção do título de mestre em Ciências no Curso de Mestrado do Programa de Pós Graduação em Ciência da Computação da UFSJ.

Trabalho aprovado. São João del-Rei, 25 de junho de 2021:

Vinícius da Fonseca Vieira
Orientador

Alexandre Gonçalves Evsukoff
Convidado 1

Daniel Ludovico Guidoni
Convidado 2

Carolina Ribeiro Xavier
Convidado 3

São João del-Rei
2021

Agradecimentos

Este trabalho representa um marco para minha carreira, o qual alcancei com o suporte de algumas pessoas que sem as quais eu provavelmente não teria feito o mesmo caminho feliz que fiz. Dessa forma, manifesto aqui, a minha gratidão eterna a esses companheiros que hora foram amigos e hora foram mentores me inspirando e me guiando nessa jornada.

Primeiramente agradeço a ela, minha mãe, que ofereceu recursos que mais ninguém poderia me oferecer para chegar até aqui. São inúmeros os recursos que eu poderia sintetizar com palavras que creio que sejam simples e de fácil compreensão de todos: o amor incondicional. Agradeço aqui também às minhas irmãs por todo amor e cuidado recebidos e também a minha família.

Ao meu orientador, talvez me falem palavras mas expresso aqui a gratidão eterna por cada momento que soube oferecer a instrução correta, a inspiração necessária e até o acolhimento humano nos momentos mais difíceis que vivenciei nos últimos meses.

Agradeço aqui aos meus amigos que não só foram amigos mas também fonte de sabedoria e conhecimento. Aqui expresso a imensa gratidão por Gilma a quem conheci no início dessa jornada e me mostrou todo entusiasmo necessário para lidar com os desafios que enfrentaríamos juntos.

Aos membros da minha banca examinadora, agradeço por tão prontamente se disporem a colaborar com o trabalho.

Aos meus companheiros do projeto Kytos que me permitiram um crescimento profissional deixo minha eterna gratidão.

Por fim, agradeço a Fundação de Amparo a Pesquisa do Estado de Minas Gerais - Fapemig pelo apoio a este trabalho através da bolsa de estudos que recebi no meu primeiro ano.

"Seja qual for o rumo que tomarmos, nosso destino está indissoluvelmente ligado à
ciência."

Carl Sagan

Resumo

Neste trabalho são utilizados registros telefônicos de duas regiões imediatas localizadas no estado de Minas Gerais, no Brasil: a região imediata de São João del-Rei (de porte pequeno) e a região imediata de Juiz de Fora (de porte médio). Desta forma, este trabalho tem como objetivos principais a caracterização das interações sociais a partir de registros de ligações telefônicas e a investigação da relação entre as interações sociais e os padrões de mobilidade exibidos por indivíduos em centros urbanos. A complexidade urbana analisada neste trabalho sob duas perspectivas: social e de mobilidade. A perspectiva social permite que se compreenda como se dá a interação entre pares de indivíduos, revelando aspectos importantes específicos sobre as relações humanas em contextos urbanos. Sob a perspectiva de mobilidade é possível compreender como os indivíduos interagem com o espaço urbano e o que pode auxiliar na caracterização da organização das cidades e dos indivíduos. Com a finalidade de investigar a relação entre as relações sociais e os padrões de mobilidade, os experimentos realizados tiveram como objetivo explorar o impacto de indivíduos que possuem relação social relevante nos padrões de mobilidade de outros indivíduos. Dessa forma foram realizadas análises quanto aos padrões de mobilidade a fim de medir a similaridade desses padrões entre indivíduos. A caracterização de cada rede também foi realizada para permitir uma compreensão das mesmas seguindo métricas básicas, além de estudos das características mais complexas como reciprocidade e detecção de comunidades da rede, que são atualmente métricas de alto interesse no campo de estudo de redes sociais. Dessa forma, este trabalho fornece um conjunto rico de informações no contexto de redes sociais para regiões de forma inédita.

Palavras-chaves: Redes Complexas. Redes Sociais. CDR. Reciprocidade.

Abstract

Human communication has undergone a drastic change with the emergence of technologies that have removed a barrier for great distances between people. These days, this evolution allows them to connect in such diverse ways that many devices can combine text, image and sound into messages that redefine social relationships. Furthermore, in recent years, new technologies for storing and processing large volumes of data allow the information collected in real time to be used for the development of innovative solutions involving algorithms and methodologies that are still the focus of research in many areas . The call data records or CDRs (Call Detail Record) that arise in the context of large volumes of available data, incorporate a huge amount of information and if the amount of data searches using complex networking tools. Information on how, when, and with whom customers communicated at any given time, plus the approximate location of the same valuable data that can be picked from CDRs. Based on data, studies have emerged proposing models of social networks and mobility analysis in recent years, offering new possibilities and a search for innovations. This work uses telephone records from two immediate regions in the state of Minas Gerais, Brazil: the immediate region of São João del-Rei (small size) and the immediate region of Juiz de Fora (medium size). Thus, this work has as main objectives the characterization of social interactions from phone calls records and an investigation of the relationship between social interactions and mobility patterns by reference in urban centers. Urban complexity analyzed in this work from two perspectives: social and mobility. A social perspective allows us to understand how an interaction between pairs of products takes place, revealing specific aspects about human relationships in urban contexts. From the perspective of mobility, it is possible to understand how owners interact with the urban space and what can help to characterize the organization of cities and individuals. With the investigation of a relationship between social relationships and mobility patterns, the experiments carried out aimed to explore the impact of disabling those who have a relevant social relationship on the mobility patterns of other guests. Thus, analyzes regarding mobility patterns were evaluated in order to measure a pattern similarity between patterns. The characterization of each network was also carried out to allow an understanding of the basic norms, in addition to studies of more complex characteristics such as reciprocity and detection of network communities, which are currently metrics of high interest in the field of study of social networks. In this way, this work integrated a rich set of information in the context of social networks for regions in an unprecedented way.

Key-words: Complex Networks. Social Networks. CDR. Reciprocity.

Lista de ilustrações

Figura 1 – Região Imediata de Juiz de Fora	31
Figura 2 – Região Imediata de São João del Rei	32
Figura 3 – Diagrama de passos metodológicos aplicados	34
Figura 4 – Visão geral da metodologia. (A) Base de dados CDR. (B) Rede social. (C) Mobilidade. (D) Matriz de localização. (E) Vetor de trajetórias. (F) Similaridade de cosseno. (G) Experimento.	38
Figura 5 – Distribuição de quantidades. (a) Grau de saída (rede direcionada). (b) Grau de entrada (rede direcionada). (c) Grau (rede não direcionada). (d) Número de ligações (rede não direcionada).	42
Figura 6 – Exemplo de uma amostra da rede usando <i>snowball sampling</i> com seis saltos (o vértice em preto é o ego).	44
Figura 7 – Histograma da distribuição de reciprocidade. Cada classe de díades está representada reciprocamente como: recíprocas na cor azul, parci- almente recíprocas na cor vermelha e não recíprocas na cor laranja (a) Histograma para rede SJDR. (b) Histograma para rede JF.	47
Figura 8 – Distribuição da reciprocidade na rede.	47
Figura 9 – Similaridade da mobilidade X <i>rank</i> dos contatos sociais. (a) Número de ligações. (b) Reciprocidade.	48
Figura 10 – Similaridade da mobilidade X similaridade na rede social. (a) Número de saltos. (b) Coeficiente de Jaccard.	49

Lista de tabelas

Tabela 1	–	Informações Básicas das Redes da Região Imediata de São João del Rei.	41
Tabela 2	–	Informações Básicas das Redes da Região Imediata de Juiz de Fora. . .	41
Tabela 3	–	Comparação das comunidades encontradas utilizando localização nós com as comunidades encontras e pelos métodos DEMON e Louvain. . .	44
Tabela 4	–	Porcentagens em cada classe de reciprocidade	46

Sumário

1	Introdução	10
1.1	Objetivos	12
1.2	Justificativa	12
1.3	Organização do texto	13
2	Revisão Bibliográfica	14
2.1	Redes Sociais	14
2.2	O experimento de mundo pequeno de Stanley Milgran	19
2.3	Snow balling	20
2.4	Reciprocidade	22
2.5	Comunidades	24
2.6	Dados de CDR	28
3	Materiais e métodos	30
3.1	A Divisão regional do Brasil	30
3.2	Descrição dos dados de CDR	31
3.3	Metodologia	34
3.3.1	Metodologia para caracterização das interações sociais	35
3.3.2	Metodologia para investigação da relação entre interações sociais e padrões de mobilidade	37
3.3.3	Ambiente computacional e ferramentas	38
4	Experimentos e discussão	40
4.1	Análise das relações sociais	40
4.1.1	Caracterizações básicas	40
4.1.2	Divisão das redes sociais em comunidades	41
4.1.3	Explorando a descoberta de comunidades na Região Imediata de São João del Rei	42
4.1.4	Reciprocidade	45
4.2	Análise da relação entre aspectos sociais e mobilidade	48
5	Conclusão e direções futuras	52
	Referências	54

1 Introdução

O crescimento da população em áreas urbanas tem imposto um grande desafio para planejadores urbanos e responsáveis por tomadas de decisão de políticas públicas, o que pode trazer enormes desigualdades, problemas sociais e de saúde pública, caso aconteça sem um planejamento adequado. A compreensão da organização das cidades é, então, fundamental para um melhor planejamento urbano e o estabelecimento de políticas públicas para que seja possível antecipar e mitigar problemas sociais na escala em que ocorrem em grandes cidades ([BROECK, 2011](#); [FARBER, 2015](#)).

Modelos computacionais capazes de integrar e processar dados de múltiplas origens podem auxiliar a caracterização da organização geográfica e socioeconômica da população com grande resolução e aplicabilidade em contextos práticos ([BLONDEL, 2015](#); [GONZALEZ, 2008](#)).

A comunicação humana sofreu uma drástica mudança com o surgimento de tecnologias que removeram a barreira imposta por grandes distâncias entre as pessoas. Nos dias de hoje, essa evolução permite que indivíduos se conectem de formas tão diversas, que muitos dispositivos podem combinar texto, imagem e som em mensagens que ajudam a redefinir as relações sociais. Mais ainda, nos últimos anos, novas tecnologias para armazenamento e processamento de grandes volumes de dados permitiram que a informação coletada em tempo real seja utilizada para o desenvolvimento de soluções inovadoras para cidades baseadas em dados, o que envolve algoritmos e metodologias que são, ainda, foco de pesquisa em muitas áreas ([ALESSANDRETTI, 2017](#); [LENORMAND](#); [RAMASCO, 2016](#)).

As estruturas que se formam a partir das relações entre pessoas, também chamadas de redes sociais, sofrem um efeito direto das mudanças que a tecnologia oferece. Essas redes são objetos de interesse de estudos desde 1930 ([NEWMAN, 2018](#)) e o desenvolvimento tecnológico que provê novas formas de interação entre pessoas, também permite a obtenção de um grande volume de dados relativos a essas interações. O estudo dessas redes sociais tem oferecido novas formas de entender diversos fenômenos sociais, por exemplo, a propagação de influência. Estruturas sociais. Essas redes são geralmente constituídas por um conjunto de elementos que representam as pessoas, também conhecidos como *nós*, e um outro conjunto de elementos conhecidos como arestas que representam a relação entre as pessoas.

Este problema está na base de um outro problema mais geral de grande interesse, conhecido como Problema da Maximização de Influência (PMI). O PMI tem como objetivo buscar, em uma rede, nós candidatos que otimizem a difusão de algum elemento (algum produto, ideia entre outros), tendo diversas aplicações, por exemplo, no *marketing* viral

(MURATA; KOGA, 2018). Além disso, semelhante ao processo de propagação de influência é possível, a partir de uma rede social, tentar entender como uma epidemia se espalha, o que tem grande utilidade na definição de estratégias para contê-la (ALTARELLI, 2014). O estudo de redes sociais também tem sido aplicado em outras áreas emergentes, como a análise forense para detecção de fraudes eleitorais (FERNÁNDEZ-GRACIA; LACASA, 2018) e o estudo de mobilidade urbana, sendo este último um dos tópicos de estudo deste trabalho.

O estudo da mobilidade urbana encontra seu lugar junto às redes sociais visto que as tecnologias de comunicação se tornaram ubíquas, permitindo que praticamente qualquer pessoa se comunique praticamente em qualquer lugar usando um dispositivo eletrônico. De fato, dentre os diversos dispositivos, os celulares se revelaram uma fonte de dados bastante rica. Os registros de dados de chamadas ou CDR (*Call Detail Records*), necessários para as operadoras de telefonia móvel para fins de faturamento, contêm uma enorme quantidade de informação.

Nos últimos anos surgiram pesquisas baseadas na análise de CDR que, na área de redes complexas, foi apresentado inicialmente como um tópico secundário. Entretanto, logo este assunto tornou-se um campo inteiro de pesquisa e foi durante alguns anos o principal tópico de diversas conferências, como a NetMob, uma conferência internacional sobre a análise de conjuntos de dados de telefones celulares. Dessa forma, é importante ressaltar que existem algumas restrições quanto à disponibilidade de algumas categorias de dados de CDR e sobre as aplicações que venham a ser projetadas com os mesmos. No Brasil por exemplo a Lei Geral de Proteção de Dados Pessoais (LGPD) impede que seja feito um uso indiscriminado de dados pessoais, o que contempla CDRs.

A primeira delas diz respeito ao conteúdo das comunicações, pois o mesmo não deve ser registrado pela operadora ou estar acessível a qualquer terceiro. Uma segunda restrição determina que pesquisadores não tenham o mesmo acesso que as operadoras de telefonia móvel têm sobre os dados relativos aos clientes, dependendo das políticas de privacidade da operadora e das leis de proteção de privacidade que se aplicam no país em questão. Nomes e números de telefone nunca devem ser transmitidos para terceiros e em alguns países, os dados de localização devem ser mantidos confidenciais (BLONDEL, 2015). Assim, é possível afirmar que, essas bases de dados, embora possam colocar uma série de questões relacionadas à ética e privacidade, constituem uma inegável fonte de informação para a compreensão de fenômenos espaciais de maneira sem precedentes.

Utilizando dados CDR, a complexidade urbana representada neste trabalho é considerada sob duas perspectivas: social e de mobilidade. A perspectiva social que se utiliza de uma rede social de ligações permite que se compreenda como se dá a interação entre pares de indivíduos, revelando aspectos importantes específicos sobre as relações humanas em contextos urbanos. Sob a perspectiva de mobilidade é possível compreender melhor

como os indivíduos interagem com o espaço urbano e o que pode auxiliar na caracterização da organização das cidades e dos indivíduos.

A organização humana em áreas urbanas pode ser investigada através da exploração de plataformas para geração e armazenamento de dados massivos com diferentes graus de resolução e originados de diferentes fontes. Por exemplo, *e-mails*, mensagens em redes sociais *online*, como Twitter e Facebook, e registros de ligações telefônicas podem evidenciar relações sociais (BLONDEL, 2015; BARBOSA, 2018). *Check-ins* por aplicativos de avaliação de serviços e sistemas de rastreamento revelam a forma como indivíduos transitam nas cidades (SAPIEZYNSKI, 2015). Apesar de haver um grande número de trabalhos na literatura com o objetivo de capturar e modelar aspectos espaciais e temporais dos indivíduos no espaço urbano (GONZALEZ, 2008; BARBOSA, 2018), que podem melhorar significativamente nossa compreensão sobre a forma como indivíduos utilizam o espaço urbano, como tem sido apontado por diversos autores (GRABOWICZ, 2014; DOMENICO, 2013; CORNACCHIA, 2020; TOOLE, 2015).

1.1 Objetivos

Este trabalho tem como objetivos principais a caracterização das interações sociais a partir de registros de ligações telefônicas e a investigação da relação entre as interações sociais e os padrões de mobilidade exibidos por indivíduos em centros urbanos.

Para que os objetivos principais sejam alcançados, tem-se os seguintes objetivos específicos:

- Modelar as redes sociais a partir de dados de registros de chamadas telefônicas;
- Caracterizar as redes obtidas de acordo com medidas básicas frequentemente consideradas;
- Investigar o processo de formação de comunidades pela localização por algoritmos capazes de identificar as comunidades automaticamente;
- Caracterizar a reciprocidade das relações sociais presentes nas redes;
- Comparar a os padrões de mobilidade a partir da representação matemática que reflita esse padrões.

1.2 Justificativa

Os estudos envolvendo modelagem e análise de redes sociais apresentam um grande potencial por sua capacidade de oferecerem ferramentas de análise de redes complexas que

buscam extrair algum entendimento do mundo real. Dessa forma, este trabalho se coloca como um estudo que aplica essas ferramentas para entender e compreender aspectos das Regiões Imediatas de Juiz de Fora e São João del Rei que até então não foram explorados na literatura do ponto de vista de análise de dados de telefonia móvel. O principal benefício desse estudo é servir como base para estudos posteriores oferecendo instâncias de redes modeladas a partir de dados não explorados anteriormente. Portanto, as informações obtidas se colocam como uma oportunidade para olhar para essas regiões através do ponto de vista da área de redes complexas.

1.3 Organização do texto

Esse trabalho está organizada seguinte forma. O Capítulo 2.1 faz uma revisão sobre o tema de redes sociais no contexto de redes complexas e descreve as principais fontes (entrevistas, registros telefônicos e outros) utilizadas para criar redes sociais. Além disso, este define o conceito comunidades e os tópicos relacionados a esse assunto que são relevantes para o desenvolvimento deste trabalho. Por fim, este capítulo oferece uma visão e definição sobre o que são os dados de registros telefônicos que são utilizados no processo de modelagem. O Capítulo 3 inicia a com a descrição da divisão regional do Brasil que foi utilizada para os experimentos envolvendo comunidades. O capítulo também descreve os dados de registros telefônicos que foram utilizados para modelar as redes sociais do trabalho. O Capítulo se encerra com descrição da metodologia do trabalho. O Capítulo 4 descreve os experimentos realizados e a discussão para os resultados obtidos. O Capítulo 5 encerra o trabalho com as considerações sobre todos os processos desenvolvidos e as implicações dos resultados encontrados nos experimentos.

