

Redes de Interesse: comparando o Google Places e Foursquare na captura da escolha de usuários por áreas urbanas

**Gustavo H. Santos¹, Fernanda R. Gubert¹,
Myriam Delgado¹, Thiago H. Silva¹**

¹Universidade Tecnológica Federal do Paraná (UTFPR)
Departamento Acadêmico de Informática
Curitiba, Brasil

gustavohenriquesantos, fernandagubert@alunos.utfpr.edu.br
myriamdelg, thiagoh@utfpr.edu.br

Abstract. *Location-Based Social Networks (LBSNs) are valuable for understanding urban behavior and provide useful data on user preferences. Modeling this data into graphs, such as Interest Network (iNET), enables significant insights. These networks can be helpful for urban area recommendations, mobility forecasts, and public policy formulation. This study uses check-in data and establishment reviews to compare the iNETs resulting from two distinct LBSNs, Foursquare and Google Places. Although the studied LBSNs differ in nature, with data varying in regularity and purpose, both modeled iNETs revealed similar urban behavior patterns, and were likewise impacted by socioeconomic and geographic factors.*

Resumo. *As Redes Sociais Baseadas em Localização (LBSNs) são úteis na compreensão do comportamento urbano, oferecendo dados valiosos sobre preferências dos usuários. A modelagem desses dados em grafos, como as Redes de Interesse, permite percepções relevantes. Essas redes podem ser úteis para, por exemplo, recomendações de áreas urbanas, previsões de mobilidade e formulação de políticas públicas. Este estudo compara redes de interesse de duas LBSNs distintas, Foursquare e Google Places, usando dados de check-ins e avaliações de estabelecimentos. Embora as LBSNs estudadas sejam diferentes em natureza, com dados diferindo em regularidade e propósito, ambas as redes de interesse modeladas revelaram padrões similares de comportamento urbano. Fatores socioeconômicos e geográficos também mostraram impacto semelhante nas redes de interesse estudadas.*

1. Introdução

Dados de Redes Sociais Baseadas em Localização (do inglês, *Location-Based Social Networks* – LBSNs) auxiliam no entendimento de diversas questões no contexto da computação urbana [Ferreira et al. 2015, Santala et al. 2017, Silva et al. 2017, Ladeira et al. 2019, Veiga et al. 2019, Ferreira et al. 2020, Senefonte et al. 2022, Silver and Silva 2023, Silva and Silver 2024]. Em particular, devido ao fato das LBSNs fornecerem dados urbanos que implicitamente têm aspectos sociais, como preferências de usuários [Silva and Loureiro 2016, Silva et al. 2019], uma modelagem apropriada pode proporcionar informações valiosas para o entendimento do comportamento de usuários em contextos urbanos.

Uma estrutura popular para representar dados de LBSNs são grafos (ou redes) [Silva et al. 2019]. Exemplos desse tipo de dado poderiam ser obtidos através de atividades geolocalizadas em LBSNs realizadas por um grupo de usuários, como as avaliações em estabelecimentos de uma cidade sob análise. Estas avaliações poderiam ser agregadas em um grafo não direcionado no qual os nós representam bairros da cidade onde os dados foram compartilhados e as arestas conectam os bairros que foram visitados pelo mesmo usuário. Este tipo de modelagem origina o que chamamos neste trabalho de *Redes de Interesse* – ver Seção 4.1 para mais detalhes.

Redes de interesse podem ser úteis em diferentes situações. Por exemplo, ao permitirem um melhor entendimento do interesse de usuários por determinados espaços físicos da cidade, elas poderiam facilitar a recomendação de áreas geográficas, promover a melhoria de previsões de mobilidade, e também fomentar novas políticas públicas para ampliar o interesse de usuários por certas áreas da cidade.

Nesse contexto, o presente trabalho tem como objetivo principal comparar redes de interesse modeladas utilizando dados provenientes de duas LBSNs: Foursquare e Google Places. No Foursquare utilizamos dados de *check-ins*, ato de compartilhar com amigos um estabelecimento onde um usuário se encontra. Já no Google Places utilizamos dados de avaliações de estabelecimentos. Ambos capturam o tipo de local onde o usuário se encontra, por exemplo, um restaurante ou uma livraria. No entanto, eles são dados distintos, pois um check-in tem uma natureza próxima de tempo real, pois indica onde o usuário se encontra em um determinado momento, já uma avaliação não necessariamente é feita enquanto o usuário está no estabelecimento. Com isso, visamos entender se LBSNs diferentes capturam informações similares na modelagem dos dados via redes de interesse.

Nosso estudo indica que o Google Places e Foursquare apresentam padrões similares quando usados para investigar o comportamento urbano. Esse resultado é interessante, pois as LBSNs contêm aspectos diferentes quanto à regularidade e no propósito de seus usos. Além disso, a influência de fatores socioeconômicos e geográficos foi investigada, e seu impacto nas redes de interesse estudadas apresentou uma grande semelhança.

O restante do artigo está organizado da seguinte maneira. A Seção 2 apresenta os trabalhos relacionados. A Seção 3 explica as bases de dados usadas e suas características. Na Seção 4, são apresentados os métodos utilizados nesse trabalho. A Seção 5 discute os resultados. Finalmente, a Seção 6 apresenta a conclusão desse artigo.

2. Trabalhos Relacionados

Nesta seção abordamos alguns estudos prévios que exploraram o potencial de se usar LBSNs para a compreensão do comportamento urbano e a investigação da dinâmica das cidades. O trabalho de [Silva et al. 2013] investigou a possibilidade de se usar dois sistemas diferentes, um do Instagram e outro do Foursquare, coletando dados para os quais a localização era compartilhada. O artigo buscou entender se a informação obtida poderia ser complementar e/ou similar para as duas bases, nos padrões relacionados às dinâmicas das cidades e comportamento urbano. A partir disso, concluíram que os dois datasets proporcionavam informações compatíveis e complementares, no qual, por exemplo, um *check-in* do Foursquare poderia trazer a informação da categoria de um estabelecimento comentado em uma publicação do Instagram, e capturavam ainda aspectos urbanos de forma similar, como as regiões mais populares das cidades.

