Estimativa de Distância em Redes Wi-Fi usando Super-sniffers

Pedro Videira Rubinstein, Fernando Dias de Mello Silva, Mohammad Imran Syed, Anne Fladenmuller, Marcelo Dias de Amorim, Luís Henrique Maciel Kosmalski Costa

Estimativa de Distância em Redes Wi-Fi usando Super-sniffers

Pedro Videira Rubinstein[1], Fernando Dias de Mello Silva[1], Mohammad Imran Syed[2], Anne Fladenmuller[2], Marcelo Dias de Amorim[2], Luís Henrique M. K. Costa[1]

[1]GTA – Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.504 – 21.941-972 – Rio de Janeiro – RJ – Brazil

[2]Sorbonne Université, CNRS, LIP6
75005 – Paris – França

{prubinstein,fernandodias,luish}@gta.ufrj.br

{mohammad-imran.syed,anne.fladenmuller,marcelo.amorim}@lip6.fr

Abstract. *Wi-Fi sniffers are devices responsible for passive data collection in wireless networks. They find application in, among others, distance estimation and location processes through the use of the RSSI metric (Received Signal Strength Indicator). Unfortunately, RSSI is sensitive to small variations in the environment and, without treatment, fails to provide a reliable distance measure. In this paper, we formulate a new approach that employs redundancy through a super-sniffer, which are multiple co-located sniffers, to enhance the distance classification process through two models: a k-Nearest Neighbors (k-NN) based and a log-distance path loss (LDPL) based model. We apply the formulated strategy to our experimental dataset and demonstrate that the method can generate a model with an average accuracy of 91.73% in addition to determining a saturation point for gains related to increasing the super-sniffer size.*

Resumo. *Sniffers Wi-Fi são dispositivos responsáveis por realizar a coleta passiva de pacotes em redes sem-fio. Sniffers possuem aplicações, entre outras, em processos de estimativa de distância e localização através do uso da métrica RSSI (Received Signal Strength Indicator). Porém, o RSSI é sensível a pequenas pertubações no ambiente e, sem tratamento, não fornece uma estimação de distância confiável. Este artigo formula uma nova abordagem que utiliza redundância através de um super-sniffer que consiste de múltiplos sniffers colocalizados para melhorar o processo de classificação de distância através de dois modelos, baseados em k-Nearest Neighbors (k-NN) e em log-distance path loss (LDPL). Aplica-se a estratégia formulada a um conjunto de dados experimental próprio e mostra-se que o método é capaz de gerar um modelo com acurácia média de 91,73%, além de determinar um ponto de saturação para os ganhos relacionados ao aumento do tamanho do super-sniffer.*

1. Introdução

Sniffers Wi-Fi são computadores equipados com interfaces de rede responsáveis por realizar a coleta passiva dos pacotes transmitidos em redes do padrão IEEE-802.11. A vantagem é que os sniffers são dispositivos de baixo custo e podem capturar os pacotes de forma passiva evitando a geração de tráfego na rede, tornando-os interessantes na

gestão e diagnóstico de redes, em ferramentas de determinação de distância, localização, reconstrução de trajetória e aplicações em Internet das Coisas (IoT – *Internet of Things*).

Diversos estudos que se dedicam à determinação de distância e aplicações relacionadas fazem uso do RSSI [Barai et al. 2017, Verma and Singh 2019]. Contudo, esse valor pode ser afetado devido a obstáculos no caminho e múltiplos percursos, o que diminui a precisão da estimativa. Ademais, a perda de quadros Wi-Fi aumenta a latência para se obter um resultado dessas aplicações. Alguns trabalhos, portanto, condenam o uso do RSSI para medir distâncias [Heurtefeux and Valois 2012, Dong and Dargie 2012]. Porém, devido à sua simplicidade e à ampla disponibilidade desse valor em interfaces de rede, métodos envolvendo RSSI continuam sendo objeto de interesse dos pesquisadores.