2 Revisão Bibliográfica

Este capítulo tem como objetivo oferecer um contexto histórico para a área de estudo de redes sociais. Esse contexto histórico é apresentado pontuando tópicos que permitiram o desenvolvimento de estudos na área. Além disso, o experimento de mundo pequeno, que é um marco no estudo de redes complexas, é descrito a fim de estabelecer a ideia por trás do conceito de mundo pequeno. O capítulo descreve o processo de *Snow Balling* que é uma técnica que permite explorar populações de indivíduos a fim de entender processos relacionados aos estudos de redes complexas. Também é apresentada a definição de reciprocidade que é um aspecto explorado nas redes estudadas neste trabalho. A definição de comunidades e o problema derivado dessa definição também são descritos neste capítulo. Por fim, é realizada uma descrição dos dados CDR utilizados neste estudo.

2.1 Redes Sociais

Embora haja análises de redes sociais na literatura desde o início do século dezenove, as origens dos estudos empíricos e quantitativos em redes sociais são atribuídas ao psiquiatra Jacob Moreno, que por volta dos anos 1930 se tornou interessado na dinâmica das relações sociais entre pessoas (NEWMAN, 2018). No trabalho de Moreno figuras desenhadas a mão, denominadas por ele de sociogramas representavam os padrões de interação entre vários grupos de pessoas sendo equivalentes ao que chamamos de redes sociais atualmente (NEWMAN, 2018). O sociograma, cuja invenção é atribuída a Moreno, de fato marcou o início da sociometria, sendo esta considerada por muitos como precursora da análise de redes sociais e muito da psicologia social (WASSERMAN, 1994). Para Wasserman (WASSERMAN, 1994) outros autores podem ser apontados, junto a Moreno, como pioneiros na análise de redes sociais. Dessa forma, através dos trabalhos desses autores, Wasserman evidencia que as origens dos estudos de redes sociais se encontram em áreas como psicologia e antropologia. Além disso, Wasserman alega que muitas pessoas atribuem o primeiro uso do termo *social network* ao trabalho *Class and committees in a Norwegian island parish* (BARNES, 1954).

Embora a análise de redes sociais tenha suas origens em mais de uma área de estudo, a perspectiva da mesma compreende em teorias, modelos e aplicações distinguindo-a de outras abordagens (WASSERMAN, 1994). No processo de evolução desse campo de estudo, os sociólogos desenvolveram termos próprios para discutir as redes sociais, onde as entidades sociais são denominadas como atores e as conexões entre eles como laços (WASSERMAN, 1994). Sendo assim, a princípio os atores são unidades sociais distintas e individuais, já os laços são uma conexão entre duas unidades também chamada de díade.

Todavia, a ideia de laço pode ser estendida para um número maior de atores como três (triáde) ou mais (subgrupos e grupos) (WASSERMAN, 1994). A literatura mais recente comumente se refere aos laços como arestas e aos atores como nós ou vértices. Isto ocorre devido à consolidação da representação de uma rede social utilizando um grafo que é uma estrutura composta por nós e arestas. A definição de grafo é oriunda da Teoria de Grafos, um ramo da matemática que estuda as relações entre os objetos de um determinado conjunto.

Um fato importante sobre redes sociais é que existem muitos significados possíveis do que uma aresta pode representar e cada definição está diretamente relacionada com as perguntas de interesse em um estudo. Essas arestas podem representar uma amizade, relações de trabalho, trocas de bens e serviços ou dinheiro, padrões de comunicação, relações românticas e sexuais ou muitos outros tipos de conexões (NEWMAN, 2018). Além disso, as técnicas usadas para explorar múltiplos tipos de interações sociais podem ser um pouco diversas, sendo assim estudos diferentes podem contemplar diferentes questões.

É possível também utilizar registros para fazer o questionamento direto de indivíduos com o intuito de se fazer um estudo de redes complexas, técnica que já em 1939. Em (GARDNER, 1942) foi aplicada por um grupo de etnógrafos para estudar a estratificação e classes sociais no sul dos Estados Unidos da América. Eles coletaram dados da participação de dezoito mulheres em eventos sociais em uma pequena cidade no Mississippi durante nove meses. Os dados foram coletados a partir da lista de convidados de eventos e relatos em páginas sociais de jornais, embora eles tenham obtido dados a partir de algumas entrevistas e relatos diretos de participantes. Esse estudo também conhecido como "O estudo das mulheres sulistas" é amplamente discutido e estudado na literatura de redes sociais desde sua publicação. Cada mulher nesse caso pode ser representada como um vértice e as arestas conectando mulheres que participaram de um mesmo evento, ou ainda uma representação mais completa como uma rede bipartida (ou rede de afiliação) conectando cada mulher aos seus respectivos eventos. Uma das razões que tornou esse estudo famoso, foi que os pesquisadores encontraram dois grupos fortemente conectados na rede formada por essas mulheres, possibilitando o uso da mesma em estudos de algoritmos que busquem agrupamentos em redes.

O processo de fazer perguntas a pessoas para obter dados de redes sociais, pode ser realizado questionamentos diretos ou através de formulários que podem ser realizados eletronicamente ou por papel. Além disso, estudos modernos que fazem uso do telefone combinam entrevistas e questionários, onde um entrevistador profissional faz a leitura do questionário para o participante. Com o uso de um questionário, os idealizadores do estudo garantem que perguntas específicas sejam feitas, com uma certa precisão, em uma ordem específica e com formulação específica. Quando um entrevistador é utilizado no estudo, este ganha flexibilidade e confiabilidade, visto que os entrevistados levam o estudo mais a

sério quando as perguntas são feitas por um ser humano, além da oportunidade propiciada ao entrevistador de tentar esclarecer qualquer resposta confusa que seja dada. Estas considerações são importantes dado que respostas confusas e inconsistentes são fontes substanciais de erros em questionários (MARSDEN, 1990). Rea e Parker (REA; PARKER, 2014) oferecem uma boa introdução de como projetar e implementar questionários deste tipo.

Um recurso que auxilia nas medidas quantitativas das pesquisas em redes sociais são conhecidos como *name generator* ou gerador de nomes. Os geradores de nome são questionários que solicitam que o respondente liste pessoas que ele tenha contato de algum tipo específico. Esse tipo de questionário pode levar a identificação de relações por meio de um processo assimétrico. Nesse caso as relações dependem que dois indivíduos se identifiquem como amigos para que a relação de amizade seja estabelecida entre os dois. Dessa forma, os dados obtidos por geradores de nomes são comumente representados como uma rede direcionada.

Outro ponto que pode ser visto sobre geradores de nome é que muitas vezes eles limitam o número de nomes a serem listados por cada respondente, sendo chamados de estudos de escolha fixada. Quando nenhum limite é imposto, o nome dado é estudo de escolha livre. Esses limites podem ser impostos para reduzir o trabalho do pesquisador, bem como facilitar o entendimento dos respondentes sobre o que é solicitado deles, evitando assim que um respondente nomeie um número grande de amigos próximos. Entretanto, limites como esses são geralmente indesejáveis por limitar o grau de saída dos nós na rede impondo um corte provavelmente não realístico e artificial. O que pode levar a impossibilidade de encontrar *hubs* que podem ser determinantes no entendimento do comportamento da rede.

As pesquisas ainda podem refinar as perguntas para se obter informações mais precisas. Além de nomear pessoas com quem um respondente tenha uma determinada relação, a pergunta ainda pode pedir para que o mesmo descreva a relação obtendo uma descrição mais rica. As perguntas ainda podem ter como objetivo mensurar a força dos laços, perguntando, por exemplo, para cada indivíduo nomeado, quantas interações ocorreram ou por quanto tempo.

Nem todo estudo que faz uso de geradores de nome dá origem a uma rede direcionada. Em alguns casos há interesse em laços que são intrinsecamente simétricos entre as duas partes envolvidas gerando uma rede não direcionada, por exemplo, em redes de contato sexual (KLOVDAHL, 1994; LILJEROS, 2001; POTTERAT, 2002; ROTHENBERG, 2001). As principais desvantagens de estudos que são baseados em questionários é o fato de serem trabalhosos e imprecisos. Mesmo com o uso de computadores o número de entrevistados ainda tem sido limitado a algumas dezenas ou no máximo centenas. Há também o problema da subjetividade das respostas que não é exclusivo de estudos de

rede. Praticamente todas as pesquisas sociais sofrem de tais problemas e um conjunto de técnicas para lidar com eles foi desenvolvido (MARSDEN, 1990) (REA; PARKER, 2014). Portanto, deve-se ter em mente, ao se tratar de qualquer rede social derivada de entrevistas ou questionários, a possibilidade de viés experimental nos dados.

Além desse tipo de estudos há também os que são conhecidos como estudos de redes pessoais ou de redes centradas no ego. Onde uma rede centrada no ego, é uma rede em torno de um determinado indivíduo, ou seja, é composta pelo indivíduo e seus vizinhos diretos. O indivíduo em questão é chamado de *ego* e cada vizinho de *alter*. Essas redes são construídas a partir de questionários ou entrevistas, ou ainda, uma combinação de ambos. E estudos que utilizam redes centradas no ego, não são capazes de revelar a estrutura completa da rede, visto que as mesmas dificilmente formam a estrutura da rede completa. Um exemplo de estudo que coleta dados de redes centradas no ego é o General Social Survey (GSS), uma pesquisa em grande escala conduzida todos os anos nos Estados Unidos começando em 1972 e a cada dois anos desde 1994 (BURT, 1984) cujo objetivo é reunir dados sobre a vida nos Estados Unidos, como ela está mudando e como ela difere ou se relaciona com a vida em outras sociedades.

Outro método para construir redes sociais é através da observação direta. Através da observação da interação de indivíduos por um período de tempo é possível construir uma rede onde os laços refletem a relação existente entre os indivíduos. Esse método tende a ser um método de estudo trabalhoso e assim é aplicado a grupos pequenos e preferencialmente a grupos onde há interação cara a cara em lugares públicos. Um exemplo clássico é o estudo da rede de caratê de Zachary (ZACHARY, 1977) composta por 34 membros de um clube de caratê construída a partir de ligações entre pares de membros que interagem fora do clube. Durante o estudo, surgiu um conflito entre o administrador e o instrutor levando a uma divisão do clube em dois. A partir dos dados coletados, Zachary foi capaz de atribuir corretamente todos os membros do clube, exceto um, aos grupos aos quais eles realmente se juntaram após a separação. Além de redes sociais de indivíduos da espécie humana há também estudos em que redes sociais de animais como macacos (FLACK, 2006), cangurus (GRANT, 1973) e golfinhos (CONNOR, 1999) e nestes casos a observação direta é o método mais viável.

Outra fonte de dados para construir redes sociais são os registros de arquivo. Esses registros de arquivo também permitem reconstruir redes que não existem mais, como redes do passado histórico. Um exemplo bem conhecido e em pequena escala de um estudo baseado em registros de arquivo é o trabalho de Padgett e Ansell sobre as famílias governantes de Florença no século XV (ACTION, 1993). Os atores dessa rede são famílias. Conforme foi observado, a família era uma importante unidade econômica e política, então a história da Florença do século quinze pode ser bem compreendida focalizando as famílias, ao invés de pessoas individuais. Nesse estudo o principal interesse

para eles era a associação entre as duas relações, casamento e negócios.

Além de registros históricos, registros de arquivo mais recentes como e-mail também são usados para construir redes sociais (EBEL, 2002; TYLER, 2003) que podem ser utilizadas para investigar padrões de troca de mensagens, entender como as informações se espalham por uma comunidade. Além de redes construídas a partir de dados de e-mail, redes similares podem ser construídas a partir de padrões de mensagens de texto ou mensagens instantâneas usando dados de telefones celulares (ONNELA, 2011; STOPCZYNSKI, 2014). Em uma rede construída a partir de dados de telefonia, os nós representam números de telefone e as arestas entre eles representam ligações de um número para outro. Os dados de telefonia normalmente são registros de chamadas telefônicas mantidos por companhias que deram origem a uma série de estudos realizados nos últimos anos, incluindo alguns em escalas maiores, com milhares ou mais de números de telefone (BLONDEL, 2015; RATTI, 2010; ONNELA, 2007; HIDALGO; RODRÍGUEZ-SICKERT, 2008). Uma das razões que tornam os dados telefones celulares atraentes se deve ao fato de que os mesmos podem revelar não apenas quem liga para quem, mas também a localização geográfica potencial dos usuários do telefone. Por fim, além das fontes de dados já mencionadas, também existem dados oriundos de serviços de rede social online, como Facebook e LinkedIn que geram uma grande quantidade de informação quanto às conexões existentes entre seus usuários. O Twitter, por exemplo, tornou seus dados disponíveis publicamente, permitindo que os pesquisadores estudassem as redes correspondentes (DODDS, 2011; GOLBECK, 2010; GONÇALVES, 2011). Entretanto, alguns dados não disponíveis publicamente deram origem a estudo de casos publicados pelas próprias empresas como por exemplo o Facebook, através de departamentos de pesquisa internos (ARAL; WALKER, 2012; UGANDER, 2011; BOND, 2012).

Algumas comunidades online ligeiramente diferentes das comunidades das redes já mencionadas podem ser estudadas usando técnicas de rede. Há estudos investigando as interações entre usuários de *sites* de namoro online, blogs, periódicos entre outros (HOLME, 2004; HOLME, 2005). Nesses *sites* o autor geralmente usa de *links* para outros sites como forma de referenciar outros materiais. Desse modo, é possível obter uma rede direcionada a partir dos *sites* e *links*.

Existe um caso especial de redes criadas a partir de registros, denominada redes de associação. Nessas redes os atores são conectados a partir da sua participação em algum tipo de grupo ou mesmo evento. (NEWMAN, 2018) Nesse tipo de rede uma representação mais completa possível é feita com dois tipos de nós representando os atores e grupos, e arestas ligando os dois tipos de nós. Diversos exemplos podem ser encontrados na literatura como o estudo de CEOs de companhias de Chicago de 1970 e as suas interações sociais em clubes, onde os cada CEO era um ator e os clubes eram os grupos (GALASKIEWICZ, 2016).

2.2 O experimento de mundo pequeno de Stanley Milgran

Na década de 1960, o psicólogo Stanley Milgran realizou experimentos que trariam contribuições importantes para a área de redes sociais. Esses experimentos, também conhecidos como experimentos de "mundo pequeno", tinham como objetivo quantificar a distância entre os atores nas redes sociais ([MILGRAM, 1967](#); [TRAVERS; MILGRAM, 1977](#)). Milgran queria testar em condições do mundo real, argumentos que defendiam a distância em dois nós quaisquer deveria ser algo pequeno em uma rede. Para tanto, Milgran enviou 96 pacotes para participantes voluntários em Omaha, no estado de Nebraska. Cada pacote continha instruções para serem entregues a um amigo de Milgran que morava em Boston no estado de Massachusetts que incluíam o nome do amigo, endereço e a sua ocupação como corretor da bolsa. Os participantes não podiam simplesmente enviar o pacote para o endereço fornecido, em vez disso, eles foram solicitados a enviá-lo para alguém que conhecia pelo primeiro nome, mais especificamente para a pessoa que eles achavam que teria uma maior chance de realizar a entrega ao alvo pretendido. Assim, eles podem decidir enviá-lo para alguém conhecido que deveria realizar as mesmas instruções.

O resultado obtido foi que 18 pacotes dos 96 pacotes enviados chegaram ao destino na corretora de valores em Boston. Cada etapa do caminho percorrido foi registrada no passaporte que estava dentro do pacote como foi solicitado por Milgran aos participantes para que ele soubesse quantos passos foram necessários. Milgran descobriu que o comprimento médio dos caminhos concluídos de Omaha até o destino em Boston era de apenas 5,9 passos, dando origem à ideia dos seis graus de separação comumente aceita de que existem apenas cerca de seis etapas entre duas pessoas quaisquer no mundo.

Mesmo com tendo muitas ressalvas em relação a problemas existentes no experimento, a conclusão fundamental de que pares de nós em redes sociais tendem em média, a serem conectados por caminhos curtos, agora é amplamente aceita. O resultado foi confirmado em muitos casos, incluindo para algumas redes sociais muito grandes como a rede de amigos no Facebook ([UGANDER, 2011](#)). Além disso, foi demonstrado que se estende a muitos outros tipos de redes também além de redes sociais levando a conclusão que resultado geral não deva ser seriamente questionado.

Os experimentos de Milgram também revelaram algumas outras características interessantes de redes de conhecidos, como por exemplo, o fato de que uma grande fração das conexões do destino do pacote com o mundo exterior parecia acontecer através de alguns de seus conhecidos, um fenômeno às vezes referido como afunilamento. Outro corolário interessante do experimento de Milgram, nunca mencionado pelo próprio Milgram, foi destacado muitos anos depois por Kleinberg ([KLEINBERG, 2000a](#); [KLEINBERG, 2000b](#)) que é possível e de fato comum que uma rede possua caminhos curtos entre os nós, porém difíceis de encontrar, a menos que se tenha informações completas sobre a estrutura de toda a rede, o que os participantes dos estudos de Milgram não tinham. Recentemente,

o experimento do mundo pequeno foi repetido por Dodds *et al.* (DODDS, 2003) usando o meio moderno de e-mail. Nesta versão do experimento, os participantes encaminharam mensagens de e-mail para seus conhecidos em um esforço para levá-los a uma pessoa-alvo específica sobre a qual foram informados alguns fatos básicos. Os resultados básicos foram semelhantes aos de Milgram com um comprimento médio de pouco mais de quatro etapas. Por causa de seus melhores dados e análises estatísticas mais cuidadosas, os autores também foram capazes de compensar os vieses devido a cadeias inacabadas e estimaram que o verdadeiro comprimento médio do caminho para o experimento estava em algum lugar entre cinco e sete etapas que é muito semelhante ao resultado de Milgram.