Em [Martí et al. 2019], os autores exploram o potencial do uso de dados de LBSNs como Foursquare, Twitter, Google Places, Instagram e Airbnb para pesquisas de fenômenos urbanos, reconhecendo não só os benefícios mas também os desafios que o uso dessas fontes de dados acarreta. O estudo apresenta pesquisas que utilizam dados das LBSNs para analisar a dinâmica da cidade e propõe uma metodologia para recuperação, seleção, classificação e análise de dados. Também identifica as principais linhas de investigação temáticas com base nas variáveis de dados fornecidas por estas plataformas, com o objetivo de oferecer um quadro abrangente para o estudo dos fenômenos urbanos através de dados LBSNs.

Outros trabalhos buscam entender quais fatores podem estar associados aos padrões de movimento das pessoas. Por exemplo, [Cheng et al. 2021] usou dados geolocalizados do Twitter para compreender as movimentações dos usuários. Os autores associaram essas informações espaciais com as características econômicas dos usuários, dos aspectos geográficos das regiões frequentadas, bem como seu posicionamento dentro da rede social e a linguagem usada em seus *check-ins*. Assim, identificaram a presença do modelo Lévy Flight em relação à mobilidade, no qual distâncias curtas são percorridas com mais frequência, e distâncias maiores acontecem de forma mais rara. Apresentaram, também, a influência da densidade populacional e da popularidade na rede social para com as distâncias percorridas pelos usuários. Além disso, empregaram o uso da análise de sentimento para agregar contexto às avaliações e aos padrões relacionados aos usuários.

O estudo realizado em [Nolasco-Cirugeda and García-Mayor 2022] mostra como os dados do Foursquare, Twitter e Google Places são cruciais para analisar o uso do espaço urbano e a dinâmica social. Ao detalhar pesquisas e estudos de caso pioneiros na última década, o artigo mostra como os dados da LBSN oferecem informações importantes sobre a vida urbana, auxiliando na compreensão da dinâmica social e intervenções urbanas específicas. A pesquisa emprega várias abordagens analíticas em escala para lidar com diversas questões urbanas, desde o dinamismo da vizinhança social até o turismo e preferências de infraestrutura verde, destacando o potencial analítico abrangente dos dados LBSN.

O trabalho [Skora et al. 2022] examinou se as informações extraídas de dados do Foursquare poderiam se assemelhar as informações divulgadas pela WTO (World Tourism Organization). O estudo encontrou o potencial de se usar LBSNs para facilitar a compreensão dos movimentos turísticos em escalas maiores e com mais detalhes do que por fontes tradicionais, apesar de limitações associadas às LBSNs, como o uso predominante de pessoas jovens com acesso à internet.

Já o estudo [Huang and Butts 2023] investigou diversas características socioeconômicas e buscou entender quais delas impactavam a migração entre condados nos Estados Unidos. Esse trabalho examinou a hipótese de que migrações aconteciam entre regiões similares, propondo uma teoria de segregação e imobilidade social em relação à essas movimentações. Para as suas análises, os autores utilizam um modelo temporal de grafos. Neste tipo de modelo, configurações específicas de parâmetros permitem uma análise do comportamento das pessoas, em simulações onde, por exemplo, a segregação política é desconsiderada.

Como abordado em alguns dos trabalhos relacionados, esse estudo também se baseia no uso de LBSNs, mas com um foco em entender se duas LBSNs diferentes podem modelar o comportamento urbano em redes de interesse de forma similar, bem como se o uso de características socioeconômicas são suficientes para entender e explicar o interesse das pessoas na escolha das áreas urbanas que elas frequentam.

3. Fontes de Dados

Esta seção descreve como foram obtidos os dados para a modelagem das redes de interesse de usuários localizados em diferentes bairros da cidade de Curitiba. Esta foi escolhida por ser uma cidade com a qual os autores estão mais familiarizados. A seção detalha ainda as características dos *datasets* resultantes, bem como as informações sobre a coleta das características socioeconômicas de cada bairro analisado.

3.1. *Datasets* utilizados

Google Places: O *dataset* do Google Places foi construído extraíndo-se avaliações realizadas por usuários da rede social Google Plus em estabelecimentos cadastrados no serviço Google Maps. Os dados¹ foram disponibilizado pelos autores de [He et al. 2017] e [Pasricha and McAuley 2018] para pesquisas. Eles buscavam viabilizar o estudo da LBSN, abrangendo o mundo todo. A partir desses estudos, foram selecionados os dados da cidade de Curitiba, para formar o nosso *dataset*. Esses dados contêm informações detalhadas como o nome do usuário, informações sobre seu nível de educação e emprego, texto da avaliação (que pode estar em diversas línguas), nota da avaliação, momento em que a avaliação ocorreu, identificador único do usuário. Há ainda informações específicas do local avaliado, como por exemplo, o nome do estabelecimento, categoria à qual o local pertence, horário de funcionamento, telefone de contato da empresa, endereço, latitude e longitude do estabelecimento.

Foursquare: Usamos, também, fonte de dados, ofertada publicamente, extraída da rede social baseada em localização Foursquare, através de check-ins da plataforma divulgados na rede social Twitter. O nosso *dataset* do Foursquare é composto por *check-ins* de usuários na cidade de Curitiba. Ele inclui informações como a data do *check-in*, o local onde ocorreu, a categoria a qual o local pertence e o identificador único do usuário. Os dados estão bem distribuídos ao longo de 3 meses, de Abril a Junho de 2014. O *dataset* é composto por 5.116 usuários que fizeram 53.253 *check-ins* em 8.523 lugares diferentes dentro da cidade.

3.2. Caracterização dos *Datasets*: Google Places × Foursquare

No Google Places existem 685 categorias distintas de estabelecimentos, como, por exemplo, restaurantes, hotéis, bares, cafeterias, loja de carros usados, parques, shoppings, padarias, universidades, cinemas e escolas. Na Figura 1a é possível visualizar as 50 categorias mais avaliadas, com um destaque especial para locais de lazer, como restaurantes de variadas especialidades. Os dados contêm 8.372 avaliações de 4.909 usuários em 2.213 estabelecimentos. Esses dados estão distribuídos ao longo dos anos de 2010 até 2014, com grande concentração nos anos de 2011-2014, sendo 2013 o ano com o maior número de avaliações, com um total de 4.003 avaliações. Já o Foursquare contém 368 categorias

¹https://cseweb.ucsd.edu/~jmcauley/datasets.html#google_local

distintas, de maneira similar ao Google Places, e as 50 categorias com mais avaliações são mostradas na Figura 1b.