Como vários problemas associados ao RSSI estão relacionados à sua inconsistência, um maior número de medições para identificação tende a gerar previsões mais precisas. É possível alcançar essa disponibilidade com a introdução de **redundância** no processo de captura. Sniffers com várias interface de rede, chamados de *super-sniffers*, permitem redundância nas medições do RSSI ao capturarem um mesmo pacote várias vezes em um mesmo canal e em uma mesma região. Esse processo de redundância ainda é pouco explorado na literatura. Métodos atuais utilizam medições sequenciais de um único sniffer para determinar um valor RSSI adequado [R et al. 2021, Jose et al. 2023]. Assim, o uso da redundância ainda pode reduzir a quantidade de pacotes necessários para se obter estimativas confiáveis.

O objetivo deste trabalho é aprimorar a confiabilidade da classificação de distância radial a partir da investigação do impacto do uso de super-sniffers. Concentra-se na estimação discreta de distância, ou seja, em classificar amostras em categorias fixas de distância. A principal contribuição é a introdução e análise da redundância trazida pelos super-sniffers para observar como ela impacta o desempenho de modelos baseados em *k-Nearest Neighbors* (k-NN) e *log-distance path loss* (LDPL). Observa-se que os resultados com super-sniffers são muito superiores a um sniffer individual. Obtém-se um modelo com acurácia média de 91,73%. Além disso, busca-se determinar os limites da introdução de redundância de sniffers e pacotes. É assim possível identificar um tamanho ideal de super-sniffer, fator indispensável no planejamento de medidas em situações reais.

O restante do artigo está organizado da seguinte maneira. A Seção 2 discute os trabalhos relacionados. A Seção 3 apresenta como estimar distância através do uso de super-sniffers, além de discutir a escolha dos modelos. A metodologia usada para aquisição do conjunto de dados, os modelos utilizados e suas particularidades são discutidos na Seção 4. A Seção 5 analisa a correlação entre os dados coletados. A Seção 6 analisa do desempenho e uma comparação entre os modelos k-NN e o modelo de perda de caminho *log-distance*. Por fim, a Seção 7 apresenta as conclusões e pistas de trabalhos futuros.

2. Trabalhos Relacionados

Medições passivas são uma tendência em aplicações de determinação de distância. Diversos trabalhos abordam métodos para se estimar distância com base em medições de RSSI. Uma parcela destes trabalhos levanta problemas de confiabilidade das medições de RSSI, ocasionando erros de estimação [Heurtefeux and Valois 2012, Dong and Dargie 2012, Mishra et al. 2023]. Por esta razão, trabalhos recentes se apoiam em processos que sejam

capazes de aumentar a confiabilidade das medições de RSSI, complementando modelos existentes. Dentre os modelos mais populares na literatura, existem modelos de perda de propagação [Chuku and Nasipuri 2021, R et al. 2021, Jose et al. 2023, Li et al. 2018] ou modelos baseados em *fingerprinting* [Saha et al. 2003, Yiu et al. 2017, Wu et al. 2016], que frequentemente são objeto de interesse de pesquisadores.

Chuku e Nasipuri sugerem e desenvolvem um método de detecção e remoção de valores RSSI discrepantes como forma de melhorar a qualidade das capturas utilizando agrupamentos [Chuku and Nasipuri 2021]. Venkatesh R *et al.* usam filtros baseados em média e mediana para melhorar a confiabilidade do modelo em redes Bluetooth [R et al. 2021]. Jose *et al.* analisam como a utilização de diversos filtros, incluindo filtro de Kalman, impacta a confiabilidade dos modelos [Jose et al. 2023]. Li *et al.* investigam como atualizar em tempo real os parâmeros de controle de propagação, melhorando a estabilidade do modelo [Li et al. 2018].

Saha *et al.* investigam a viabilidade de uso do *fingerprinting* com diferentes modelos de aprendizado de máquina, apresentando bons resultados para o k-NN [Saha et al. 2003]. Yiu *et al.* comparam diversos modos de gerar *fingerprints*, e investigam como o desempenho se deteriora com o tempo [Yiu et al. 2017]. Wu *et al.* usam filtros de partículas para suavizar valores discrepantes e melhorar a qualidades das *fingerprints* [Wu et al. 2016].