2.3 Snow balling

Estudos de algumas populações enfrentam um problema especial devido a dificuldade de identificar os membros dessas populações. Usuários de drogas ou mesmo imigrantes ilegais são exemplos de indivíduos que tendem a evitar serem encontrados e identificados devido suas condições. A técnica começa com a identificação de um membro inicial da população de interesse e realiza uma entrevista com o mesmo. Após ganhar a confiança do entrevistado, você o convida a nomear outros conhecidos que sejam membros da população alvo. Em seguida, os mesmos passos são realizados para cada conhecido apontado e por meio de uma sucessão de ondas de amostragem a técnica permite ir revelando os membros da população. O processo ocorre de forma semelhante a formação de uma bola de neve que dá origem ao nome em inglês da técnica *Snow balling*.

Esta técnica se mostra como a melhor maneira de encontrar uma população que está oculta em comparação com as pesquisas, uma vez que os indivíduos nomeados provavelmente são membros da população. Além disso, o fato do entrevistador ser apresentado aos outros possíveis membros por um de seus conhecidos faz com que seja mais provável que eles se disponham a serem entrevistados.

No entanto, a técnica apresenta alguns problemas, como por exemplo, fornecer amostras altamente tendenciosas. Apesar disso, a técnica é suficientemente útil para encontrar populações que de outra forma seriam difíceis de encontrar e vem sendo amplamente utilizada em estudos nas últimas décadas. Para populações alvo pequenas, algumas ondas de amostragem podem encontrar essencialmente todos os membros de uma população local, caso especial em que a técnica é capaz de revelar a estrutura da rede.

Há uma variante da técnica de amostragem *snowballing* que lida, até certo ponto, com os problemas de viés na amostra. A técnica é chamada de amostragem *random-walk* (KLOVDAHL, 1989; THOMPSON; FRANK, 2000). A técnica inicia igual ao *snowballing* e novamente com um único membro de uma comunidade a entrevista é feita para determinar seus contatos. Entretanto, ao invés de rastrear todos os contatos do entrevistado, é feita a

escolha aleatória dentre um deles para ser realizada a próxima entrevista. Se a pessoa em questão não for encontrada ou se recusar a ser entrevistada, escolhe-se outro conhecido e a etapa é repetida.

Um detalhe importante é que deve-se determinar todos os contatos de cada indivíduo, embora na maioria das vezes apenas um deles seja procurado. Isso é necessário para que a escolha do próximo entrevistado seja realmente aleatória entre o conjunto completo de contatos. A vantagem do método é que a probabilidade de amostragem assintótica de nós em um passeio aleatório é simplesmente proporcional ao grau do nó. Além disso, o regime assintótico nessa técnica, ao contrário da amostragem em *snow balling*, é alcançado muito rapidamente para tamanhos de amostra relativamente pequenos. Sabendo disso, e dado que determinamos o grau como parte do processo de entrevista, é possível compensar o viés de amostragem por meio de uma ponderação adequada dos resultados e fazer estimativas populacionais de quantidades em uma forma que é, pelo menos em teoria, imparcial. Dessa forma, a amostragem *random-walk* é uma melhoria significativa na amostragem de *snow balling* e a sua principal desvantagem é que é relativamente lento. Uma vez que os participantes só podem ser entrevistados em série e não em ondas de entrevistas paralelas. Esse problema pode ser contornado até certo ponto com várias caminhadas curtas aleatórias em paralelo em vez de uma longa, mas as caminhadas não podem ser muito curtas ou não atingirão o regime assintótico em que a amostragem é proporcional ao grau. Além disso tudo, a técnica é boa para explorar a estrutura da própria rede uma vez que ela retorna uma cadeia linear de contatos.

Em alguns casos, é considerado antiético fazer com que os participantes nomeiem seus contatos, especialmente quando o tópico de um estudo é de legalidade duvidosa, e a permissão para a realização de tais estudos pode ser negada pelas autoridades. Para contornar este problema, pode-se fazer uso da amostragem dirigida por respondentes em inglês *respondent-driven sampling* (SALGANIK; HECKATHORN, 2004). Nesta técnica, os participantes são geralmente os voluntários e a inscrição é feita com a entrega de ingressos aos entrevistados. Para evitar pedir aos participantes que mencionem seus contatos, os entrevistados são simplesmente convidados a dar os ingressos aos seus conhecidos e informados de que eles e os conhecidos receberão o pagamento se o conhecido levar o ingresso ao pesquisador e concordar em participar do estudo. Desta forma, nunca se pede a ninguém que diga nomes e todos os participantes voluntariamente participaram ativamente. Acredita-se que essa técnica fornece amostras populacionais superiores à amostragem em *snow balling* e é o método de escolha para estudos em que não se pode pedir às pessoas que nomeiem seus contatos.

2.4 Reciprocidade

Os estudos iniciais que investigavam a reciprocidade das relações entre dois indivíduos, tentavam verificar se ambos indivíduos da relação indicavam o outro como um amigo, ideia que ia de encontro a teoria de equilíbrio proposta Heider ([CHAWLA, 2011](#)). A teoria de Heider permitiu estabelecer uma definição inicial para entender as variações existentes em relacionamento em uma díade, isto é um par de atores ou indivíduos. Nessa teoria um estado (ou situação) era dito equilibrado quando este era um estado harmonioso, no qual as entidades que compõem a situação e os sentimentos a respeito delas se encaixam sem estresse ([HARMON, 1959](#)).

A investigação do equilíbrio em díades através de análises sociométricas e também em redes sociais deu início ao estudo de caracterização de relações recíprocas ([CHAWLA, 2011](#)). Questões importante como, Díades assimétricas são mais ou menos estáveis do que díades mútuas? e Uma a relação existente em díade assimétrica muda, ela se torna mútua ou ela se desfaz? foram investigadas utilizando dados sociométricos de alunos de uma escola primária revelando a existência de maior estabilidade nas díades mútuas em comparação com as não mútuas e uma tendência de relações assimétricas se desfazerem com o tempo ([HALLINAN, 1978](#)).

A reciprocidade pode ser estabelecida por meio do censo diádico a partir da definição da existência arestas com valores binários. O cálculo das quantidades de díades assimétricas, mútuas e nulas no conjunto total de díades existentes define uma tripla $\langle M, A, N \rangle$ conhecida como censo diádico. O nome censo diádico é dado devido ao fato de se percorrer todas as díades existentes no grafo a fim calcular os valores da tripla ([WASSERMAN, 1994](#)).

É possível entender que altos níveis de não reciprocidade ocorrem quando a proporção de díades assimétricas é maior do que a proporção que seria encontrada ao acaso, em uma rede com propriedades topológicas semelhantes na definição tradicional de reciprocidade ([CHAWLA, 2011](#)). Entretanto, essa definição carrega algumas limitações intrínsecas que não permitem que a mesma possa mensurar os níveis de reciprocidade de qualquer tipo de rede ([GARLASCHELLI; LOFFREDO, 2004](#)). A primeira limitação deriva do fato de se comparar o valor de reciprocidade da rede que está sendo estudada, com o valor de reciprocidade de uma rede topologicamente similar, essa comparação caracteriza o resultado encontrado como um resultado de significado relativo. Deste fato, deriva a segunda limitação, que é a não possibilidade de ordenar redes de acordo o seu valor de reciprocidade uma vez que estas, podem não possuir características topológicas similares. Por fim, se a rede possui laços, a definição pode gerar resultados inconsistentes uma vez que laços em rede nunca podem ocorrer em pares recíprocos.

Para transpor as limitações da definição tradicional de reciprocidade, ([GARLAS-](#)

(CHELLI; LOFFREDO, 2004) propuseram uma definição de reciprocidade baseada no coeficiente de correlação entre as entradas da matriz de adjacência de uma rede direcionada. A questão relativa à primeira e a segunda limitação se resolve uma vez que o coeficiente de correlação é uma medida absoluta. Além disso, o coeficiente não considera as posições da matriz que representam os laços resolvendo a terceira limitação mencionada. A medida de reciprocidade denominada por ρ expressa reciprocidade para valores maiores que zero, anti reciprocidade para valores menores que zero e um caso neutro para valores iguais a zero.

Já Chawla *et al.* (CHAWLA, 2011) buscaram definir uma métrica para reciprocidade a fim de mensurar o grau de equilíbrio em uma relação social. Com essa métrica, é possível analisar a distribuição da reciprocidade em um sistema de grande escala que possui arestas ponderadas. Redes ponderadas são típicas quando se trata de dados de CDR que possuem registros de durações de chamadas.

Para serem consistentes com a ideia de reciprocidade como equilíbrio, Chawla *et al.* buscaram atender as seguintes propriedades:

- Possuir um valor mínimo quando os arcos em direções opostas entre dois vértices, possuírem pesos com valores próximos;
- Deve aumentar monotonicamente com a diferença de peso entre os dois arcos direcionados;
- Deve normalizar o valor da diferença dos pesos para ajustar o fato de que algumas pessoas são simplesmente mais comunicativas do que outras (elas entram em contato com todos os seus parceiros com mais ou menos frequência);
- A medida deve ser a mesma, independentemente da direcionalidade ($R_{ij} = R_{ji}$).

Com base nessas propriedades os autores propuseram a seguinte definição:

$$R_{ij} = |\ln(p_{ij}) - \ln(p_{ji})|, \quad (2.1)$$

onde,

$$p_{ij} = \frac{w_{ij}}{w_{i+}}, \quad (2.2)$$

sendo que w_{ij} é o peso da aresta de i para j , e w_{i+} é a força de saída do i -ésimo vértice (BARRAT, 2004) isto é, o somatório dos pesos das arestas de saída.

Através dessa definição, Chawla *et al.* apontaram características estruturais e a tendência de formar relacionamentos com outras pessoas semelhantes, que ajudam ou

atrapalham a capacidade das pessoas de obter equilíbrio comunicativo em seus relacionamentos. Dessa forma este trabalho utiliza a fórmula 2.1 será utilizada como o objetivo de se ter uma medida de reciprocidade mais granular da rede modelada.

2.5 Comunidades

O conceito de comunidade em redes complexas é amplamente estudado por permitir análises que tentam entender a organização estrutural. A princípio, uma comunidade pode ser definida como um subgrafo conectado localmente denso em uma rede (BARABÁSI; PÓSFAL, 2016). Isto quer dizer, todos os membros de uma comunidade devem ser alcançados por meio de outros membros da mesma comunidade, estabelecendo uma conexão existente neste grupo. Além disso, também esperamos que os nós que pertencem a uma comunidade tenham uma probabilidade maior de se vincular aos outros membros dessa comunidade do que aos nós que não pertencem à mesma comunidade, estabelecendo uma característica de densidade no grupo. Embora essa descrição seja bem precisa no que se refere a uma definição de comunidade, ela não a define de maneira única.

O problema de detecção de comunidades tem estado sob intensa investigação nos últimos anos (LANCICHINETTI; FORTUNATO, 2009; ZHANG, 2007; JORA; CHIRA, 2016). Assim, têm surgido métodos e ferramentas de diversas disciplinas de física, biologia, matemática aplicada, ciência da computação e ciências sociais na busca pela solução, (LANCICHINETTI; FORTUNATO, 2009), bem como a utilização de diferentes abordagens (JORA; CHIRA, 2016). Nestas abordagens deve-se considerar que não existe restrição a tamanho de grupos como acontece em particionamento de grafos (BARABÁSI; PÓSFAL, 2016).

A detecção de comunidades pode ter aplicações concretas como agrupar clientes da Web que têm interesses semelhantes e estão geograficamente próximos uns dos outros pode melhorar o desempenho dos serviços prestados na *World Wide Web*, utilizando um servidor espelho dedicado para cada cluster de clientes (KRISHNAMURTHY; WANG, 2000). Identificar clusters de clientes com interesses semelhantes na rede de relacionamentos de compra entre clientes e produtos permite configurar sistemas de recomendação eficientes oferecendo oportunidades de negócios (REDDY, 2002). Há também a possibilidade de criar redes auto configuráveis onde, encontrar comunidades possibilitam gerar tabelas de roteamento compactas enquanto a escolha da comunicação por caminhos ainda é eficiente (STEENSTRUP, 2001). A detecção da comunidade se destaca também por permitir a classificação de vértices nas comunidades de acordo com sua posição. Vértices com uma posição central, ou seja, compartilhando um grande número de arestas com os outros parceiros do grupo, podem ter uma função importante de controle e estabilidade dentro do grupo. E vértices situados nos limites entre os módulos desempenham um

papel importante de mediação nas relações e trocas entre diferentes adjacentes (FORTUNATO, 2010). Encontrar vértices centrais tem importância significativa em redes sociais (GRANOVETTER, 1973; BURT, 1976; FREEMAN, 1977) e metabólicas (GUIMERA; AMARAL, 2005). Por fim, outra importância em encontrar comunidades é a possibilidade de estudar o grafo onde os vértices são as comunidades e as arestas são definidas quando houver conexões entre alguns de seus vértices no grafo original ou se as comunidades se sobrepõem.

Embora haja muitas abordagens diferentes para detecção de comunidades disponíveis, é recorrente compará-las para avaliar seus pontos fortes e fracos relativos. Uma das formas de se comparar, é avaliando a capacidade de um método para identificar as chamadas comunidades que são vistas no mundo real para uma dada partição dos nós da rede em grupos. Essa abordagem para avaliar métodos de detecção de comunidade funciona bem em redes geradas artificialmente, cujas arestas são colocadas explicitamente de acordo com essas comunidades de verdade terrestre e um processo de geração de dados conhecido. Por este motivo, a partição dos nós em comunidades que são vistas no mundo real, nas redes sintéticas são chamadas de partição plantada. Já para redes do mundo real, tanto as comunidades verdadeiras quanto o processo que as geram são desconhecidos, o que dificulta a avaliação objetiva. Normalmente, alguns dados observados nos nós de uma rede, que chamamos de metadados dos nós são tratados como se fossem comunidades de verdade (PEEL, 2017).

Considerar metadados dos nós como equivalentes comunidades de verdade para redes do mundo real é comumente justificado por meio de uma crença errônea de que o objetivo da detecção de comunidade é recuperar grupos que correspondem aos rótulos de metadados (YANG; LESKOVEC, 2012; YANG; LESKOVEC, 2015; HRIC, 2014). Consequentemente, a recuperação de metadados é frequentemente usada para medir o desempenho de detecção da comunidade, e os metadados são frequentemente considerados como uma verdade fundamental. No entanto, a organização de redes reais normalmente se correlaciona com vários conjuntos de metadados, tanto observados como não observados. Assim, rotular qualquer conjunto específico como verdade fundamental é uma decisão arbitrária e geralmente injustificada. Além disso, quando um algoritmo de detecção de comunidade falha em identificar as comunidades que correspondem aos metadados conhecidos existem três possibilidades que justificam isso. A primeira se deve ao fato de que os metadados podem ser irrelevantes para a estrutura da rede, a segunda que os metadados e as comunidades capturam diferentes aspectos da estrutura de rede e por fim pode ser que a rede não possua estrutura de grupo.

A definição mais simples de comunidade carrega algumas limitações como a dificuldade de quantificar a qualidade das comunidades encontradas. Nesse contexto, surge a necessidade quantificar a qualidade de cada partição e o conceito de modularidade. Por-

tanto, a modularidade nos permite decidir se uma partição de comunidade em particular é melhor do que outra.

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta_{c_i c_j}, \quad (2.3)$$

Valores de modularidade podem ser positivos ou negativos. Quando positivos, indicam que possivelmente é encontrada estrutura de comunidade (BARABÁSI; PÓSFAL, 2016). Valores no intervalo de 0.3 a 0.7, são mais comuns de serem encontrados, por outro lado, valores acima deste intervalo são raros (NEWMAN; GIRVAN, 2004). Desta forma, para (NEWMAN; GIRVAN, 2004) modularidade Q , avalia a qualidade de uma partição considerando a fração de arestas dentro da mesma comunidade com o número de arestas inseridas ao acaso. A modularidade pode ser descrita como:

onde pode-se verificar, que a probabilidade de uma aresta entre dois nós i e j é $k_i k_j / 2m$, onde k_i é o grau do nó i e m é o total de número de arestas na rede. Desta forma, A_{ij} é um elemento da matriz de adjacência, δ_{ij} é o símbolo do delta de Kronecher, e c_i é o rótulo da comunidade a qual o nó i é atribuída (LEICHT; NEWMAN, 2008). Neste contexto, não é considerado a direção de arestas.

Uma das abordagens possíveis é usar modularidade como um problema de maximização. Considerando um grafo não direcionado e não ponderado, é possível tratar como um problema de otimização, atribuindo uma pontuação para cada divisão possível da rede em comunidades (NEWMAN, 2018). Valores altos de pontuação são atribuídos a detecções consideradas boas e valores baixos são atribuídos a detecções consideradas ruins. O problema de maximização de modularidade consiste em encontrar uma divisão da rede em comunidades, que possua o maior valor de pontuação. No caso mais simples, por exemplo, é possível dividir uma rede em duas comunidades. Para uma rede com n nós, existe um número de possibilidades de divisões da ordem de 2^n . Já no caso mais simples é possível perceber a complexidade exponencial do problema se for atacado por um método de força bruta, tornando essa abordagem inviável. Mesmo para abordagens que limitam a pesquisa apenas às divisões que têm uma chance de ser as melhores, segundo (NEWMAN, 2018), existem resultados fundamentais no campo da ciência da computação que nos dizem que nenhum algoritmo desse tipo será capaz de encontrar a melhor divisão da rede em todos os casos. Dessa forma, na prática o problema é tratado buscando algoritmos que encontram soluções aproximadas, mas aceitáveis, isto é, algoritmos heurísticos ou apenas heurísticas. Uma variedade de algoritmos heurísticos foi testada para o problema de maximização de modularidade.

Definições precisas do problema de detecção de comunidades como um problema de otimização são conhecidas por serem computacionalmente intratáveis. A tentativa de encontrar partições razoavelmente boas, levou a busca por algoritmos rápidos atraindo

muito interesse nos últimos anos com o surgimento de grandes conjuntos de dados de rede e o crescimento dessa área de estudo. Com o surgimento de vários algoritmos para encontrar partições que fossem boas soluções para o problema fez-se necessário avaliar a qualidade dessas partições utilizando a medida de modularidade de partição. Blondel et. al (BLONDEL, 2008) propõe um algoritmo que encontra partições de alta modularidade de grandes redes em curto espaço de tempo e que desdobra uma estrutura de comunidade hierárquica completa para a rede, dando acesso a diferentes resoluções de detecção de comunidade. Em contraste com outros algoritmos de detecção de comunidade, as limitações desse método estão relacionadas ao tamanho da rede e não ao tempo de computação limitado.