Figura 1. Word Cloud das 50 Categorias Mais Avaliadas na Cidade de Curitiba

Ao comparar as categorias mais avaliadas de ambas as LBSNs, é evidente a diferença nas categorias dos dados do Google Places em relação aos dados do Foursquare. Por exemplo, o Foursquare traz informações sobre as residências e locais de trabalho dos usuários, algo não presente no Google Places.

Como este trabalho é focado na modelagem, através de redes de interesse, de estabelecimentos que as pessoas frequentam, verificamos também quantas vezes um determinado usuário costuma avaliar para cada um dos dois *datasets*, através do identificador único do usuário em cada plataforma.

Analizando o comportamento dos usuários do Google Places com presença em Curitiba, 23% deles avaliaram pelo menos 2 vezes, 4,7% avaliaram pelo menos 5 vezes e 1,49% avaliaram pelo menos 10 vezes. Portanto, esse *dataset* apresenta uma baixa recorrência em seu uso pelos usuários. Buscamos entender, também, os intervalos entre as avaliações dos usuários. Entre todas as avaliações registradas na cidade, analisando os usuários isoladamente, 50,3% aconteceram em um intervalo menor do que 6 horas, 2,93% aconteceram entre 6 e 24 horas, 9,45% aconteceram entre 1 dia e 1 semana, e 37,3% aconteceram em um intervalo maior do que 1 semana. Assim, verifica-se um uso variado dessa LBSN pelos seus usuários. A maioria das avaliações aconteceram em intervalos curtos, entretanto, ocorreu, também, uma quantidade significativa de avaliações em intervalos maiores do que uma semana.

Assim como o Google Places, investigamos o potencial do Foursquare para uma modelagem através de redes de interesse. Considerando seus usuários, 76,9% fez *check-in* pelo menos 2 vezes, 47,6% fez *check-in* pelo menos 5 vezes, e 28,1% fez *check-in* pelo menos 10 vezes. Diferente do Google Places, os usuários do Foursquare apresentam uma recorrência maior no uso do serviço, bem como são caracterizados por intervalos entre avaliações mais bem distribuídos. De forma a entender melhor os comportamentos dos usuários, investigamos também os intervalos entre *check-ins* dos usuários. Entre todos os *check-ins* realizados em Curitiba, considerando os usuários separadamente, 36,8% aconteceram em um intervalo menor do que 6 horas, 25,5% aconteceram entre 6 e 24 horas, 28,7% aconteceram entre 1 dia e 1 semana, e 9% aconteceram em um intervalo maior do que 1 semana. É possível notar que as avaliações no Foursquare se concentram em um período muito mais curto quando comparado ao Google Places.

Portanto, além das diferenças nas categorias, identificam-se diferenças nos dados coletados das LBSNs em relação ao uso, principalmente no quesito da regularidade no emprego por parte dos usuários, das ferramentas disponibilizadas por cada LBSN.

É importante ressaltar que o uso de LBSNs deve sempre estar acompanhado de uma compreensão sobre as consequências de sua aplicabilidade. Por exemplo, os dados do Google Places e do Foursquare podem não representar adequadamente as redes de interesse da população curitibana, pois os usuários dessas plataformas são, frequentemente, pessoas jovens e com acesso a internet móvel. Porém, ainda permite *insights* sobre o comportamento desse grupo que pode ou não representar padrões mais gerais. Observe, também, que esses dados foram coletados há 10 anos, e, assim sendo, os resultados da nossa análise podem ser diferentes quando comparados com *datasets* mais recentes. Entretanto, a metodologia é geral e em razão das fontes de dados utilizadas nesse trabalho estarem disponíveis publicamente, elas estabelecem um bom ponto de partida para explorar abordagens para a computação urbana.

3.3. Dados socioeconômicos e eleitorais

As informações dos aspectos socioeconômicos de cada bairro de Curitiba foram adquiridas por meio do Censo Demográfico Brasileiro de 2010 realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE)². Esta pesquisa nacional teve o objetivo de retratar a população brasileira com suas características socioeconômicas e também fundamentar o planejamento público e privado para a década entre 2010-2020. De forma a entender as diferenças entre os bairros de Curitiba, extraímos os dados da renda média mensal e da composição racial, formada pelas categorias: Branca; Preta; Amarela; Parda; Indígena. Encontramos essas informações para todos os bairros de Curitiba.

Para coleta dos dados eleitorais, usamos os dados disponibilizados pelo Tribunal Regional Eleitoral (TRE)³ em relação ao 2º turno das eleições para presidente em Curitiba, para o ano de 2014. A escolha desses dados é motivada pelos trabalhos de [Liu et al. 2019] e [Huang and Butts 2023], nos quais, a discrepância nos percentuais de votos para um candidato presidencial nos Estados Unidos nos anos de eleições federais no período de 2004-2016, e em 2008, respectivamente, foi usada para avaliar a polarização política entre regiões. Por exemplo, um alto percentual de votos para democratas em uma região e um baixo percentual em outra, resultava numa maior polarização.

Os dados do TRE contêm as informações da quantidade total de eleitores e a quantidade que votou em cada candidato, em nível de local de votação organizados em 10 zonas eleitorais. Para extrair os dados para o nível de bairro, buscamos uma fonte da prefeitura de Curitiba que relacionasse os locais de votação aos bairros, mas encontramos apenas para os anos de 2012 e 2016. Portanto, para determinar o percentual de eleitores que votaram em determinado candidato em cada bairro, usamos a API do Google Maps para associar os endereços dos locais de votação, de um total de 418 locais de votação, a seus respectivos bairros, exigindo que os resultados da API fossem da cidade de Curitiba. Com esse método, conseguimos associar todos os locais de votação aos seus respectivos bairros. Entretanto, alguns bairros ficaram sem informação relativa aos dados eleitorais. São eles: Centro Cívico, Campina do Siqueira, Alto Da Rua XV, Riviera, São Miguel,

²<https://sidra.ibge.gov.br>

³<https://www.tre-pr.jus.br/eleicoes/eleicoes-anteriores/eleicoes-2014>

Caximba, Lamenha Pequena, São João, Cascatinha.