A escolha dos pacotes monitorados não está associada apenas ao desempenho de coleta, mas também à privacidade dos usuários. Assim, as aplicações de coleta de pacotes Wi-Fi optam por coletar pacotes públicos, como *probe-requests* e *beacons*. No entanto, a taxa de envio de *probes* pode variar significativamente, podendo ser tão baixa quanto 55 pacotes por hora ou tão alta quanto 2.000 pacotes por hora, dependendo do dispositivo e de seu estado [Gupta et al. 2007, Jaisinghani et al. 2017, Freudiger 2015]. Essa variabilidade apresenta desafios, especialmente considerando a latência exigida por diferentes aplicações. Além disso, a perda de pacotes durante a transmissão também precisa ser considerada ao explorar métodos que forneçam medidas de RSSI confiáveis.

Diferentemente dos trabalhos anteriores, este artigo concentra seus esforços na análise dos impactos da introdução de redundância por meio do uso de super-sniffers, visando aprimorar a qualidade e confiabilidade das medidas de RSSI associadas aos pacotes. Essas melhorias são então aplicadas para aperfeiçoar modelos de classificação de distância baseados em modelos de perda de propagação e *fingerprinting*, de modo a determinar os ganhos trazidos pelo uso de super-sniffers.

3. Super-sniffers e Estimativa Discreta de Distância

Super-sniffer é um termo empregado para descrever um conjunto de sniffers colocalizados que operam de maneira colaborativa [Syed et al. 2022b]. Essa abordagem busca introduzir redundância no processo de captura de pacotes, gerando assim uma representação mais precisa do tráfego da rede. Ao trabalhar em conjunto, esses sniffers proporcionam maior disponibilidade de medidas do RSSI, contribuindo para aprimorar a confiabilidade dos modelos utilizados na estimativa de distância em redes sem fio. Essa introdução deliberada de redundância visa mitigar os desafios associados à variabilidade na taxa de envio de pacotes junto à possibilidade de perda durante a transmissão, resultando em uma maior disponibilidade de informação para cada pacote capturado.

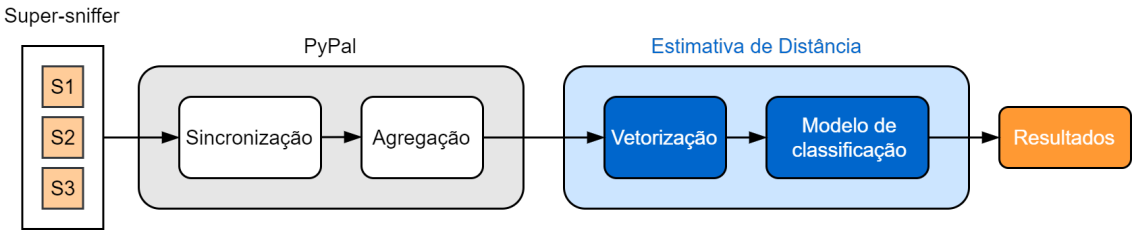


Figura 1. Método de estimativa de distância a partir do uso de super-sniffers.

Super-sniffers realizam a agregação dos *traces* individuais dos sniffers co-localizados. O diagrama dessa arquitetura está presente na Figura 1. Cada *trace* capturado é um arquivo de pacotes capturados por um sniffer (S_1, S_2, S_3, ..., S_N) e contém apenas *probes* e *beacons*. Os *traces* passam por etapas de sincronização e agregação através do software PyPal [Syed et al. 2022c] para associar as capturas de um mesmo pacote por diferentes sniffers e indicar o RSSI detectado por cada dispositivo (um exemplo dessa saída de dados está representada na tabela 1). Para a aplicação de estimação de distância, as medidas são vetorizadas e entregues para um modelo adaptado para serem classificadas, determinando a distância do dispositivo detectado ao super-sniffer.