O algoritmo é conhecido como método de Louvain e foi criado pelo professor e pesquisador Vincent Blondel da Universidade de Louvain. O método é dividido em duas fases que são executadas iterativamente. A partir de uma rede ponderada, cada nó é atribuído a uma comunidade diferente. No primeiro momento, existem tantas comunidades quanto nós. A partir daí, para cada nó, é avaliado o ganho de modularidade que ocorreria removendo o nó de sua comunidade e o colocando na comunidade de um de seus vizinhos. O nó é então colocado na comunidade para a qual esse ganho é máximo (em caso de empate é usado uma regra para desempatar). Se nenhum ganho positivo for possível, o nó é mantido em sua comunidade original. Este processo é aplicado repetida e sequencialmente para todos os nós até que nenhuma melhoria adicional possa ser alcançada, encerrando assim a primeira. Esta primeira fase finaliza quando um máximo local da modularidade é atingido, ou seja, quando nenhum movimento individual pode melhorar a modularidade. A saída obtida para esta etapa depende da ordem em que os nós são considerados, o que também pode influenciar o tempo de cálculo.

A segunda fase do algoritmo consiste na construção de uma nova rede cujos nós são as comunidades encontradas na primeira fase. E cada nova aresta tem um peso dado pela soma dos pesos das arestas entre os nós nas duas comunidades correspondentes. As arestas entre os nós da mesma comunidade levam a *self-loops* para esta comunidade na nova rede. Quando a segunda fase é concluída, a primeira fase do algoritmo é aplicada à rede resultante. As duas fases são aplicadas iterativamente (consulte a Figura 1) até que não haja mais alterações e um máximo de modularidade seja atingido.

Este algoritmo simples tem várias vantagens como possuir etapas que são intuitivas e fáceis de implementar. Além disso, o algoritmo é extremamente rápido, e isso se deve ao fato de que os ganhos possíveis na modularidade ao longo das etapas são fáceis de calcular e ao fato de que número de comunidades diminui drasticamente após apenas algumas repetições das duas etapas, de modo que a maior parte do tempo de execução é concentrado nas primeiras iterações.

Uma abordagem diferente de descoberta de comunidades envolve a ideia de per-

mitir que *nós* possamos estar presentes em mais de uma comunidade. É possível encontrar diversos algoritmos que exploram essa abordagem, sendo um deles conhecido como DEMON. O algoritmo DEMON (*Democratic Estimate of the Modular Organization of a Network*) proposto Coscia *et al.* (COSCIA, 2012) parte da ideia de explorar a estrutura local de comunidades, e então encontrar a estrutura a nível global. Sendo assim, o algoritmo DEMON propõe uma abordagem simples *local-first* que é explorada localmente com o objetivo de encontrar a estrutura modular que surge de padrões locais (COSCIA, 2012). Para alcançar esse objetivo um *nó* (considerado o *nó* ego da rede) pode democraticamente, votar nas comunidades dada sua visão limitada do seu entorno. O algoritmo de propagação de rótulo é utilizado para este processo e, assim, comunidades locais são mescladas em um conjunto global.

2.6 Dados de CDR

Os registros detalhados de chamadas de telefones celulares ou CDR (do inglês Call Detail Record) são fontes de dados de empresas de telecomunicações que controlam o tráfego de telefones celulares em uma determinada região. O número de chamadas em um determinado local é um forte indicador de quantas pessoas existem naquele local em um determinado momento. O objetivo dos dados de CDR é cobrar dos clientes o uso do celular. No entanto, eles também podem ser usados em investigações para tentar colocar o usuário de telefone celular em uma localização geográfica baseada na torre usada pelo telefone celular para enviar ou receber uma chamada, mensagem de texto ou conexão de dados pela Internet.

Os dados de CDR podem incluir informações sobre a localização aproximada da chamada e recebimento, hora e duração da chamada, estatísticas anônimas e dados demográficos dos clientes e identificadores anônimos usados para se referir aos indivíduos que ligam ou recebem uma chamada. Atualmente no Brasil, a Lei Geral de Proteção de Dados Pessoais (LGPD) impõe que dados como esses só podem ser disponibilizados quando o cliente autoriza. Uma prática comum realizada para esse tipo de dados é o processo de anonimização. A anonimização é uma técnica de processamento de dados que remove ou modifica informações que possam identificar uma pessoa. A modificação pode incluir inclusive aplicação de uma função *hash* criptográfica as informações tidas como mais sensíveis.

No início dos anos 2000 Aiello *et. all* (AIELLO, 2000) propôs um grafo baseado em chamadas telefônicas considerando todas as ligações feitas em um dia. Cada chamada telefônica concluída é uma aresta no grafo. Cada número de telefone que origina ou recebe uma chamada é um *nó* no grafo. As arestas são direcionadas com a direção indicando a origem e o destino da chamada. Desde então, foram realizados muitos trabalhos explorando

base de dados de CDR a fim de estudar redes sociais nesse contexto. Onnela *et. all* (ONNELA, 2007), por exemplo, construíram uma rede social com 3,9 milhões de nós a partir de dados de CDR, a fim de obter uma rede de comunicação humana que pudesse refletir uma organização à nível social. Neste trabalho tanto a duração da chamada e o número cumulativo de chamadas feitas entre os indivíduos ao longo de 18 semanas foram utilizadas para quantificar a interação entre os indivíduos.

Em (ONNELA, 2007) é feito um levantamento sobre os estudos usando dados de telefonia e suas aplicação. Neste trabalho, a partir de uma rede social estática construída a partir das chamadas entre os usuários são realizadas diversas análises. Aspectos das propriedades topológicas e uma visão da rede no contexto de sua posição geográfica são explorados oferecendo um entendimento inicial dos dados. O trabalho também discute os estudos considerando as redes de dados CDR como redes temporais, uma vez que as chamadas possuem duração e estão localizadas no tempo. Além de estudos de mobilidade humana, possibilitada pelas informações espaço-temporais contidas nos dados de CDR e estudos de difusão de informações e a disseminação de vírus de telefones celulares também são considerados neste trabalho, além de uma combinação dessas abordagens. O trabalho ainda faz considerações sobre as questões de privacidade levantadas pela disponibilidade e uso de dados pessoais. Dessa forma, Onnela *et. all* oferecem um trabalho denso e rico com um conjunto de técnicas para análise de redes sociais construídas a partir de dados de CDR.

3 Materiais e métodos

Neste capítulo na seção 3.1, será descrita a motivação de usar a divisão regional que foi escolhida das possibilidades existentes propostas pelo IBGE . Esse capítulo também descreve as bases de dados CDR utilizadas como seus tamanhos, as regiões geográficas que esses dados cobrem e os filtros aplicados sobre esses dados para eliminar possíveis dados ruidosos na seção 3.2. Os requisitos para realizar os experimentos são descritos na 3.3.3. Por fim, na seção 3.3 é descrita a metodologia de investigação das interações sociais e padrões de mobilidade.

3.1 A Divisão regional do Brasil

Desde a década de 1940 três propostas de Divisão Regional do Brasil foram elaboradas pelos geógrafos do Instituto Brasileiro de Geografia e Estatística (IBGE, 2017) - IBGE, ao longo do Século XX: Zonas Fisiográficas, da década de 1940 e suas revisões; Microrregiões Homogêneas, de 1968; e Mesorregiões e Microrregiões Geográficas, de 1989. Segundo o IBGE, o aumento na diferenciação interna do território brasileiro é um resultado das transformações econômicas, demográficas, políticas e ambientais ocorridas ao longo das últimas décadas. Desse processo de diferenciação surge a demanda por uma nova divisão regional do Brasil. Nesse contexto, a maior heterogeneidade do território ressalta a urgência de uma revisão das unidades sub-estaduais do espaço brasileiro. Assim surge como prioridade uma revisão das unidades mesorregionais e microrregionais, que passaram a ser denominadas de Regiões Geográficas Intermediárias e Regiões Geográficas Imediatas, respectivamente. Ainda segundo o IBGE, as escalas dessa divisão oferecem um retrato mais detalhado do território brasileiro e dos seus atributos.

Os dados de CDR usados como fonte de informação neste trabalho estão cobrem um intervalo de tempo compreendido entre 21/03/2013 a 19/04/2013. Desta forma uma divisão geográfica do Brasil que seja recente e coerente com o intervalo temporal dos dados de CDR se torna fundamental para realizar parte da investigação deste trabalho. Portanto a divisão que define subdivisões estaduais em regiões Intermediárias e Imediatas proposta no documento "Divisão Regional do Brasil em Regiões Geográficas Imediatas e Regiões Geográficas Intermediárias 2017" do IBGE pode ser entendido como o mais adequado dois motivos. O primeiro, se deve ao fato de ser a divisão regional mais recente realizada pelo IBGE e o segundo por ser contemporânea aos dados se levado em conta o período das demais divisões.

Figura 2: Região Imediata de São João del Rei



Belmiro Braga, Bias Fortes, Bocaina de Minas, Bom Jardim de Minas, Chácara, Chiador, Coronel Pacheco, Ewbank da Câmara, Goianá, Juiz de Fora, Liberdade, Lima Duarte, Matias Barbosa, Olaria, Oliveira Fortes, Paiva, Passa Vinte, Pedro Teixeira, Piau, Rio Novo, Rio Preto, Santa Bárbara do Monte Verde, Santa Rita de Jacutinga, Santana do Deserto, Santos Dumont, Simão Pereira), dos quais estima-se uma população de 749569 habitantes, sendo 573285 no municípios base de Juiz de Fora.

Dados CDR são gerados por diferentes operadoras de telefonia móvel, que devem trocar informações com o propósito de efetuar a cobrança devida pela utilização das antenas por seus clientes e geralmente armazenam informação de um ID do usuário que realizou a chamada, um ID do usuário que recebeu a chamada, assim como sua duração. Dados CDR também armazenam o ID da antena que o dispositivo móvel utilizou para realizar a chamada o que, juntamente com sua localização, pode ser utilizado para estimar a localização do usuário.

Os dados considerados consistem em 9.293.317 registros de ligações telefônicas na região de São João del-Rei e 63.588.996 registros na região de Juiz de Fora, observados

durante um período de 30 dias. Com base nesses registros, de maneira imediata, é possível construir redes de comunicação direcionadas, representando os usuários como nós e as ligações telefônicas como arestas. Cada aresta (v_i, v_j) indica que o indivíduo representado pelo vértice v_i realizou uma chamada para o indivíduo representado pelo vértice v_j e armazena duas informações: a quantidade e o tempo total das chamadas de v_i para v_j .

Diversos trabalhos podem ser encontrados na literatura com o objetivo de obter uma rede social a partir de uma rede de comunicação construída através de CDRs (ONNELA, 2007; BLONDEL, 2015) e uma das maiores preocupações nesse sentido é evitar que a rede social represente relações que claramente não refletem interações entre pessoas. Assim, tomando como base os resultados dos trabalhos de Robin Dunbar (CARRON, 2016), mais especificamente o conhecido número de Dunbar, eliminamos da rede vértices que mantinham mais do que 150 relações sociais, que poderiam indicar centrais de atendimento e números com ramais. Além disso, eliminamos da rede arestas entre usuários com menos do que três ligações telefônicas e arestas entre usuários com duração total de chamada inferior a 30 segundos, com o objetivo de reduzir o número de registros de ligações telefônicas realizadas por engano ou interações esporádicas. Considerando n como o número de nós e m como o número de arestas, os tamanhos das redes direcionadas foram $n = 1.303.294/m = 689.952$ para a rede SJDR e $n = 3.080.014/m = 4.395.760$ para a rede JF, após essas filtrações.

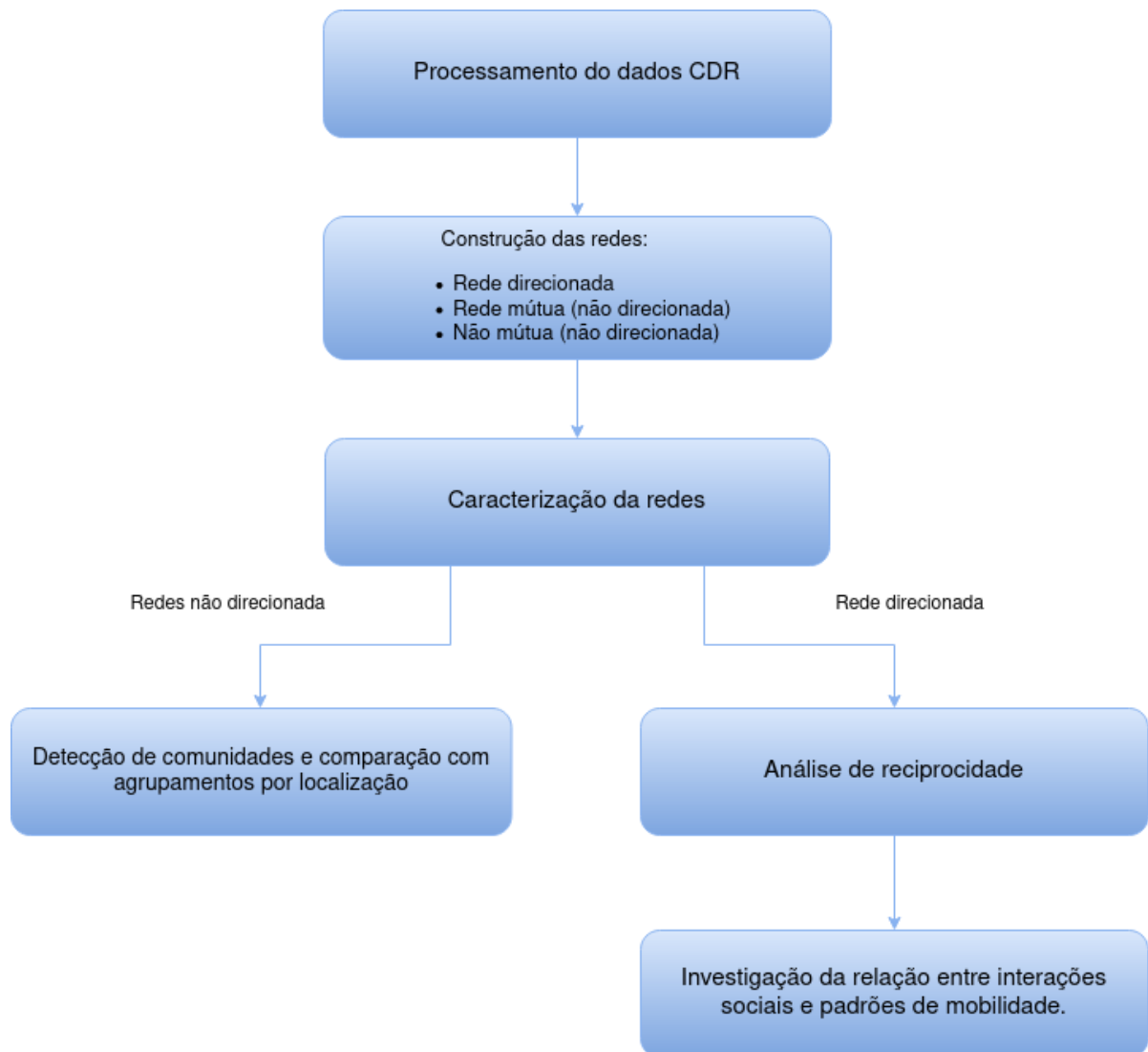
Um aspecto particular observado na base de dados CDR considerada neste trabalho é o fato de haver o registro da localização apenas da antena que realizou a chamada, mas não da antena que recebeu a chamada. Embora isso não seja um grande problema para a caracterização das relações sociais entre os usuários, é impossível analisar qualquer característica dos indivíduos em relação à utilização do espaço, caso não tenham realizado chamadas, o que impacta diretamente na investigação de padrões de mobilidade proposta neste trabalho. Sendo assim, foram também eliminados da rede direcionada vértices que apresentavam grau de saída nulo, ou seja, não realizaram nenhuma ligação. Após essa filtração, as redes direcionadas resultantes possuem $n = 161.948/m = 297.427$, para a região de SJDR e $n = 662.212/m = 2.770.314$, para a região de JF.

Apesar de, em um primeiro momento, a comunicação entre indivíduos em uma rede de telefonia se dar de maneira direcionada, já que um indivíduo realiza a ligação e outro recebe, neste trabalho consideramos que, em diversas situações, as relações sociais, mesmo provenientes de chamadas telefônicas, podem ser consideradas não-direcionadas. Por isso, obtivemos também uma versão não-direcionada da rede. Assim, as arestas da rede não direcionada armazenam o número de ligações entre os indivíduos e o tempo total das chamadas realizadas por eles, não importando quem realizou e que recebeu as chamadas.

3.3 Metodologia

Após o pré-processamento dos dados CDR e a construção da rede social em suas versões direcionada e não-direcionada, foi adotada uma série de passos metodológicos para que fossem cumpridos os objetivos do trabalho relacionados à caracterização das relações sociais e à investigação da relação entre interações sociais e padrões de mobilidade. As seções seguintes apresentam uma visão geral dos passos metodológicos adotados que estão representados na Figura 3

Figura 3: Diagrama de passos metodológicos aplicados



A construção das redes sociais direcionadas que foram utilizadas nos experimentos explorou os conceitos de relações não mútuas e mútuas. Para o primeiro caso, definido como uma rede não mútua, uma aresta é criada entre dois *nós* se existe pelo menos um registro de chamada entre os dois nós, i.e., se o nó i liga para j ou se j liga para i . Para o segundo caso, definido como uma rede mútua, uma aresta é criada entre dois *nós*, se

existe pelo menos um par de registros de chamada recíproco entre os dois nós, i.e., se o nó i liga para j e se j liga para i . A primeira abordagem pode conter muitas chamadas que são unilaterais, que geralmente indica que o chamador pode não conhecer pessoalmente quem está sendo chamado (ONNELA, 2007). Dessa forma, a segunda abordagem é uma alternativa para remover esses casos.