Os dados socioeconômicos analisados neste trabalho e os dados eleitorais coletados para os experimentos estão mostrados na Figura 2. A distribuição da renda média nos bairros de Curitiba está na Figura 2a. Para representar a composição racial, a Figura 2b apresenta a porcentagem de pessoas declaradas *Pretas* para cada bairro de Curitiba. Com os dados eleitorais dos bairros, foi contabilizado o número de eleitores que votaram naquele bairro e calculado o percentual que votou na candidata presidencial *Dilma Rousseff*. A polarização política pode ser visualizada na Figura 2c. É possível perceber na Figura 2c que alguns dos bairros de Curitiba estão em regiões periféricas e, para aqueles que não apresentaram dados, provavelmente seus moradores votaram em bairros próximos que pertencem às mesmas zonas eleitorais. Ademais, outros bairros, sem informações eleitorais encontradas nesse trabalho, como o Centro Cívico, são pequenos, apesar de estarem em regiões centrais. Uma explicação possível para a ausência de dados é que os habitantes desses bairros votaram, também, em bairros próximos. Porém, outra possibilidade está na imprecisão do método empregado para associar os locais de votação à seus respectivos bairros.

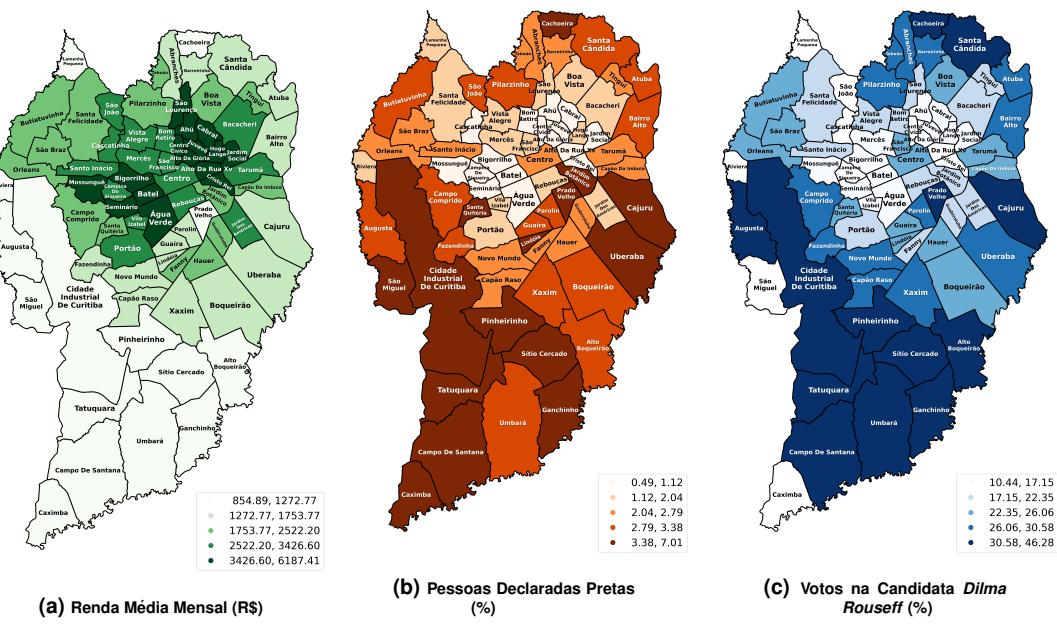


Figura 2. Distribuição das Características Socioeconômicas em Curitiba

Em 2010, a variação de renda entre os bairros de menor e maior renda girava em torno de 7 vezes e a população que se declara preta não ultrapassava 7,2% do total. O percentual de eleitores da candidata *Dilma Rousseff*, no ano eleitoral de 2014, não alcançou 45% em nenhum dos bairros analisados. Nos bairros mais próximos ao centro, houve maior concentração de renda, de pessoas brancas e menor percentual de votação na candidata. O maior percentual de votação dela ocorreu na região mais afastada (em especial na região sul), em bairros de menor poder aquisitivo e onde há maior participação da população autodeclarada preta.

4. Redes de Interesse: Modelagem e Análise

Esta seção detalha como, neste trabalho, as redes de interesse são modeladas através dos dados das avaliações do Google Places e do Foursquare, bem como será verificada a similaridade entre estas redes. A seção descreve ainda o processo usado para averiguar quais fatores podem estar associados com as conexões mais fortes da rede.

4.1. Construção das Redes de Interesse

Para a elaboração da rede de interesse dos usuários do Google Places, utilizamos apenas os dados dos usuários que avaliaram pelo menos 2 vezes, independente do estabelecimento, resultando em uma análise com 1.127 usuários e no conjunto de 4.590 avaliações realizadas por eles. Com essas informações, construímos um grafo não dirigido ponderado $G = (V, E)$, no qual o conjunto V de nós identifica os bairros de Curitiba, e uma aresta $e_{i,j} \in E$ conecta o bairro $v_i \in V$ ao bairro $v_j \in V$, com peso $w_{i,j} \in \mathbb{N}$ indicando a quantidade de usuários que têm avaliações de estabelecimentos localizados em ambos os bairros. Estas arestas representam então o interesse de usuários por dois bairros distintos. A rede construída conta também com *self-loops*, ou seja, uma aresta $e_{i,i} \in E$ conecta um bairro $v_i \in V$ a ele mesmo, nesta aresta o peso $w_{i,i} \in \mathbb{N}$ contabiliza a quantidade de usuários que avaliou estabelecimentos localizados naquele bairro pelo menos duas vezes. Esta construção analisa redes de interesse, portanto, considera apenas arestas não nulas. Desse modo, obtemos um grafo para o Google Places com 75 nós, contendo 1.287 arestas, sendo 53 *self-loops*.

Utilizando a mesma metologia descrita anteriormente, construimos a rede de interesse dos usuários do Foursquare, considerando 3.933 usuários que avaliaram pelo menos 2 vezes e os 52.033 *check-ins* realizados por eles. Desta forma, foi obtido para o Foursquare um grafo com 75 nós, contendo 1.618 arestas, sendo 58 delas *self-loops*.