Decide-se utilizar o super-sniffer para investigar o impacto em modelos com grande popularidade na literatura. Nesse sentido, escolheu-se dois algoritmos da literatura para avaliar o desempenho: O primeiro consiste em uma técnica baseada em *fingerprinting*, composta por um modelo k-NN, enquanto o segundo consiste em um modelo teórico de perda de propagação baseado no modelo log-distância de perda de caminho (LDPL).

4. Experimentos e Coleta de Dados

Este estudo foi conduzido por meio de uma abordagem experimental. Foram coletados dados para formar um *trace* composto por valores RSSI obtidos por super-sniffers. O conjunto de dados foi gerado a partir de um super-sniffer de tamanho 10, composto por Raspberry Pis (RPIs) modelo 4B, conectados a interfaces externas Wi-Fi modelo Alfa AWUS 051NH. Os sniffers individuais que compõem o super-sniffer estavam espaçados por 20 cm e permaneceram estáticos durante toda a coleta (conforme Figura 2).

Realizaram-se um total de seis coletas, durante as quais um RPI "fonte" conectado a um módulo de Wi-Fi similar enviava um total de 10 pacotes por segundo. Em cada uma das coletas, a fonte foi posicionada a uma distância fixa do super-sniffer. Assim, obtiveram-se dados para as seguintes distâncias: 1, 10, 20, 30, 40 e 50 metros. Todas as medições foram realizadas no canal 1 em uma área urbana externa a um edifício (campus da Sorbonne Université - Paris) com linha de visada para os dispositivos em todas as distâncias. Cada pacote capturado recebeu um rótulo, correspondente à distância da fonte ao super-sniffer no momento da coleta. Para cada distância, foram coletados 2.500 pacotes em média, o que totaliza 15.000 pacotes no conjunto de dados.

Os dados coletados foram organizados em tabelas referentes a cada distância, como mostrado na Tabela 1. As tabelas foram então manipuladas de modo a gerar os vetores $\vec{V}$ de tamanho $N_s \times N_p$ para uso nos modelos. N_s e N_p são parâmetros de controle e representam o tamanho do super-sniffer (ou seja, a quantidade de sniffers considerados) e o número de pacotes por vetor, respectivamente.

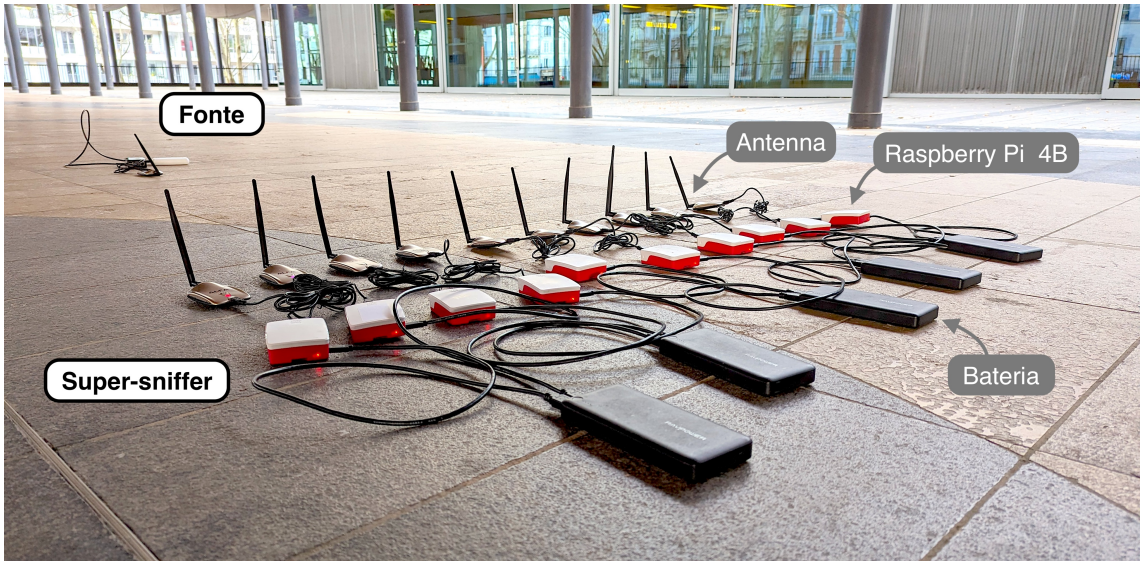


Figura 2. Super-sniffer de tamanho 10 no ambiente de captura dos pacotes.