É importante ressaltar que as redes sociais obtidas a partir dessas abordagens, capturam obviamente apenas uma parte da rede social subjacente que consiste em todas as formas de interações sociais. Essa rede social subjacente é definida a partir das interações que acontecem pessoalmente, e-mail e comunicação por telefone fixo, mensagens de texto entre outras formas de comunicação existentes. No entanto, a investigação sobre o impacto dos meios de comunicação e da Internet na conectividade entre as pessoas, sugere que o uso de um meio para a comunicação implica no uso de outros meios também (HAYTHORNTHWAITE, 2005). Além disso, o fato de um número de telefone celular não estar listado em catálogos telefônicos aumenta a possibilidade das interações serem confiáveis, i.e., as pessoas tendem a compartilhar seus números de celular apenas com indivíduos em quem confiam. Portanto, esses fatores possibilitam que redes sociais criadas a partir de dados de CDR possam ser utilizadas como um meio de entender a rede social subjacente que se forma a partir de todos tipos de interações possíveis entre as pessoas (ONNELA, 2007).

A primeira etapa foi gerar uma rede considerando todas as ligações telefônicas registradas a partir dos dados de CDR. Essa rede serviu como base para gerar os dois tipos de rede descritos no início desta seção. Para criar rede mútua ligações entre dois nós podem ser removidas caso não haja a reciprocidade entre os mesmos. Essa remoção pode fazer com que a rede fique fragmentada, desta forma aplicou-se um passo adicional para encontrar a componente gigante para utilizá-la nos experimentos.

3.3.1 Metodologia para caracterização das interações sociais

Além das ferramentas básicas, frequentemente exploradas na literatura para a caracterização de redes sociais, neste trabalho, as redes sociais obtidas a partir dos dados de registros de comunicação por ligações telefônicas são investigadas considerando a forma como os indivíduos se organizam em grupos, que é caracterizada pela estrutura de comunidades e o balanceamento das relações que é caracterizado pela reciprocidade.

O processo de caracterização da estrutura de comunidades realizado pode ser dividido em duas etapas. A primeira consistiu em definir a localização de cada nó da rede e criar uma estrutura que representasse os agrupamentos de municípios. Na segunda etapa foi realizada uma comparação da estrutura encontrada na primeira com outros dois agrupamentos encontrados pelo algoritmo DEMON e o algoritmo de Louvain.

Na primeira etapa a localização atribuída a cada nó se dá por meio da identificação da residência da pessoa representada por aquele nó. A identificação da possível residência para um determinado nó se baseia em identificar qual localização o nó tem um número maior de ligações realizadas em dias úteis no intervalo de 18h às 5h. Desta forma, foi definido o primeiro agrupamento por localização.

A segunda etapa consistiu em comparar as comunidades encontradas por algoritmos de detecção de comunidades bem estudados nesta área. O primeiro algoritmo, também chamado de método de Louvain, que é descrito ao final da seção 2.5 foi utilizado uma implementação de Networkx. Essa implementação possui um parâmetro *randomize* que permite randomizar a ordem de avaliação do nó e a ordem de avaliação da comunidade para obter partições diferentes em cada chamada. Esse parâmetro foi explorado com 5 valores para verificar o impacto do mesmo na rede em questão. O segundo algoritmo utilizado foi algoritmo de detecção de comunidades sobrepostas conhecido como DEMON. Foi utilizada a implementação desse algoritmo, fornecida pela biblioteca Networkx. O método que implementa este algoritmo possui dois parâmetros que são o tamanho mínimo para as comunidades encontradas e o valor de ϵ (epsilon) que permite variar a porcentagem de elementos comuns fornecidos por cada par de comunidades sobrepostas (COSCIA, 2012). O primeiro parâmetro foi definido usando o tamanho da menor comunidade encontrada no agrupamento por localização, realizado na primeira etapa do experimento, a fim de evitar comunidades com um tamanho que não é observado no agrupamento. Para o segundo parâmetro, que admite valores no intervalo de 0 a 1, buscou-se realizar experimentos que explorassem este intervalo. Portanto, foi realizada uma busca exploratória no intervalo utilizando passos de tamanho 0.1 (i.e. 0.1, 0.2 até 1) a fim de entender o impacto que este parâmetro causa no agrupamento realizado na rede através de medidas de modularidade.

A caracterização da reciprocidade foi desenvolvida na aplicação *Jupyter* fazendo o cálculo de uma medida de reciprocidade geral da rede e também seguindo uma metodologia proposta por Chawla *et. al.* (CHAWLA, 2011) que permitisse avaliar como essa característica estaria distribuída na rede.

Para o cálculo da reciprocidade geral foi utilizado o método "*overall_reciprocity*" da biblioteca *NetworkX*. A definição de reciprocidade de um grafo direcionado usada nesta implementação é a razão entre o número de arestas apontando em ambas as direções e o número total de arestas no gráfico. Formalmente $r = |(u, v) \in G| / |(v, u) \in G|$.

Para o cálculo da reciprocidade de Chawla *et. al.*, foram implementadas as fórmulas presentes no artigo do mesmo e as distribuições de seus respectivos cálculos de reciprocidade. A distribuição foi dividida em três classes, assim como no trabalho de referência, a fim de realizar uma caracterização da distribuição das reciprocidades encontradas.

3.3.2 Metodologia para investigação da relação entre interações sociais e padrões de mobilidade

Para a caracterização dos padrões de mobilidade dos indivíduos no espaço urbano, é preciso definir uma representação matemática da maneira como eles visitam cada um dos lugares, estimados pelas posições das antenas onde as ligações telefônicas ocorreram. Neste trabalho, a representação adotada foi baseada na proposta de Toole *et al.* (TOOLE, 2015). A cada vértice da rede é atribuída uma matriz de localização $\mathbf{L}$, definida por como uma matriz $48 \times L$, onde L é o número de antenas de localização única no espaço observado e cada linha de $\mathbf{L}$ corresponde a uma hora t de seu dia de semana típico e seu fim de semana típico. Cada elemento $\mathbf{L}_{t,j}$ da matriz de localização de um usuário armazena a quantidade de vezes que um indivíduo realizou uma ligação utilizando a antena da localização j , o que é utilizado para estimar uma visita do usuário à localização j . Como discutido em (TOOLE, 2015), a matriz de localização pode ser de grande utilidade para o cálculo de diversas propriedades de mobilidade dos usuários, como número de ligações telefônicas realizadas por um indivíduo e a frequência de visitação dos usuários a determinadas localizações. Uma dessas propriedades, de particular importância neste trabalho, é o vetor de trajetórias $\mathbf{v}$, obtido através da soma das linhas da matriz de localização. O vetor de trajetórias, calculado para cada um dos usuários da rede, armazena a quantidade de vezes que um indivíduo visitou cada uma das localizações durante o período de observação e pode ser utilizado como uma forma de representar seu padrão de interação com o espaço urbano.

Considerando usuários representados por dois vértices v_i e v_j , pode-se medir sua similaridade de mobilidade através do cálculo da similaridade de seus vetores de trajetória, $\mathbf{v}_i$ e $\mathbf{v}_j$ respectivamente. Assim como proposto em (TOOLE, 2015), neste trabalho a similaridade entre os vetores de localização $\mathbf{v}_i$ e $\mathbf{v}_j$ é calculada através do cosseno do ângulo formado entre eles, como

$$\cos\theta_{i,j} = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{|\mathbf{v}_i||\mathbf{v}_j|}. \quad (3.1)$$

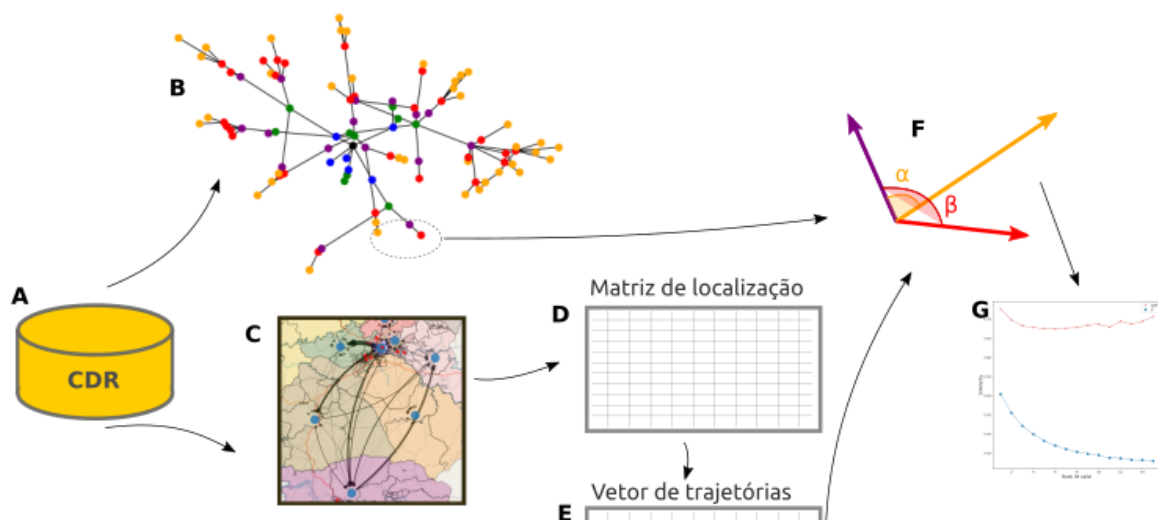
A similaridade de cosseno é frequentemente utilizada para comparar vetores em espaços de alta dimensionalidade, especialmente quando seus elementos possuem muitas posições vazias, representando estruturas muito esparsas. Assim, essa medida mostra-se bastante adequada para o cálculo da similaridade entre dois vetores de similaridade, quando comparada a outras medidas de similaridade, já que não leva em conta a magnitude dos vetores de trajetória, desconsiderando, assim, o número de ligações realizadas pelos usuários. Além disso, a similaridade de cosseno entre dois vetores não é afetada pelas posições vazias nos vetores, concentrando a análise sobre os lugares visitados.

Particularmente, neste trabalho, estamos interessados em investigar a relação entre as interações sociais e os padrões de mobilidade. Assim, os experimentos realizados

visam explorar o impacto de indivíduos que possuem relação social relevante nos padrões de mobilidade de outros indivíduos. Por isso, a rede social será fundamental para a sistematização da escolha dos indivíduos a serem comparados.

A metodologia aqui descrita é apresentada de maneira geral na Figura 4. Primeiramente, é realizado um pré-processamento dos dados CDR (Figura 4A), definindo as relações sociais, modeladas pela rede social (Figura 4B). São também identificadas as localizações visitadas pelos usuários no espaço urbano (Figura 4C), que são representadas pelas matrizes de localização (Figura 4D), para que em seguida, os vetores de trajetórias sejam calculados (Figura 4E). Tomando como base os vetores de trajetórias e as interações sociais a serem investigadas, são calculadas as similaridades entre os indivíduos (Figura 4F), para que os experimentos possam ser realizados (Figura 4G).

Figura 4: Visão geral da metodologia. (A) Base de dados CDR. (B) Rede social. (C) Mobilidade. (D) Matriz de localização. (E) Vetor de trajetórias. (F) Similaridade de cosseno. (G) Experimento.



3.3.3 Ambiente computacional e ferramentas

As análises e processamentos realizados neste trabalho foram desenvolvidas na linguagem Python versão 3.6. Os módulos necessários para execução das análises e processamentos estão disponíveis no repositório "[cdr_research](#)" armazenado no GitHub. Para utilização ou replicação dos experimentos é altamente recomendável a criação de um ambiente virtual utilizando o módulo 'venv'. O módulo venv fornece suporte para a criação de ambientes virtuais com seus próprios diretórios de *site*, opcionalmente isolados dos diretórios de *site* do sistema. O ambiente virtual limpo criado servirá como um ecossistema seguro para a instalação das dependências necessárias para executar as análises e processamentos. Todas as dependências necessárias estão listadas no arquivo *requirements.txt* contido no diretório raiz do projeto.

Para a construção das redes sociais a partir dos dados de CDR foi utilizada a biblioteca *igraph* (CSARDI; NEPUSZ, 2006). Para realizar a caracterização com medidas básicas da redes como grau, grau máximo, distribuição de graus entre outros, foi utilizada a biblioteca *NetworkX* (HAGBERG, 2008). As detecções de comunidades foram realizadas com implementações disponíveis na biblioteca CDlib (GIULIO, 2019). As análises de reciprocidade foram realizadas na plataforma *Jupyter* (KLUYVER, 2016) e utilizando a biblioteca *matplotlib* (HUNTER, 2007).

4 Experimentos e discussão

Neste capítulo serão apresentados os resultados para a metodologia aplicada neste trabalho que foi descrita na seção 3.3. Sendo assim, os resultados podem ser divididos em duas partes principais: A análise das relações sociais e a análise da relação entre os aspectos sociais e a mobilidade. A primeira parte compreende a caracterização, a análise de comunidades e a investigação do fenômeno de reciprocidade das redes na seção 4.1. A segunda parte descreve a investigação dos padrões de mobilidades da redes na seção 4.2.

4.1 Análise das relações sociais

4.1.1 Caracterizações básicas

As Tabelas 1 e 2 contém descrições das informações básicas para redes modeladas das regiões imediatas de São João del Rei e Juiz de Fora respectivamente. A primeira coluna contém as métricas calculadas para cada rede obtida, sendo as seguintes métricas: número de nós (n), número de arestas (m), grau médio ($\hat{k}$), diâmetro (d), densidade (D) e coeficiente de *clustering* ($\hat{c}$). A primeira linha das tabelas contém a identificação para cada uma das redes obtidas sendo elas: a direcionada com todas as chamadas, a rede mútua e a rede não mútua.

Uma das medidas obtidas foi o diâmetro da rede que informa a maior distância (calculada usando o número de arestas) entre quaisquer dois vértices e é tipicamente pequeno em redes sociais e redes criadas a partir de dados de CDR (BLONDEL, 2015). Para a Região Imediata de São João del Rei, o diâmetro encontrado para a rede não mútua exibe essa característica de diâmetro pequeno, sendo ele composto por 53 arestas em uma rede de 37678 nós. Já a rede mútua apresenta um diâmetro de 31 arestas que é pequeno mas não na mesma proporção que a primeira, uma vez que esta rede possui 228 nós. Para a rede direcionada, esta métrica não está disponível devido ao fato da implementação não ser capaz de calcular o diâmetro para uma rede direcionada que não é fortemente conectada.

Embora seja observado em redes sociais um alto coeficiente de *clustering* (WATTS; STROGATZ, 1998) os valores observados não correspondem a esse padrão.

Os baixos níveis de densidade estão diretamente relacionados ao tipo de relação que a rede representa, i.e. as relações pessoais, e refletem o fato de que no mundo real as pessoas atingem um limite baixo na quantidade de relacionamentos.

A Figura 5 apresenta a distribuição de graus para as versões direcionada e não-

Tabela 1: Informações Básicas das Redes da Região Imediata de São João del Rei.

redes/atributos	n	m	$\hat{k}$	d	D	$\hat{c}c$
não mútua	37678	52937	2.81	53	7.45×10^{-5}	0.10
mútua	228	245	2.14	31	0.009	0.10
direcionada	184339	255538	1.38	indisponível	7.52×10^{-6}	0.043

Tabela 2: Informações Básicas das Redes da Região Imediata de Juiz de Fora.

redes/atributos	n	m	$\hat{k}$	d	D	$\hat{c}c$
não mútua	1351566	3663519	2.7	indisponível	4.011×10^{-6}	0.053
mútua	373500	627189	3.35	indisponível	8.99×10^{-6}	0.11
direcionada	3080014	4395760	1.42	indisponível	4.63×10^{-7}	0.03

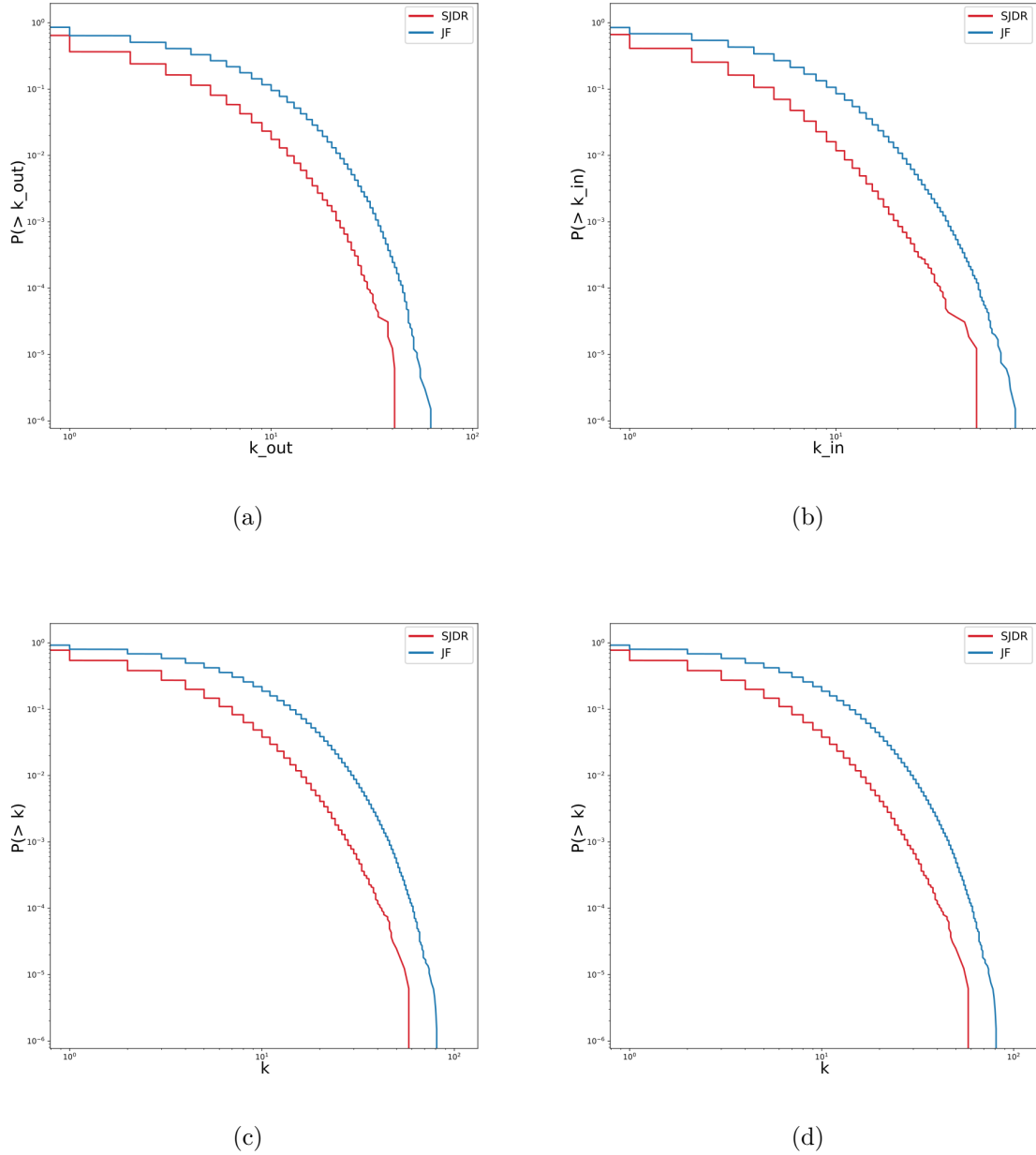
direcionada da rede e a distribuição do número de chamadas entre indivíduos. É possível perceber que as distribuições de graus são semelhantes a distribuições de graus tipicamente observadas em outras redes sociais relatadas na literatura (NEWMAN; PARK, 2003). Também é importante notar que há uma queda acentuada nas probabilidades observadas para os graus e para o número de ligações maiores que 10, indicando que apenas uma pequena quantidade de indivíduos apresenta atividade mais intensa, o que é particularmente relevante na discussão sobre a relação entre características sociais e de mobilidade conduzidas neste trabalho, mais especificamente na Seção 4.2.