A Figura 3 ilustra as redes de interesse formadas com os dados do Google Places (lado esquerdo) e Foursquare (lado direito). Nos grafos mostrados, o tamanho do nó é proporcional ao seu grau, ponderado em relação ao peso das arestas conectadas a ele, e a largura da aresta é proporcional ao seu peso. A diferença visual mais aparente é que na rede proveniente dos dados do Google Places, há um interesse por lugares um pouco mais afastados da região central da cidade, como os bairros Alto Boqueirão e Xaxim, que não está presente na rede modelada pelos dados do Foursquare. Apesar disso, é possível observar uma semelhança visual considerável entre as redes de interesse construídas pelas duas LBSNs estudadas. Em ambos os grafos, podemos visualizar uma grande concentração de avaliações na região central da cidade, em bairros como o Centro, Batel, Água Verde, Rebouças, Alto da Rua XV, onde predominam os setores de comércio varejista, restaurantes e escritórios. Bairros como Santa Felicidade e São Francisco são caracterizados pela ampla presença de restaurantes típicos, bares, hamburguerias e *pubs* casuais com música ao vivo. Outros bairros como Bigorrilho, Centro Cívico, Portão, Seminário, Jardim Botânico, Bacacheri e Cabral, são regiões que oferecem área verde extensa com seus parques e praças, além de oferecer infraestruturas de comércio e lazer, como shopping centers.

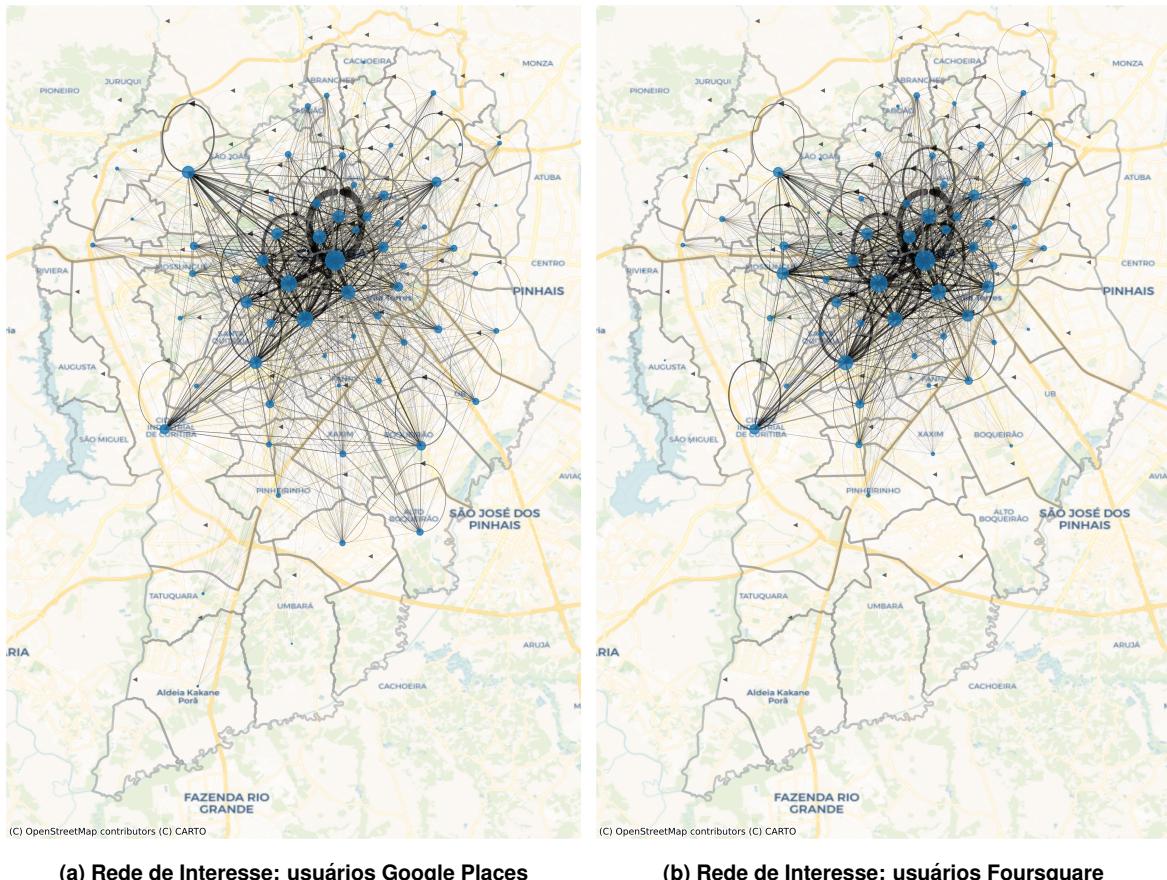


Figura 3. Redes de Interesse na Cidade de Curitiba

5. Resultados

5.1. Similaridade das Redes

Para verificar a similaridade das redes construídas através dos dados das LBSNs analisadas, empregamos a correlação de Pearson, um método que avalia a presença e intensidade de uma relação linear entre dois conjuntos de variáveis. Valores próximos de 1 indicam uma relação linear positiva forte, enquanto valores próximos de -1 indicam uma relação linear negativa forte. Na comparação das redes, associamos o peso da aresta da rede do Google Places com a sua respectiva aresta na rede do Foursquare, e, em seguida, calculamos a correlação de Pearson, usando todo o conjunto de arestas que estão presentes em ambos os grafos. O resultado da correlação foi de 0,875. Testamos, também, outras métricas, como a correlação de Spearman, mas os resultados foram similares. Portanto, existe uma forte indicação, através das arestas, de similaridade dessas redes de interesse.

Além disso, verificamos a similaridade em relação à centralidade por autovetor de ambas as redes. Isto é, computamos a centralidade por autovetor de cada bairro em cada uma das redes de interesse. Esse método indica a importância de um nó na rede tendo por base a importância dos nós com os quais ele está conectado. Para investigar a similaridade, empregamos o coeficiente de correlação de Kendall's Tau. Este coeficiente é utilizado para avaliar a semelhança entre as ordens de dois conjuntos de dados. Valores próximos de 1 indicam uma ordenação semelhante, enquanto valores próximos de

-1 sugerem ordenações bastante distintas. Atribuindo um identificador para cada bairro, comparamos o *ranking* dos bairros, em ordem decrescente, entre as duas redes de interesse. O resultado foi uma correlação de 0.646, indicando uma moderada conexão entre as características de centralidade para as duas fontes de dados. Uma comparação entre o *ranking* dos bairros é exposta na Figura 4, na qual, apresentamos os 20 bairros mais centrais, pela centralidade de autovetor, na rede de interesse do Google Places e seus respectivos *ranks* na rede de interesse do Foursquare. Observa-se, visualmente, pequenas variações na ordem dos bairros na comparação das bases. Entretanto, elas mantêm uma relativa proximidade, representando a moderada relação entre as centralidades, já indicada pelo coeficiente de Kendall's Tau. A diferente natureza das LBSNs pode ser uma possível explicação para essas pequenas diferenças encontradas na questão de quais bairros são mais valorizados pelos seus respectivos usuários. Com essa análise é possível concluir que os 4 bairros mais centrais são os mesmos para as duas LBSNs e, para esses bairros, uma análise a partir de uma rede de interesse poderia ser complementada por dados da outra.