Tabela 1. Forma de organização dos dados adquiridos para uma determinada distância. Cada coluna representa um sniffer e cada linha representa o nível de RSSI (em dBm) da recepção deste pacote em cada um dos sniffers componentes do super-sniffer. Neste exemplo, $N_s = 10$ e $N_p = 5$. O termo NC indica que o pacote não foi capturado pelo sniffer em questão.

Pacotes	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
1	-21	-21	NC	-21	NC	-21	-7	-9	NC	-21
2	NC	-21	-7	-21	-7	NC	NC	-21	NC	-21
3	-7	-7	-21	NC	-9	-7	NC	NC	-21	NC
4	-21	NC	-21	-21	-7	-21	-21	-7	NC	NC
5	NC	NC	NC	-21	NC	NC	-21	NC	NC	NC

4.1. Modelo de perda de caminho *log-distance*

A perda de caminho Log-Distância (*log-distance path loss*, ou LDPL) é um modelo utilizado para estimar a perda de potência entre um transmissor e um receptor dada sua distância em diversos ambientes e condições [Rappaport 2002, sec. 3.9]. A equação pode ser rearranjada para o cálculo da distância da seguinte forma:

$$D(\text{RSSI}) = D_0 \times 10^{-\frac{\text{RSSI}-\text{RSSI}_0}{10\eta}}, \tag{1}$$

onde RSSI é o RSSI recebido pelo dispositivo, RSSI_0 é o RSSI recebido na distância de referência D_0 e η é um coeficiente de ajuste dessa fórmula para se adequar a diferentes ambientes. Deve-se escolher uma distância D_0 para medir o valor de RSSI_0 correspondente. Assim, com essa fórmula é possível chegar a um valor de distância para cada entrada recebida na Tabela 5. Segundo Rappaport, o valor de η destinado a ambientes urbanos com visada direta varia entre $1,6$ e $1,8$ [Rappaport 2002]. No trabalho de Imran *et al.*, escolheu-se o valor de $\eta = 1,75$ [Syed et al. 2022a].

Caracterização do η – De posse das medidas de RSSI em diferentes distâncias, decidiu-se encontrar o valor de η que melhor se ajusta aos dados coletados. Faz-se um ajuste de

curva com o uso dos dados coletados e a Equação (1) para o conjunto nos dados coletados com variação do η de modo a achar o valor que minimize o erro quadrático médio. Para tornar a otimização mais robusta, os *outliers* de cada classe são removidos da seguinte forma: para cada classe de distância D, é calculada a média $\overline{\text{RSSI}}_D$ e o desvio padrão σ_{RSSI_D}, e são descartados todos os valores de RSSI_i naquela classe tais que $|\text{RSSI}_i - \overline{\text{RSSI}}_D| > z\dot{\sigma}_{\text{RSSI}_D}$, onde z é o z-score. É calculada uma tabela com valores de η para cada z-score, para uso na busca de parâmetros de classificação na Seção 6.

Para chegar ao resultado de distância, separa-se um conjunto de dados de tamanho N_s sniffers e N_p pacotes (assim como a Tabela 1) e calcula-se a distância segundo a Equação (1) para cada medida. O resultado de distância é a média desse conjunto, e o desvio padrão indica a precisão desse resultado. A quantidade de elementos no conjunto final de dados utilizados para estimar a distância será $N_sN_p - N_{\text{NC}}$, onde N_{NC} é a quantidade de medidas não coletadas, que pode variar por resultado.