4.1.2 Divisão das redes sociais em comunidades

Uma das características mais frequentemente investigadas em redes é a organização de seus vértices em comunidades, isto é, grupos com uma alta densidade interna e uma baixa densidade externa. Considerando o contexto explorado neste trabalho, a estrutura de comunidades pode revelar como os contatos sociais estão distribuídos, o que poderia evidenciar aspectos relacionados à organização espacial e administrativa da cidade e, integrada a outras fontes de dados, poderia auxiliar a compreensão de aspectos sociológicos e culturais da população.

Aplicando o método de Louvain (BLONDEL, 2008) às redes SJDR e JF foram obtidas estruturas de comunidades que, considerando a modularidade Q de Newman (NEWMAN; PARK, 2003), com $Q = 0,8437$ para a rede SJDR e $Q = 0,6014$ para a rede JF. Entretanto, as partições apresentaram um número de comunidades extremamente alto (42704 e 58293 para SJDR e JF, respectivamente) e as comunidades resultantes não são muito significativas. Conforme observado por Tibely *et al.* (TIBÉLY, 2011) e discutido por Blondel *et al.* (BLONDEL, 2015), as pequenas estruturas de árvores presentes na rede de telefonia são frequentemente consideradas como comunidades, mesmo que sejam extremamente esparsas.

Figura 5: Distribuição de quantidades. (a) Grau de saída (rede direcionada). (b) Grau de entrada (rede direcionada). (c) Grau (rede não direcionada). (d) Número de ligações (rede não direcionada).



4.1.3 Explorando a descoberta de comunidades na Região Imediata de São João del Rei

Os experimentos desta seção se concentram na Região Imediata de São João del Rei. Das redes modeladas para esta região, a que se mostrou mais apta para a realização dos experimentos foi a rede não mútua. Por meio de inspeção foi verificado nesta rede a maior ocorrência de nós com localizações diferentes, o que permite explorar se a descoberta de comunidades utilizando algoritmos de detecção revelaria a organização geoespacial.

Os experimentos para comparar os agrupamentos de localização com os algoritmos de detecção de comunidades escolhidos, geraram resultados que estão na Tabela 3. Esses experimentos buscaram explorar duas abordagens dos estudos de detecção de comunidades, sendo elas: comunidades com sobreposição utilizando o algoritmo DEMON e comunidades sem sobreposição utilizando o algoritmo de Louvain.

Cada resultado tabelado está associado ao nome do algoritmo que o gerou, aos parâmetros utilizados e ao número de comunidades geradas na execução do algoritmo. Para o algoritmo de Louvain, a tabela apresenta os 5 valores testados na coluna semente para o parâmetro *randomize* que é usado como semente pelo gerador de números aleatórios e geram diferentes execuções do algoritmo. Para o algoritmo DEMON a tabela apresenta os valores de ϵ utilizados, que permite variar a porcentagem de elementos nas sobreposições fornecidas por cada par de comunidades sobrepostas. Os resultados de cada comparação se encontram na coluna "pontuação" que foram calculados através "*Overlapping Normalized Mutual Information*" (LANCICHINETTI, 2009) para lidar com o caso de partições sobrepostas.

Os resultados encontrados para o algoritmo Louvain variando as sementes se mostraram sem grande variação. Os valores do número de comunidades encontradas para cada execução não variaram muito, isso corrobora a ideia de que uma variação na escolha do nó para iniciar o algoritmo, não influencia a modularidade (BLONDEL, 2008). Dessa forma, não foram encontrados agrupamentos com grande diferença, o que é sugerido pelo valor de pontuação que também não teve grandes variações.

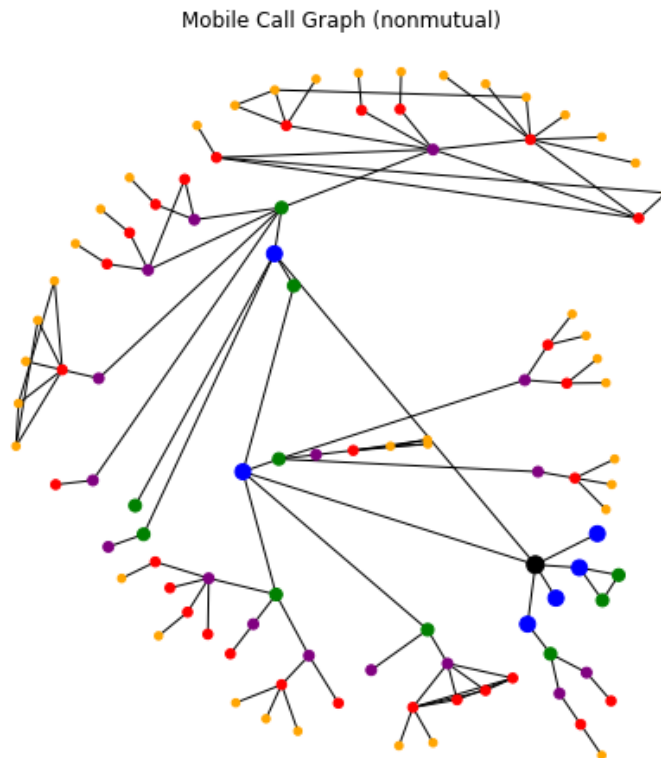
A inspeção da rede mostrou a recorrência de pequenas estruturas semelhantes a árvores que segundo Tibely *et al.* (TIBÉLY, 2011) frequentemente são consideradas como comunidades, embora sua estrutura seja esparsa. Dessa forma essas estruturas não impedem que a detecção de comunidades reflita os agrupamentos por localização. Este fato, segundo Blondel *et al.* (BLONDEL, 2015), embora seja contra-intuitivo deve ser considerado com cautela uma vez que as comunidades podem fornecer informações significativas, quando combinadas de forma útil com informações externas.

Entretanto, a estrutura das redes sociais obtidas a partir de ligações telefônicas apresenta características que dificultam a identificação das comunidades, como discutido por Tibely *et al.* (TIBÉLY, 2011) e Blondel *et al.* (BLONDEL, 2015). Conforme apontado por Onnela *et al.* (ONNELA, 2007), redes sociais baseadas em ligações telefônicas apresentam um aspecto parecido com uma árvore, o que pode fazer com que algoritmos clássicos para detecção de comunidades produzam resultado insatisfatório neste tipo de redes. A Figura 6 apresenta uma amostra em *snowball* da rede social SJDR. O vértice marcado na cor preta representa o ego, a partir do qual são obtidos vértices em até seis saltos. É possível notar que a estrutura topológica em torno do ego é bastante esparsa, com formação de poucos triângulos e organizada, essencialmente, quase como uma árvore.

Tabela 3: Comparação das comunidades encontradas utilizando localização nós com as comunidades encontradas ϵ pelos métodos DEMON e Louvain.

método	ϵ	semente	num de comunidades	pontuação
Louvain	não possui	2	415	0.022
Louvain	não possui	3	420	0.023
Louvain	não possui	5	416	0.021
Louvain	não possui	7	406	0.024
Louvain	não possui	11	421	0.023
DEMON	0.1	não possui	179	0.0
DEMON	0.2	não possui	198	0.0
DEMON	0.3	não possui	197	0.0
DEMON	0.4	não possui	196	0.0
DEMON	0.5	não possui	203	0.0
DEMON	0.6	não possui	201	0.0
DEMON	0.7	não possui	215	0.0
DEMON	0.8	não possui	224	0.0
DEMON	0.9	não possui	286	0.0
DEMON	1.0	não possui	293	0.0

Figura 6: Exemplo de uma amostra da rede usando *snowball sampling* com seis saltos (o vértice em preto é o ego).



Tais resultados levantam dúvidas sobre a utilidade de métodos para detecção de comunidades em redes de telefonia. Embora Blondel *et al.* argumentem que resultados

desse tipo sejam mais uma particularidade do que um problema, a discussão sobre a aplicabilidade das comunidades identificadas automaticamente em redes vai ao encontro de um conjunto de trabalhos que investigam o que é, de fato, uma comunidade. Uma noção básica e consensual existente na área de ciência de redes é a de que uma comunidade deve estar unicamente codificada nas conexões de uma rede e devem ser identificadas apenas através dessas conexões, o que implica que a estrutura topológica da rede deva conter uma espécie de gabarito sobre a estrutura de comunidades, sua *ground truth*. Entretanto, como discutido por Peel *et al.* (PEEL, 2017), em redes reais a partição correta e processo de geração dos dados são tipicamente desconhecidos, tornando possível que não haja um gabarito codificado nos dados. Assim, a aplicação de métodos para detecção de comunidades, apesar de conveniente, pode levar a resultados incorretos, muitas vezes enviesados pelos próprios métodos empregados ou pelas medidas consideradas para a avaliação da qualidade das partições. Nesse sentido, é bastante comum que uma estrutura de comunidades apresente uma boa qualidade quando são consideradas medidas de avaliação como a modularidade, mas reflitam partições nitidamente enviesadas pelos métodos aplicados para identificá-las, e não pela estrutura codificada nas conexões (VIEIRA, 2020).

4.1.4 Reciprocidade

Ainda com o objetivo de investigar o comportamento de usuários com base no comportamento de suas ligações telefônicas, é possível quantificar a reciprocidade das relações. Tradicionalmente, a reciprocidade é utilizada como uma forma de quantificar as arestas do tipo (v_j, v_i) que ocorrem em razão de arestas (v_i, v_j) em um grafo direcionado. Considerando a medida de reciprocidade tradicional, foram observados valores de 0,4708 e 0,4885 para as redes SJDR e JF, respectivamente, para a versão da rede direcionada, mas sem vértices com grau de saída nulo. Por outro lado, alguns autores discutem a caracterização da reciprocidade de relações de uma forma mais refinada. Wang *et al.* (CHAWLA, 2011) propõem uma medida de reciprocidade para avaliar o desbalanceamento no relacionamento entre dois usuários:

$$R_{ij} = |\ln(p_{ij}) - \ln(p_{ji})|, \quad (4.1)$$

onde p_{ij} é a probabilidade de ligação de um v_i para um v_j . A Figura 8 apresenta a distribuição da reciprocidade dos usuários nas redes SJDR e JF, onde valores mais baixos indicam relações mais balanceadas. É possível perceber que a maior parte das relações apresenta reciprocidade moderada, enquanto poucas relações são extremamente balanceadas ou desbalanceadas. Esses resultados, quando comparados com os apresentados em (CHAWLA, 2011), em que os autores também realizam uma análise de características sociais em uma base de telefonia celular, revelam uma forte semelhança entre as redes so-

Tabela 4: Porcentagens em cada classe de reciprocidade

Redes	Recíprocas	Parcialmente Recíprocas	Não recíprocas
Chawla <i>et al.</i>	28%	58%	14%
SJDR	28%	65%	7%
JF	27%	65%	8%

ciais investigadas neste trabalho e as redes investigadas por outros autores, considerando essa propriedade.

Como visto em (CHAWLA, 2011), a distribuição dos valores de reciprocidade calculados usando a Equação 4.1 pode ser dividida em três classes, sendo elas:

Díades recíprocas cuja a razão da probabilidade está compreendida no intervalo 1.0 a 1.5 (0 a 0.41 utilizando o logaritmo natural).

Díades parcialmente recíprocas cuja a razão da probabilidade está compreendida no intervalo 1.5 a 9.0 (0.41 a 2.20 utilizando o logaritmo natural).

Díades não recíprocas cuja a razão da probabilidade seja maior que 9.0 (2.20 utilizando o logaritmo natural).

Os valores percentuais encontrados em cada classe de reciprocidade para as redes de SJDR e JF estão listados na tabela 4 juntos aos valores encontrados no estudo de Chawla *et. all* (CHAWLA, 2011). Os valores tabelas estão apresentados nos histogramas de distribuição dos valores de reciprocidade na Fig 7.

A rede de (CHAWLA, 2011) possui 8 milhões de vértices e 17.2 milhões de arestas simétricas. A rede de SJDR possui 161.948 de vértices e 297.427 de arestas simétricas. Por fim, a rede de JF possui 662.212 de vértices e 2.770.314 de arestas simétricas.

Para a primeira classe foram verificados valores percentuais muito próximos para as três redes, sugerindo que a formação de díades dessa classe (classe das díades recíprocas) não são sensíveis às diferenças topológicas como tamanho da rede, por exemplo, que varia nas três redes observadas.

Já para as demais classes, é possível verificar um movimento no percentual de díades da segunda classe para a terceira, na rede de (CHAWLA, 2011). A rede em questão possui número de arestas e vértices substancialmente maior que as demais, o que pode contribuir para um aumento da terceira classe.

As distribuições em cada uma das redes, embora não sejam idênticas, possuem uma divisão em classes próximas sugerindo que os desequilíbrios de comunicação entre os nós em cada uma das três são similares.

Figura 7: Histograma da distribuição de reciprocidade. Cada classe de díades está representada reciprocamente como: recíprocas na cor azul, parcialmente recíprocas na cor vermelha e não recíprocas na cor laranja (a) Histograma para rede SJDR. (b) Histograma para rede JF.

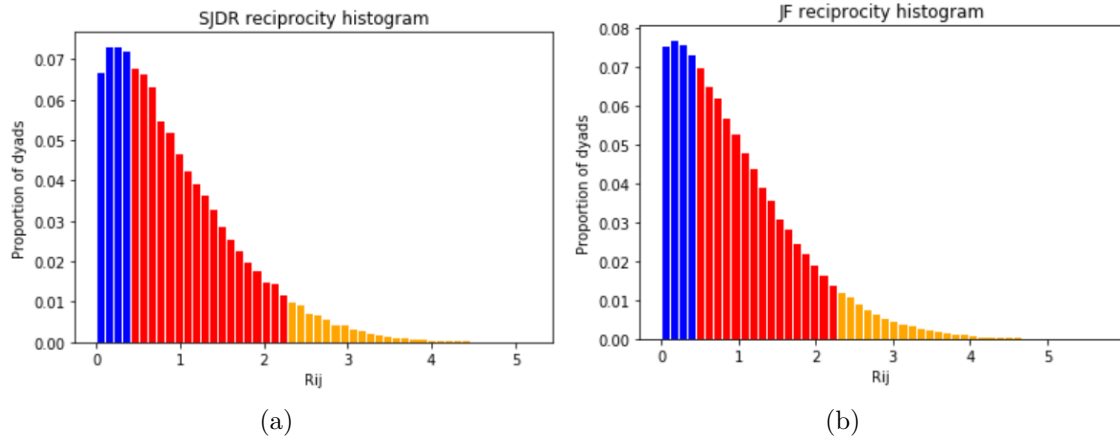
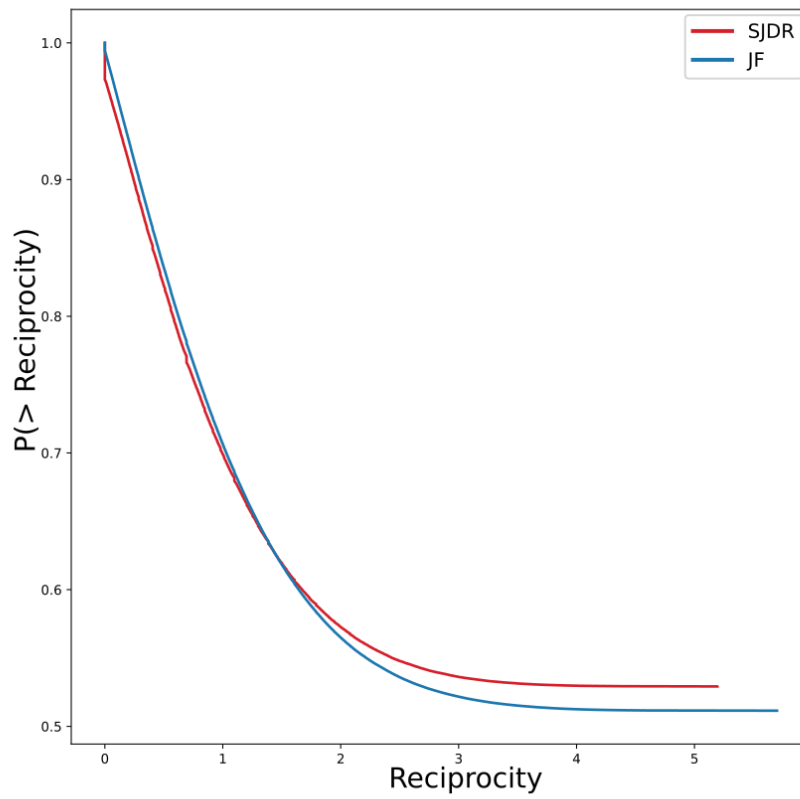


Figura 8: Distribuição da reciprocidade na rede.