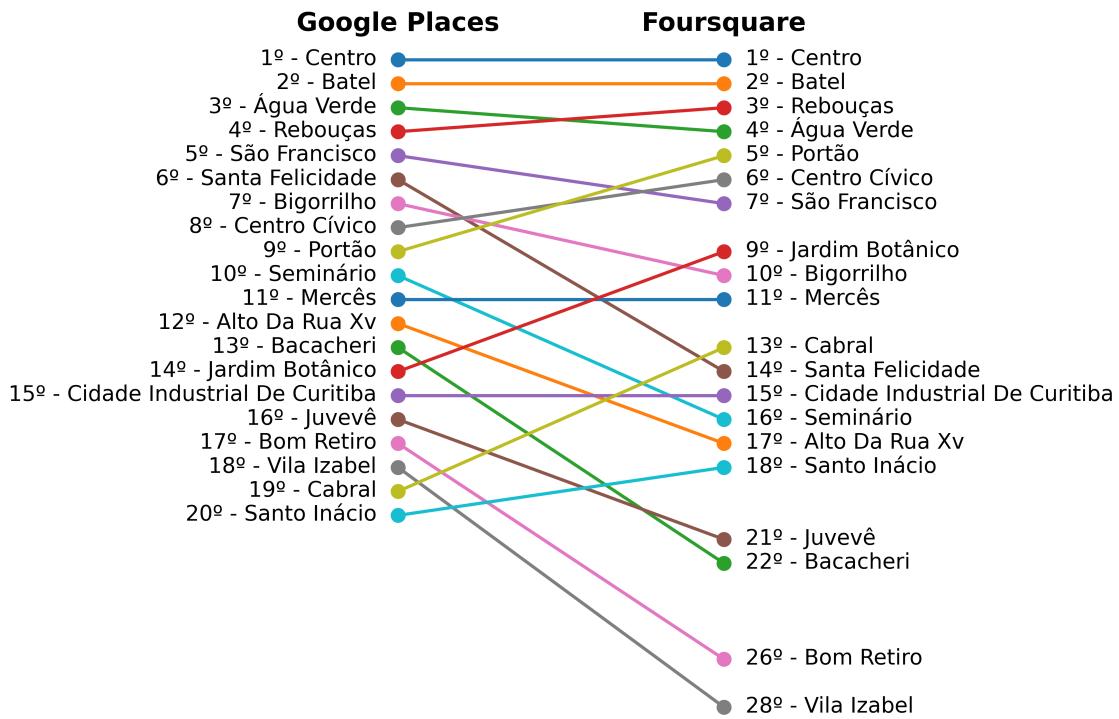


Figura 4. Comparação Dos Bairros Mais Centrais De Curitiba Em Relação À Centralidade Por Autovetor

5.2. Cálculo das Distâncias entre Fatores Socioeconômicos e Políticos

Nesta seção, o objetivo da análise é entender se há fatores extras (socioeconômicos, eleitorais e geográficos) nas regiões analisadas influenciando o interesse dos frequentadores dessas regiões. Para isso, relacionamos os pesos das arestas com as distâncias entre um determinado fator. Os fatores socioeconômicos considerados em cada região foram, a renda média mensal, a composição racial, o fator político foi a polarização e o último fator extra considerado foi a posição geográfica.

Para tanto, usamos a diferença absoluta para a renda média mensal entre dois bairros e a escolha política para eleição presidencial, de maneira similar aos trabalhos de [Huang and Butts 2023] e [Liu et al. 2019]. Para a diferença entre as composições raciais, usamos o mesmo método de [Huang and Butts 2023], no qual a diferença entre os bairros A e B é definida como

$$R_{A,B} = \frac{1}{2} \sum_{i=1}^n \left| \frac{P(A)_i}{P(A)} - \frac{P(B)_i}{P(B)} \right|$$

onde $R_{A,B}$ é a diferença das composições raciais entre os bairros A e B , $P(A)$ representa o tamanho da população do bairro A e $P(A)_i$ o tamanho da população que pertence à i -ésima categoria racial no bairro A . Usamos as categorias definidas pelo Censo Demográfico Brasileiro de 2010: Branca, Preta, Amarela, Parda e Indígena.

De modo a compreender melhor como essas distâncias se aplicam para a cidade de Curitiba, a Figura 5 mostra um sub-grafo das distâncias entre fatores políticos e socioeconômicos para alguns dos bairros mais centrais de ambas as redes de interesse. A cor do nó é proporcional ao valor do fator do aspecto analisado. Desse modo, cores mais escuas representam valores mais altos e cores mais claras estão associadas com valores mais baixos, levando-se em consideração apenas o subconjunto de bairros apresentado. Para a polarização política, cores mais escuras representam um percentual maior de votos para a candidata presidencial *Dilma Rousseff*. Na composição racial, foi escolhido o percentual de pessoas declaradas *Preta* naquele bairro para a representação. Enquanto que na renda média mensal, nós mais escuros estão associados com maiores rendas. A cor da aresta é proporcional à diferença entre os fatores dos bairros, cores mais escuras estão associadas com distâncias maiores, e cores mais claras representam diferenças menores. Nas Figuras 5b e 5c, as diferenças usadas nas arestas consideram, como explicado anteriormente, toda a composição racial e de renda mensal, respectivamente.