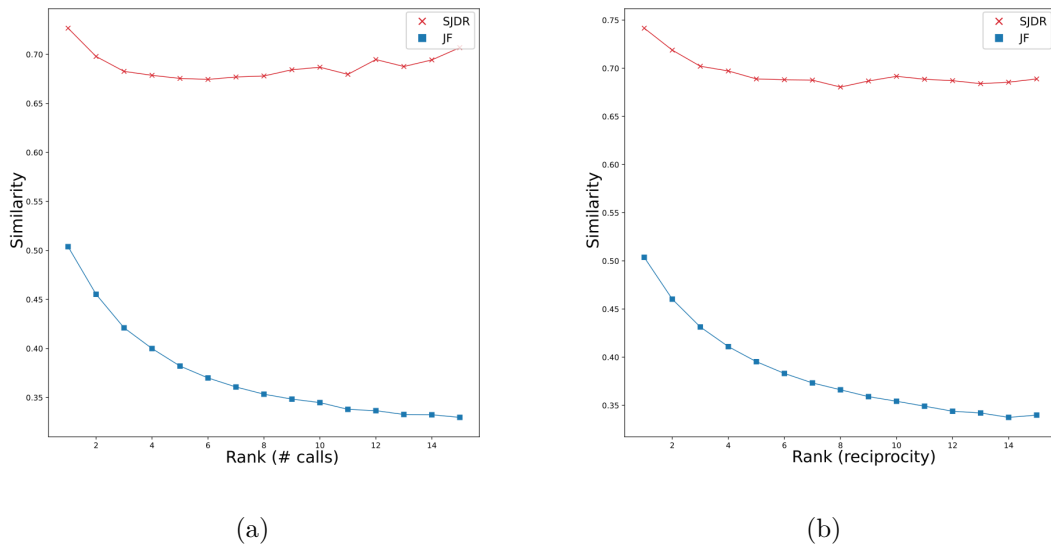


4.2 Análise da relação entre aspectos sociais e mobilidade

A partir da definição de vetores de trajetórias entre quaisquer pares de indivíduos, é possível calcular a similaridade entre seus padrões de mobilidade, tomando como base a metodologia descrita na Seção 3.3. Nesta seção, são apresentados os resultados de uma série de experimentos que têm como objetivo mostrar que aspectos relacionados à mobilidade dos indivíduos estão fortemente correlacionados às suas interações sociais. Para isso, as relações sociais são exploradas sob diferentes pontos de vista.

Primeiramente, investigamos como indivíduos que possuem relação direta com um outro afetam sua mobilidade, o que é apresentado na Figura 9, sob duas perspectivas: números de ligações (Figura 9(a)) e reciprocidade (Figura 9(b)). Tomando como base cada indivíduo v_i de uma rede, identificamos todos os seus vizinhos v_j e elaboramos um *rank* de amizade (com base no número de ligações e na reciprocidade) e calculamos as similaridades entre os vetores de trajetória t_i e t_j .

Figura 9: Similaridade da mobilidade X *rank* dos contatos sociais. (a) Número de ligações. (b) Reciprocidade.



Como é possível observar pela Figura 9, os indivíduos tendem a apresentar uma maior similaridade de trajetória com seus contatos mais altos no *rank* e essa similaridade é reduzida à medida em que a força das relações cai. Na rede JF, esse comportamento é preservado até a posição 15 do *rank*, enquanto na rede SJDR isso é observado apenas nas primeiras posições do *rank*. Uma possível explicação pode ser obtida pela observação da Figura 5, que mostra que, talvez pelo menor volume de dados, os vértices da rede SJDR apresentam graus sensivelmente menores que os vértices da rede JF. Assim, é interessante observar, combinando os resultados das Figuras 5 e 9, a enorme queda no

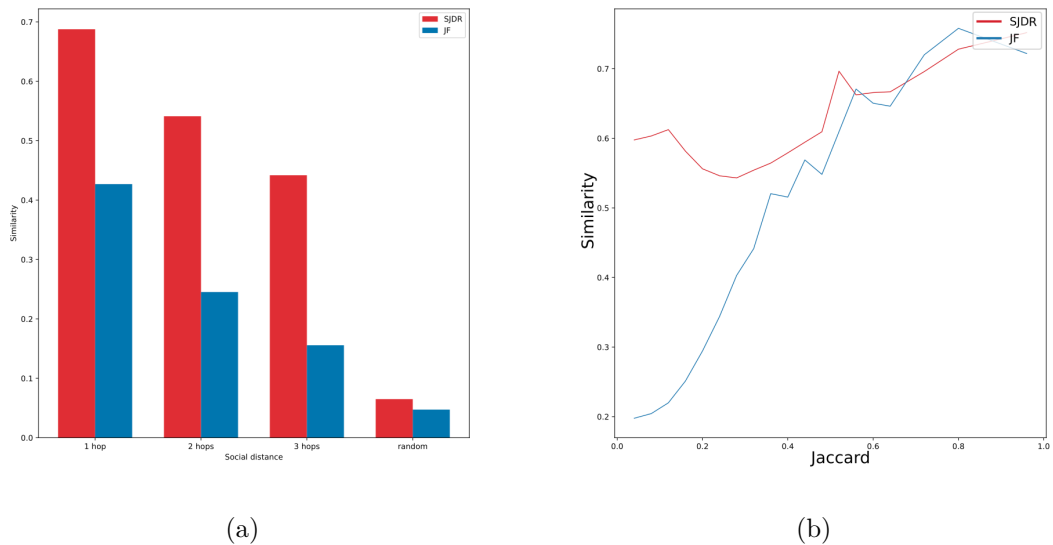
número de contatos sociais, e o seu impacto na definição de padrões de mobilidade, em torno de 15. Esse número corrobora as descobertas de Dunbar (CARRON, 2016), mais especificamente no que diz respeito ao que o autor afirma que, tipicamente, as pessoas são capazes de manter contato com alguma regularidade com apenas outras 15. Assim, como há apenas poucos vértices com grau alto, qualquer comportamento desses indivíduos acaba por enviesar a análise de maneira desproporcional, o que é conhecido como “viés do sobrevivente” (*“survivor bias”*) (MEHRABI, 2019).

Foram também realizados experimentos para avaliar se, além do comportamento dos usuários com os quais um determinado indivíduo mantém contato direto, isto é, seus vizinhos na rede, a topologia da rede social tem um impacto significativo sobre a mobilidade dos indivíduos. Ou seja, investigamos se padrões de mobilidade de indivíduos v_j com os quais um vértice v_i sequer mantém relações sociais afetam o comportamento de v_i , o que é apresentado na Figura 10. Para isso, duas ideias foram exploradas. A Figura 10(a) apresenta uma investigação da variação da similaridade entre indivíduos em razão de sua distância na rede social, considerando 1 salto (seus vizinhos), 2 saltos e 3 saltos, além de sua similaridade com indivíduos aleatórios. Além disso, investigamos como a similaridade entre dois indivíduos v_i e v_j é afetada pelo número de vizinhos que v_i e v_j têm em comum (Figura 10(b)). Para isso, utilizamos o coeficiente de Jaccard:

$$J_{i,j} = \frac{|F_i \cap F_j|}{|F_i \cup F_j|}, \quad (4.2)$$

onde F_i é o conjunto dos vizinhos do vértice v_i .

Figura 10: Similaridade da mobilidade X similaridade na rede social. (a) Número de saltos. (b) Coeficiente de Jaccard.



Primeiramente, é possível verificar pela Figura 10 que a mobilidade de um indivíduo v_i pode ser similar a de outro indivíduo v_j , mesmo que eles não mantenham relações na rede social e, assim, é possível começar a tentar mapear os comportamentos em espaços urbanos. Outro ponto relevante, observado na Figura 10(a), é que mesmo com três saltos de separação, a similaridade entre indivíduos ainda é significativamente mais alta do que entre dois indivíduos aleatoriamente selecionados. Além disso, as curvas apresentadas para os coeficientes de Jaccard na Figura 10(b) mesmo não apresentando comportamento monotônico, indicam tendência crescente, corroborando as observações anteriores de que a similaridade social está positivamente correlacionada à similaridade de mobilidade.

É importante observar que os resultados dos experimentos conduzidos neste trabalho são coerentes com resultados obtidos por outros autores que também investigam a relação entre a interação social e padrões de mobilidade (TOOLE, 2015; GRABOWICZ, 2014; DOMENICO, 2013). Mais especificamente, é interessante comparar os resultados obtidos neste trabalho com os resultados descritos em Toole *et al.* (TOOLE, 2015). Lá, os autores investigam três regiões, das quais duas se assemelham muito com o comportamento exibido para a região de JF deste trabalho e uma se assemelha ao comportamento exibido para a região de SJDR. Em (TOOLE, 2015), os autores argumentam que a região que apresentou comportamento diferente é a única para a qual os dados CDR não possuem geolocalização dos receptores de chamadas telefônicas e que isso justificaria a distinção de comportamento. Se isso fosse verdade, era esperado que ambas regiões exploradas neste trabalho (SJDR e JF) apresentassem comportamento similar, já que nenhuma delas possui tal informação para os receptores. Assim, outras hipóteses precisam ser investigadas para avaliar tal diferença nos experimentos.

Na tentativa de compreender as diferenças no comportamento dos usuários de diferentes regiões, pode-se levantar como hipótese o baixo volume de dados para a região de SJDR, que faz com que o "viés do sobrevivente" tenha um impacto muito mais significativo. Ainda nesse sentido, é impossível ignorar o fato que, em todos os experimentos realizados, a similaridade observada para os usuários de telefonia na região de SJDR é muito maior do que para os usuários de JF. Uma das justificativas para isso pode vir do fato de que o centro urbano da região de SJDR possui uma extensão muito menor do que o centro urbano da região de JF, fazendo com que todos os indivíduos, de fato, tenham uma tendência a apresentar padrões mais similares de mobilidade, ou seja, uma inadequação na granularidade dos dados. Outra justificativa pode vir do fato de o centro urbano de SJDR possuir uma menor complexidade socioeconômica em relação ao centro urbano de JF, reduzindo o impacto das relações sociais nos padrões de mobilidade. De qualquer forma, o presente trabalho levanta uma série de questões que precisam ser exploradas mais profundamente em trabalhos futuros. De qualquer forma, é possível afirmar que os resultados obtidos neste trabalho são coerentes com outros resultados encontrados na literatura e que as relações sociais e de mobilidade observadas nos experimentos realizados

vão ao encontro de resultados observados por outros autores. Assim, esses resultados permitem uma mais clara compreensão dos dados de telefonia armazenados no CDR e da forma como as pessoas se relacionam e utilizam o espaço urbano;

5 Conclusão e direções futuras

Este trabalho apresenta uma investigação da relação entre redes sociais e aspectos de mobilidade de indivíduos em espaços urbanos, tomando como base registros de ligações telefônicas (*Call Detail Records* - CDR). Para isto foram estabelecidos objetivos específicos na seção 1.1. O primeiro objetivo foi a modelagem de redes sociais a partir de critérios que tentassem explorar as possíveis relações existentes entre os indivíduos. Dessa forma foram obtidas 3 redes para cada região estudada (São João del Rei e Juiz de Fora). O segundo objetivo foi a caracterização através de métricas básicas das redes obtidas na etapa anterior que foram agrupadas tabelas e descritas na seção 4.1.1. O objetivo seguinte foi investigar o processo de formação de comunidades pela localização por algoritmos de detecção de comunidades de duas abordagens diferentes. Este objetivo foi realizado para a rede não mútua da região imediata de São João del Rei verificando que as abordagens não oferecem agrupamentos que refletem a organização espacial. O último objetivo listado foi realizado a partir das redes direcionadas, realizando a análise de reciprocidade e a investigação entre interações sociais e padrões de mobilidade.

Os experimentos realizados considerando as duas cidades permitiu caracterizar aspectos sociais relevantes dos indivíduos nas cidades. Essa caracterização sugere uma forte interdependência entre seus padrões de utilização do espaço urbano e os padrões exibidos pelos seus contatos sociais. Mais ainda, foi possível verificar uma forte similaridade entre os padrões de mobilidade de um indivíduo e os padrões de mobilidade dos indivíduos de sua rede social, mesmo que eles não estabeleçam um contato direto. É importante enfatizar que esse é o primeiro trabalho dessa natureza realizado para essas cidades do Brasil e que os resultados encontrados são bastante coerentes com os resultados encontrados nos trabalhos de outros autores realizados em espaços urbanos em outras regiões do mundo e permitem um melhor entendimento sobre as cidades aqui investigadas, mas também fornecem uma forma de validação dos dados CDR em outras aplicações relacionadas a redes sociais e mobilidade urbana.

Uma série de questões, principalmente relacionadas ao impacto de características socioeconômicas dos indivíduos sobre sua mobilidade, ainda precisam ser investigadas em trabalhos futuros. Sendo assim, algumas possibilidades possuem grande potencial no contexto deste trabalho. Uma possibilidade é aplicar a metodologia apresentada a outras cidades com tamanhos diferentes (como por exemplo grandes centros) e perfis diferentes, com o objetivo de generalizar os resultados encontrados. Outra possibilidade é usar o modelo de rede para propagação de doenças envolvendo cenários pandêmicos intensamente estudados como o caso da COVID-19 que teve um impacto significativo na sociedade nos últimos tempos. Por fim, ainda é possível investigar a interdependência dos padrões

de utilização do espaço urbano e os padrões de contatos nos diferentes estratos sociais buscando entender se existe correlação desses padrões para as regiões investigadas.

Referências

- BROECK, V. d. W.; GIOANNINI, C.; GONÇALVES, B.; QUAGGIOTTO, M.; COLIZZA, V.; VESPIGNANI, A. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC infectious diseases*, p. 37–37, 2011. Citado na página 10.
- FARBER, S.; O’KELLY, M.; MILLER, H.; NEUTENS, T. Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography*, v. 49, p. 26–38, 12 2015. Citado na página 10.
- BLONDEL, V. D.; DECUYPER, A.; KRINGS, G. A survey of results on mobile phone datasets analysis. *EPJ data science*, Springer, v. 4, n. 1, p. 10, 2015. Citado 8 vezes nas páginas 10, 11, 12, 18, 33, 40, 41 e 43.
- GONZALEZ, M. C.; HIDALGO, C. A.; BARABASI, A.-L. Understanding individual human mobility patterns. *Nature*, Nature Publishing Group, v. 453, n. 7196, p. 779–782, June 2008. Citado 2 vezes nas páginas 10 e 12.
- ALESSANDRETTI, L.; SAPIEZYNSKI, P.; LEHMANN, S.; BARONCHELLI, A. Multi-scale spatio-temporal analysis of human mobility. *PLOS ONE*, Public Library of Science, v. 12, n. 2, p. 1–17, 02 2017. Citado na página 10.
- LENORMAND, M.; RAMASCO, J. J. Towards a better understanding of cities using mobility data. *Built Environment*, v. 42, p. 356–364(9), 10 2016. Citado na página 10.
- NEWMAN, M. *Networks*. [S.l.]: Oxford university press, 2018. Citado 5 vezes nas páginas 10, 14, 15, 18 e 26.
- MURATA, T.; KOGA, H. Extended methods for influence maximization in dynamic networks. *Computational Social Networks*, Springer, v. 5, n. 1, p. 8, 2018. Citado na página 11.
- ALTARELLI, F.; BRAUNSTEIN, A.; DALLASTA, L.; WAKELING, J. R.; ZECCHINA, R. Containing epidemic outbreaks by message-passing techniques. *Physical Review X*, APS, v. 4, n. 2, p. 021024, 2014. Citado na página 11.
- FERNÁNDEZ-GRACIA, J.; LACASA, L. Bipartisanship breakdown, functional networks, and forensic analysis in spanish 2015 and 2016 national elections. *Complexity*, Hindawi, v. 2018, 2018. Citado na página 11.
- BARBOSA, H.; BARTHELEMY, M.; GHOSHAL, G.; JAMES, C. R.; LENORMAND, M.; LOUAIL, T.; MENEZES, R.; RAMASCO, J. J.; SIMINI, F.; TOMASINI, M. Human mobility: Models and applications. *Physics Reports*, v. 734, p. 1–74, 2018. ISSN 0370-1573. Human mobility: Models and applications. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S037015731830022>>. Citado na página 12.
- SAPIEZYNSKI, P.; STOPCZYNSKI, A.; GATEJ, R.; LEHMANN, S. Tracking human mobility using wifi signals. *PLOS ONE*, Public Library of Science, v. 10, n. 7, p. 1–11,

07 2015. Disponível em: <<https://doi.org/10.1371/journal.pone.013082>>. Citado na página 12.

GRABOWICZ, P. A.; RAMASCO, J. J.; GONÇALVES, B.; EGUÍLUZ, V. M. Entangling mobility and interactions in social media. *PLOS ONE*, Public Library of Science, v. 9, n. 3, p. 1–12, 03 2014. Disponível em: <<https://doi.org/10.1371/journal.pone.009219>>. Citado 2 vezes nas páginas 12 e 50.

DOMENICO, M. D.; SOLÈ-RIBALTA, A.; COZZO, E.; KIVELÄ, M.; MORENO, Y.; PORTER, M.; GOMEZ, S.; ARENAS, A. Mathematical formulation of multi-layer networks. *Physical Review X*, v. 3, 07 2013. Citado 2 vezes nas páginas 12 e 50.

CORNACCHIA, G.; ROSSETTI, G.; PAPPALARDO, L. *Modelling Human Mobility considering Spatial, Temporal and Social Dimensions*. 2020. Citado na página 12.

TOOLE, J. L.; HERRERA-YAQUÉ, C.; SCHNEIDER, C. M.; GONZÁLEZ, M. C. Coupling human mobility and social ties. *The Royal Society*, v. 12, n. 105, p. 20141128, 2015. Citado 3 vezes nas páginas 12, 37 e 50.