5.3. Correlações Entre Redes de Interesse

Para compreender melhor os fatores associados com as regiões de interesse dos freqüentadores, relacionamos cada peso de aresta com a sua distância nas características analisadas (renda média mensal, composição racial, polarização política, posição geográfica), para ambos os grafos. Utilizando a correlação de Spearman, os resultados para ambas as redes, considerando a informação tanto das redes de interesse completas (mostradas na Figura 3) quanto de redes de interesse filtradas, são apresentados na Tabela 1. Para a distância geográfica, usamos os centroides de cada bairro e calculamos a distância usando a biblioteca *geopandas* [Jordahl et al. 2020] do Python, onde transformamos as coordenadas geográficas de latitude e longitude do sistema de coordenadas WGS84, para a projeção plana UTM22S que traz a projeção mais precisa da região onde se encontra Curitiba, e calculamos a distância geográfica entre dois bairros, olhando para os seus centroides, em metros. A biblioteca utiliza a distância Euclidiana para os pontos na projeção plana e, baseado na projeção, retorna a distância geográfica entre os centroides dos bairros em metros.

Nas redes de interesse completas, conforme já destacado, observamos o impacto de todas as arestas encontradas. Já na rede de interesse filtrada, consideramos apenas as arestas com peso superior ou igual a 5, focando apenas nas áreas conectadas por interesse mais elevado. Para a rede de interesse filtrada, a modelagem do Google Places apresentou

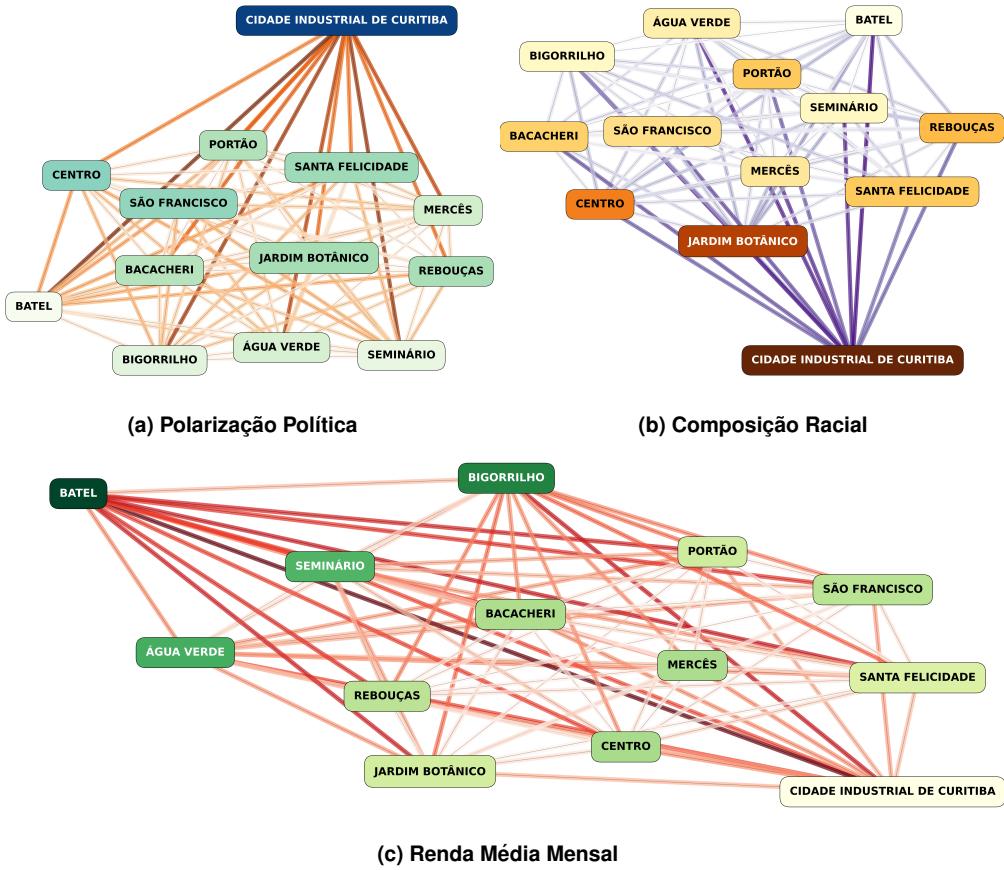


Figura 5. Sub-Grafo das Distâncias Entre Fatores socioeconômicos

apenas 418 arestas do total de 1.287 (32,5%) arestas da rede completa, enquanto os dados do Foursquare apresentaram 1.169 arestas do total de 1.618 (72,2%).

Os resultados indicam que os fatores de renda média mensal, composição racial e polarização política não ajudam a explicar o interesse de frequentadores por determinadas regiões. Isso significa que as pessoas não costumam frequentar, necessariamente, lugares que sejam similares em renda, raça, ou opiniões políticas, quando seus interesses são modelados por essas LBSNs. Apesar da composição racial apresentar um coeficiente de -0.25 , o maior valor absoluto entre os fatores de renda média, composição racial e polarização política, quando analisando a rede de interesse completa do Google Places, pode-se perceber que é um valor inferior quando comparado ao impacto da distância geográfica em todos os cenários analisados na Tabela 1. Isso indica que quanto maior a distância geográfica entre dois bairros, a tendência é de se encontrar arestas com pesos menores. No mais, é possível observar que a distância geográfica tem um peso maior na rede modelada pelo Foursquare. Uma explicação plausível é a diferente natureza no uso das LBSNs pelos seus usuários. Entretanto, o resultado com a maior influência da distância geográfica já era esperado, pois diversos outros trabalhos [Cheng et al. 2021] [González et al. 2008] [Rhee et al. 2008] [Brockmann et al. 2006] explicam a tendência das distâncias percorridas pelos usuários em LBSNs, e em outros formatos de dados similares, apresentarem uma maior frequência em distâncias curtas e serem mais raras em distâncias maiores como demonstramos.

Tabela 1. Resultados das Correlações de Spearman com os Fatores Analisados

	Google Places Completo	Google Places Filtrado	Foursquare Completo	Foursquare Filtrado
Renda média mensal	-0.09	-0.10	0.06	-0.04
Composição Racial	-0.25	-0.19	-0.11	-0.20
Polarização Política	-0.23	-0.14	-0.09	-0.15
Distância Geográfica	-0.38	-0.30	-0.55	-0.44

6. Conclusão

Entender o comportamento urbano é uma tarefa desafiadora. Nesse trabalho, investigamos se diferentes LBSNs poderiam fornecer informações similares ao modelar o interesse de usuários por áreas geográficas. Concluímos, através de uma análise nas características dos dois *datasets* analisados (Google Places e Foursquare) e dados coletados em Curitiba, que os grafos obtidos a partir das escolhas dos usuários, denominados neste trabalho de redes de interesse, modelam a dinâmica nas ações de seus usuários de maneira bastante parecida, em especial para as conexões entre bairros mais centrais da cidade.