WASSERMAN, S.; FAUST, K. *Social network analysis: Methods and applications*. Cambridge university press, 1994. Citado 3 vezes nas páginas 14, 15 e 22.

BARNES, J. A. Class and committees in a norwegian island parish. *Human relations*, Sage Publications Sage CA: Thousand Oaks, CA, v. 7, n. 1, p. 39–58, 1954. Citado na página 14.

GARDNER, B. B.; GARDNER, M. R.; LOEB, M. B. Social status and education in a southern community. *The School Review*, The University of Chicago Press, v. 50, n. 3, p. 179–191, 1942. Citado na página 15.

MARSDEN, P. V. Network data and measurement. *Annual review of sociology*, JSTOR, p. 435–463, 1990. Citado 2 vezes nas páginas 16 e 17.

REA, L. M.; PARKER, R. A. *Designing and conducting survey research: A comprehensive guide*. [S.l.]: John Wiley & Sons, 2014. Citado 2 vezes nas páginas 16 e 17.

KLOVDAHL, A. S.; POTTERAT, J. J.; WOODHOUSE, D. E.; MUTH, J. B.; MUTH, S. Q.; DARROW, W. W. Social networks and infectious disease: The colorado springs study. *Social science & medicine*, Elsevier, v. 38, n. 1, p. 79–88, 1994. Citado na página 16.

LILJEROS, F.; EDLING, C. R.; AMARAL, L. A. N.; STANLEY, H. E.; ÅBERG, Y. The web of human sexual contacts. *Nature*, Nature Publishing Group, v. 411, n. 6840, p. 907–908, 2001. Citado na página 16.

POTTERAT, J. J.; PHILLIPS-PLUMMER, L.; MUTH, S. Q.; ROTHENBERG, R.; WOODHOUSE, D.; MALDONADO-LONG, T.; ZIMMERMAN, H.; MUTH, J. Risk network structure in the early epidemic phase of hiv transmission in colorado springs. *Sexually transmitted infections*, The Medical Society for the Study of Venereal Disease, v. 78, n. suppl 1, p. i159–i163, 2002. Citado na página 16.

ROTHENBERG, R.; BALDWIN, J.; TROTTER, R.; MUTH, S. The risk environment for hiv transmission: Results from the atlanta and flagstaff network studies. *Journal of Urban Health*, Springer, v. 78, n. 3, p. 419–432, 2001. Citado na página 16.

- BURT, R. S. Network items and the general social survey. *Social networks*, Elsevier, v. 6, n. 4, p. 293–339, 1984. Citado na página 17.
- ZACHARY, W. W. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, University of New Mexico, v. 33, n. 4, p. 452–473, 1977. Citado na página 17.
- FLACK, J. C.; GIRVAN, M.; WAAL, F. B. D.; KRAKAUER, D. C. Policing stabilizes construction of social niches in primates. *Nature*, Nature Publishing Group, v. 439, n. 7075, p. 426–429, 2006. Citado na página 17.
- GRANT, T. Dominance and association among members of a captive and a free-ranging group of grey kangaroos (*macropus giganteus*). *Animal Behaviour*, Elsevier, v. 21, n. 3, p. 449–456, 1973. Citado na página 17.
- CONNOR, R. C.; HEITHAUS, M. R.; BARRE, L. M. Superalliance of bottlenose dolphins. *Nature*, Nature Publishing Group, v. 397, n. 6720, p. 571–572, 1999. Citado na página 17.
- ACTION, R. the rise of the medici. *American Journal of Sociology*, v. 98, p. 1259–1319, 1993. Citado na página 17.
- EBEL, H.; MIELSCH, L.-I.; BORNHOLDT, S. Scale-free topology of e-mail networks. *Physical review E*, APS, v. 66, n. 3, p. 035103, 2002. Citado na página 18.
- TYLER, J. R.; WILKINSON, D. M.; HUBERMAN, B. A. Email as spectroscopy: Automated discovery of community structure within organizations. In: SPRINGER. *Communities and technologies*. [S.l.], 2003. p. 81–96. Citado na página 18.
- ONNELA, J.-P.; ARBESMAN, S.; GONZÁLEZ, M. C.; BARABÁSI, A.-L.; CHRISTAKIS, N. A. Geographic constraints on social network groups. *PLoS one*, Public Library of Science, v. 6, n. 4, p. e16939, 2011. Citado na página 18.
- STOPCZYNSKI, A.; SEKARA, V.; SAPIEZYNSKI, P.; CUTTONE, A.; MADSEN, M. M.; LARSEN, J. E.; LEHMANN, S. Measuring large-scale social networks with high resolution. *PloS one*, Public Library of Science, v. 9, n. 4, p. e95978, 2014. Citado na página 18.
- RATTI, C.; SOBOLEVSKY, S.; CALABRESE, F.; ANDRIS, C.; READES, J.; MARTINO, M.; CLAXTON, R.; STROGATZ, S. H. Redrawing the map of great britain from a network of human interactions. *PloS one*, Public Library of Science, v. 5, n. 12, p. e14248, 2010. Citado na página 18.
- ONNELA, J.-P.; SARAMÄKI, J.; HYVÖNEN, J.; SZABÓ, G.; LAZER, D.; KASKI, K.; KERTÉSZ, J.; BARABÁSI, A.-L. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 104, n. 18, p. 7332–7336, 2007. Citado 2 vezes nas páginas 18 e 43.
- HIDALGO, C. A.; RODRÍGUEZ-SICKERT, C. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 387, n. 12, p. 3017–3024, 2008. Citado na página 18.

- DODDS, P. S.; HARRIS, K. D.; KLOUMANN, I. M.; BLISS, C. A.; DANFORTH, C. M. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, Public Library of Science, v. 6, n. 12, p. e26752, 2011. Citado na página 18.
- GOLBECK, J.; GRIMES, J. M.; ROGERS, A. Twitter use by the us congress. *Journal of the American society for information science and technology*, Wiley Online Library, v. 61, n. 8, p. 1612–1621, 2010. Citado na página 18.
- GONÇALVES, B.; PERRA, N.; VESPIGNANI, A. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, Public Library of Science, v. 6, n. 8, p. e22656, 2011. Citado na página 18.
- ARAL, S.; WALKER, D. Identifying influential and susceptible members of social networks. *Science*, American Association for the Advancement of Science, v. 337, n. 6092, p. 337–341, 2012. Citado na página 18.
- UGANDER, J.; KARRER, B.; BACKSTROM, L.; MARLOW, C. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011. Citado 2 vezes nas páginas 18 e 19.
- BOND, R. M.; FARISS, C. J.; JONES, J. J.; KRAMER, A. D.; MARLOW, C.; SETTLE, J. E.; FOWLER, J. H. A 61-million-person experiment in social influence and political mobilization. *Nature*, Nature Publishing Group, v. 489, n. 7415, p. 295–298, 2012. Citado na página 18.
- HOLME, P.; EDLING, C. R.; LILJEROS, F. Structure and time evolution of an internet dating community. *Social Networks*, Elsevier, v. 26, n. 2, p. 155–174, 2004. Citado na página 18.
- HOLME, P. Core-periphery organization of complex networks. *Physical Review E*, APS, v. 72, n. 4, p. 046111, 2005. Citado na página 18.
- GALASKIEWICZ, J. *Social organization of an urban grants economy: A study of business philanthropy and nonprofit organizations*. [S.l.]: Elsevier, 2016. Citado na página 18.
- MILGRAM, S. The small world problem. *Psychology today*, New York, v. 2, n. 1, p. 60–67, 1967. Citado na página 19.
- TRAVERS, J.; MILGRAM, S. An experimental study of the small world problem. In: *Social networks*. [S.l.]: Elsevier, 1977. p. 179–197. Citado na página 19.
- KLEINBERG, J. M. Navigation in a small world. *Nature*, Nature Publishing Group, v. 406, n. 6798, p. 845–845, 2000. Citado na página 19.
- KLEINBERG, J. The small-world phenomenon: An algorithmic perspective. In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. [S.l.: s.n.], 2000. p. 163–170. Citado na página 19.
- DODDS, P. S.; MUHAMAD, R.; WATTS, D. J. An experimental study of search in global social networks. *science*, American Association for the Advancement of Science, v. 301, n. 5634, p. 827–829, 2003. Citado na página 20.

- KLOVDAHL, A. S. Urban social networks: Some methodological problems and possibilities. *The small world*, Ablex Norwood, NJ, p. 176–210, 1989. Citado na página 20.
- THOMPSON, S. K.; FRANK, O. Model-based estimation with link-tracing sampling designs. *Survey Methodology*, v. 26, n. 1, p. 87–98, 2000. Citado na página 20.
- SALGANIK, M. J.; HECKATHORN, D. D. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, Wiley Online Library, v. 34, n. 1, p. 193–240, 2004. Citado na página 21.
- CHAWLA, N. V.; HACHEN, D.; LIZARDO, O.; TOROCZKAI, Z.; STRATHMAN, A.; WANG, C. Weighted reciprocity in human communication networks. *CoRR*, abs/1108.2822, 2011. Disponível em: <<http://arxiv.org/abs/1108.2822>>. Citado 5 vezes nas páginas 22, 23, 36, 45 e 46.
- HARMON, J. *The Psychology of Interpersonal Relations. By Fritz Heider. New York: John Wiley and Sons, Inc., 1958. 322 pp. \$6.25.* [S.l.]: The University of North Carolina Press, 1959. Citado na página 22.
- HALLINAN, M. T. The process of friendship formation. *Social networks*, Elsevier, v. 1, n. 2, p. 193–210, 1978. Citado na página 22.
- GARLASCHELLI, D.; LOFFREDO, M. I. Patterns of link reciprocity in directed networks. *Physical review letters*, APS, v. 93, n. 26, p. 268701, 2004. Citado 2 vezes nas páginas 22 e 23.
- BARRAT, A.; BARTHELEMY, M.; PASTOR-SATORRAS, R.; VESPIGNANI, A. The architecture of complex weighted networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 101, n. 11, p. 3747–3752, 2004. Citado na página 23.
- BARABÁSI, A.; PÓSFAL, M. *Network Science*. [S.l.]: Cambridge University Press, 2016. ISBN 9781107076266. Citado 2 vezes nas páginas 24 e 26.
- LANCICHINETTI, A.; FORTUNATO, S. Community detection algorithms: a comparative analysis. *Physical review E*, APS, v. 80, n. 5, p. 056117, 2009. Citado na página 24.
- ZHANG, S.; WANG, R.-S.; ZHANG, X.-S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 374, n. 1, p. 483–490, 2007. Citado na página 24.
- JORA, C.; CHIRA, C. Evolutionary community detection in complex and dynamic networks. In: IEEE. *2016 IEEE 12th international conference on Intelligent computer communication and processing (ICCP)*. [S.l.], 2016. p. 127–134. Citado na página 24.
- KRISHNAMURTHY, B.; WANG, J. On network-aware clustering of web clients. In: *Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*. [S.l.: s.n.], 2000. p. 97–110. Citado na página 24.

- REDDY, P. K.; KITSUREGAWA, M.; SREEKANTH, P.; RAO, S. S. A graph based approach to extract a neighborhood customer community for collaborative filtering. In: SPRINGER. *International Workshop on Databases in Networked Information Systems*. [S.l.], 2002. p. 188–200. Citado na página 24.
- STEENSTRUP, M. Cluster-based networks. In: *Ad hoc networking*. [S.l.: s.n.], 2001. p. 75–138. Citado na página 24.
- FORTUNATO, S. Community detection in graphs. *Physics reports*, Elsevier, v. 486, n. 3–5, p. 75–174, 2010. Citado na página 25.
- GRANOVETTER, M. S. The strength of weak ties. *American journal of sociology*, University of Chicago Press, v. 78, n. 6, p. 1360–1380, 1973. Citado na página 25.
- BURT, R. S. Positions in networks. *Social forces*, The University of North Carolina Press, v. 55, n. 1, p. 93–122, 1976. Citado na página 25.
- FREEMAN, L. C. A set of measures of centrality based on betweenness. *Sociometry*, JSTOR, p. 35–41, 1977. Citado na página 25.
- GUIMERA, R.; AMARAL, L. A. N. Functional cartography of complex metabolic networks. *nature*, Nature Publishing Group, v. 433, n. 7028, p. 895–900, 2005. Citado na página 25.
- PEEL, L.; LARREMORE, D. B.; CLAUSET, A. The ground truth about metadata and community detection in networks. *Science advances*, American Association for the Advancement of Science, v. 3, n. 5, p. e1602548, 2017. Citado 2 vezes nas páginas 25 e 45.
- YANG, J.; LESKOVEC, J. Community-affiliation graph model for overlapping network community detection. In: IEEE. *2012 IEEE 12th international conference on data mining*. [S.l.], 2012. p. 1170–1175. Citado na página 25.
- YANG, J.; LESKOVEC, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, Springer, v. 42, n. 1, p. 181–213, 2015. Citado na página 25.
- HRIC, D.; DARST, R. K.; FORTUNATO, S. Community detection in networks: Structural communities versus ground truth. *Physical Review E*, APS, v. 90, n. 6, p. 062805, 2014. Citado na página 25.
- NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical review E*, APS, v. 69, n. 2, p. 026113, 2004. Citado na página 26.
- LEICHT, E. A.; NEWMAN, M. E. Community structure in directed networks. *Physical review letters*, APS, v. 100, n. 11, p. 118703, 2008. Citado na página 26.
- BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2008, n. 10, p. P10008, Oct 2008. ISSN 1742-5468. Citado 3 vezes nas páginas 27, 41 e 43.

- COSCIA, M.; ROSSETTI, G.; GIANNOTTI, F.; PEDRESCHI, D. Demon: a local-first discovery method for overlapping communities. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2012. p. 615–623. Citado na página 28.
- AIELLO, W.; CHUNG, F.; LU, L. A random graph model for massive graphs. In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. [S.l.: s.n.], 2000. p. 171–180. Citado na página 28.
- ONNELA, J.-P.; SARAMÄKI, J.; HYVÖNEN, J.; SZABÓ, G.; MENEZES, M. A. D.; KASKI, K.; BARABÁSI, A.-L.; KERTÉSZ, J. Analysis of a large-scale weighted network of one-to-one human communication. *New journal of physics*, IOP Publishing, v. 9, n. 6, p. 179, 2007. Citado 3 vezes nas páginas 29, 33 e 35.
- IBGE. *Divisao regional do Brasil em regioes geograficas imediatas e regioes geograficas intermediarias, 2017*. IBGE, Instituto Brasileiro de Geografia e Estatística, 2017. Disponível em: <<https://biblioteca.ibge.gov.br/visualizacao/livros/liv100600.pd>>. Citado na página 30.
- CARRON, P. M.; KASKI, K.; DUNBAR, R. Calling dunbars numbers. *Social Networks*, Elsevier BV, v. 47, p. 151155, Oct 2016. ISSN 0378-8733. Disponível em: <<http://dx.doi.org/10.1016/j.socnet.2016.06.00>>. Citado 2 vezes nas páginas 33 e 49.
- HAYTHORNTHWAITE, C. Social networks and internet connectivity effects. *Information, Community & Society*, Taylor & Francis, v. 8, n. 2, p. 125–147, 2005. Citado na página 35.
- COSCIA, M.; ROSSETTI, G.; GIANNOTTI, F.; PEDRESCHI, D. DEMON: a local-first discovery method for overlapping communities. *CoRR*, abs/1206.0629, 2012. Disponível em: <<http://arxiv.org/abs/1206.0629>>. Citado na página 36.
- CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal*, Complex Systems, p. 1695, 2006. Disponível em: <<https://igraph.org>>. Citado na página 39.
- HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using networkx. In: VAROQUAUX, G.; VAUGHT, T.; MILLMAN, J. (Ed.). *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA: [s.n.], 2008. p. 11 – 15. Citado na página 39.
- GIULIO, R.; MILLI, L.; RÉMY, C. Cdlib: a python library to extract, compare and evaluate communities from complex networks. 2019. Citado na página 39.
- KLUYVER, T.; RAGAN-KELLEY, B.; PÉREZ, F.; GRANGER, B.; BUSSONNIER, M.; FREDERIC, J.; KELLEY, K.; HAMRICK, J.; GROUT, J.; CORLAY, S.; IVANOV, P.; AVILA, D.; ABDALLA, S.; WILLING, C.; TEAM, J. development. Jupyter notebooks ? a publishing format for reproducible computational workflows. In: LOIZIDES, F.; SCHMIDT, B. (Ed.). *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, 2016. p. 87–90. Disponível em: <<https://eprints.soton.ac.uk/403913>>. Citado na página 39.
- HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 39.

- WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado na página 40.
- NEWMAN, M.; PARK, J. Why social networks are different than other types of networks. *Physical review. E*, v. 68, p. 036122, 10 2003. Citado na página 41.
- TIBÉLY, G.; KOVANEN, L.; KARSAI, M.; KASKI, K.; KERTÉSZ, J.; SARAMÄKI, J. Communities and beyond: Mesoscopic analysis of a large social network with complementary methods. *Physical Review E*, American Physical Society (APS), v. 83, n. 5, May 2011. ISSN 1550-2376. Disponível em: <<http://dx.doi.org/10.1103/PhysRevE.83.05612>>. Citado 2 vezes nas páginas 41 e 43.
- LANCICHINETTI, A.; FORTUNATO, S.; KERTÉSZ, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, IOP Publishing, v. 11, n. 3, p. 033015, mar 2009. Disponível em: <<https://doi.org/10.1088/1367-2630/11/3/03301>>. Citado na página 43.
- VIEIRA, V. d. F.; XAVIER, C. R.; EVSUKOFF, A. G. A comparative study of overlapping community detection methods from the perspective of the structural properties. *Appl. Netw. Sci.*, v. 5, n. 1, p. 51, 2020. Disponível em: <<https://doi.org/10.1007/s41109-020-00289->>. Citado na página 45.
- MEHRABI, N.; MORSTATTER, F.; SAXENA, N.; LERMAN, K.; GALSTYAN, A. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019. Disponível em: <<http://arxiv.org/abs/1908.09635>>. Citado na página 49.