Além disso, examinamos a possibilidade de compreender as escolhas dos usuários de LBSNs, em relação ao seu interesse por regiões urbanas, usando características socioeconômicas, de polarização política e distância geográficas das regiões frequentadas por eles. Desse modo, descobrimos que, para os dados analisados, fatores como renda média, composição racial e polarização política das diferentes regiões analisadas (bairros de Curitiba) não influenciam a escolha por essas regiões. Entretanto, inferimos que a distância geográfica entre os bairros analisados pode atuar como um limitante em suas interações, pois existe uma tendência de se frequentar lugares que sejam mais próximos.

Em trabalhos futuros, seria interessante explorar *datasets* mais recentes, bem como outras cidades, tanto para comparar se as diferentes LBSNs continuam fornecendo informações similares, quanto para investigar quais fatores influenciam as redes de interesse ao longo do tempo. Desse modo, poderiam ser investigados outros fatores, como a cultura, os tipos de estabelecimento, o conteúdo das avaliações realizadas, estudando, por exemplo, sentimentos e tópicos expressados, entre outros. Estes novos fatores complementariam a análise, ajudando a responder perguntas como se os locais que as pessoas frequentam no cenário urbano são similares em cultura, se os usuários costumam frequentar os mesmos tipos de estabelecimentos em diferentes bairros, ou se cada bairro é mais caracterizado por uma classe específica de locais, e se as regiões mais conectadas nas redes de interesse formadas compartilham de um mesmo sentimento por parte de seus usuários. Isto permitiria um melhor entendimento do fenômeno do comportamento urbano e a dinâmica nas cidades em relação às regiões que as pessoas costumam frequentar. Com uma melhor compreensão sobre os fatores associados com os interesses das populações de cada cidade, essas análises poderiam integrar sistemas de recomendação, trazendo o aspecto das características de cada bairro. Ademais, esse tipo de investigação tem o potencial de indicar aos órgãos públicos, quais regiões estão mais interligadas entre si, permitindo políticas públicas voltadas, por exemplo, ao combate a epidemias ou à promoção da integração social entre regiões antes pouco conectadas.

7. Agradecimentos

Esta pesquisa foi parcialmente apoiada pelo projeto SocialNet (processo 2023/00148-0 da Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (processos 313122/2023-7, 314603/2023-9 e 441444/2023-7).

Referências

- Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.
- Cheng, Z., Caverlee, J., Lee, K., and Sui, D. (2021). Exploring millions of footprints in location sharing services. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):81–88.
- Ferreira, A. P., Silva, T. H., and Loureiro, A. A. (2020). Uncovering spatiotemporal and semantic aspects of tourists mobility using social sensing. *Computer Communications*, 160:240–252.
- Ferreira, A. P. G., Silva, T. H., and Loureiro, A. A. F. (2015). Beyond sights: Large scale study of tourists' behavior using foursquare data. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1117–1124.
- González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.
- He, R., Kang, W.-C., and McAuley, J. (2017). Translation-based recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17. ACM.
- Huang, P. and Butts, C. T. (2023). Rooted america: Immobility and segregation of the intercounty migration network. *American Sociological Review*, 88(6):1031–1065.
- Jordahl, K., den Bossche, J. V., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L. J., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, and Leblanc, F. (2020). geopoland/gopoland: v0.8.1.
- Ladeira, L., Souza, A., Filho, G. R., Silva, T. H., and Villas, L. (2019). Serviço de sugestão de rotas seguras para veículos. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 608–621, Porto Alegre, RS, Brasil. SBC.
- Liu, X., Andris, C., and Desmarais, B. A. (2019). Migration and political polarization in the u.s.: An analysis of the county-level migration network. *PLOS ONE*, 14(11):e0225405.
- Martí, P., Serrano-Estrada, L., and Nolasco-Cirugeda, A. (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, 74:161–174.
- Nolasco-Cirugeda, A. and García-Mayor, C. (2022). Social dynamics in cities: Analysis through lbsn data. *Procedia Computer Science*, 207:877–886.
- Pasricha, R. and McAuley, J. (2018). Translation-based factorization machines for sequential recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18. ACM.
- Rhee, I., Shin, M., Hong, S., Lee, K., and Chong, S. (2008). On the levy-walk nature of human mobility. In *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*. IEEE.
- Santala, V., Miczevski, S., de Brito, S. A., Baldykowski, A. L., Gadda, T., Kozievitch, N., and Silva, T. H. (2017). Making sense of the city: Exploring the use of social media data for urban planning and place branding. In *Anais do I Workshop de Computação Urbana*, Porto Alegre, RS, Brasil. SBC.
- Senefonte, H. C. M., Delgado, M. R., Lüders, R., and Silva, T. H. (2022). Predictour: Predicting mobility patterns of tourists based on social media user's profiles. *IEEE Access*, 10:9257–9270.
- Silva, T. H., de Melo, P. O. S. V., Almeida, J. M., and Loureiro, A. A. F. (2017). Uma fotografia do instagram: Caracterização e aplicação. *Revista Brasileira de Redes de Computadores e Sistemas Distribuídos*.
- Silva, T. H. and Loureiro, A. A. (2016). Computação urbana: Técnicas para o estudo de sociedades com redes de sensoriamento participativo. *Sociedade Brasileira de Computação*.
- Silva, T. H. and Silver, D. (2024). Using graph neural networks to predict local culture. *arXiv*.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013). A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *Proc. ACM SIGKDD Int. Workshop on Urban Computing (UrbComp'13)*, Chicago, USA.
- Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., and Quercia, D. (2019). Urban computing leveraging location-based social network data: A survey. *ACM Computing Surveys*, 52(1):1–39.
- Silver, D. and Silva, T. H. (2023). Complex causal structures of neighbourhood change: Evidence from a functionalist model and yelp data. *Cities*, 133:104130.
- Skora, L. E., Senefonte, H. C., Delgado, M. R., Lüders, R., and Silva, T. H. (2022). Comparing global tourism flows measured by official census and social sensing. *Online Social Networks and Media*, 29:100204.
- Veiga, D. A. M., Frizzo, G. B., and Silva, T. H. (2019). Cross-cultural study of tourists mobility using social media. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*, WebMedia '19, page 313–316, New York, NY, USA. Association for Computing Machinery.