Para adequar os resultados ao problema de classificação e mitigar os efeitos da precisão, é feita a quantização do resultado de distância para se obter as categorias de cada uma das coletas feitas. A quantização é feita pela função $Q(x)$, onde x é a média do resultado de distância. Para esse problema, a função é definida da seguinte forma:

$$Q(x) = \begin{cases} 1, \text{ se } x \leq 5, \\ a, \text{ se } x \in]a-5, a+5], \quad \forall a \in \{10, 20, 30, 40\}, \\ 50, \text{ se } x \geq 45. \end{cases}$$

Para encontrar os parâmetros que fornecem o melhor resultado, é feita a classificação dos dados para diferentes valores de N_p, N_s e η. Tanto N_s quanto N_p variam entre 1 e 10 enquanto os valores de eta η são aqueles cujo, no resultado da caracterização do η, o z-score varia entre 0,5 e 3,5, com passo de 0,1. Para reforçar os resultados, é feita a validação cruzada por K-folding com $K = 5$ e o desempenho é medido com a parcela de teste de cada iteração.

O desvio padrão da média encontrada por medida pode ser utilizado para indicar a precisão da estimativa e assim descartar resultados de baixa precisão. Para avaliar o ganho de acurácia, será feito o K-folding com $K = 10$ dos resultados para o melhor modelo. Os dados de *teste* (ou seja, $1/K$ dos dados) servirão como amostras para determinar o valor de σ_η na qual, por classe, dados com valores de σ_η maiores representem $X\%$ dos dados de teste. Os valores por classe obtidos serão utilizados para descartar resultados do conjunto de *treino* (ou seja, $K - 1/K$ dos dados) com σ_η maior e a métrica de acurácia é calculada nos dados restantes.

4.2. *k*-Nearest Neighbors

O algoritmo *k-Nearest Neighbors* (k-NN) consiste em um algoritmo de aprendizado de máquina supervisionado capaz de solucionar problemas de classificação. Busca-se classificar os vetores gerados em classes que representam as categorias de distância. O k-NN funciona partindo da avaliação de uma métrica de distância, no caso Euclidiana, entre determinada amostra e seus k vizinhos mais próximos.

O parâmetro k influencia diretamente nos resultados obtidos e é escolhido via validação cruzada por K-folding, com $K = 5$. Esse processo é repetido para diferentes

valores de k. O valor de k que rende a maior acurácia média é o valor utilizado nos modelos finais referentes aos diversos parâmetros de vetor.

Para a entrada de parâmetros no k-NN são escolhidas $N_s \times N_p$ medidas organizadas de forma sequencial, ordenadas primeiro por pacote e depois por sniffer. Para uma medida x_{ij} do pacote i feita pelo sniffer j, o vetor resultante fica ordenado dessa forma:

$$\mathbf{x} = \left[x_{1,1}, x_{1,2}, \ldots, x_{1,N_s}, x_{2,1}, \ldots, x_{N_p-1,N_s}, x_{N_p,1} \ldots, x_{N_p,N_s}\right].$$

Ao se organizar os vetores de forma sequencial, o k-NN se torna capaz de representar padrões na captura do RSSI referente a cada sniffer. Por exemplo, se um sniffer apresenta consistentemente valores de RSSI mais elevados, esse padrão aparece consistentemente em uma coordenada específica. No entanto, como o k-NN considera a proximidade dos vizinhos, o modelo naturalmente tende a mitigar o impacto desse padrão.

Também é importante observar que a escolha dos parâmetros N_p e N_s influencia diretamente os resultados obtidos. O efeito é causado tanto pelas diferenças entre a informação a ser representada por cada vetor, quanto pelo volume disponível. Uma preocupação é a questão de concentração de pacotes perdidos em um vetor, que agem como ruído, e podem atrapalhar o processo de classificação. Para explorar essas de possibilidades, varia-se N_s de modo a testar todas as possíveis combinações para um super-sniffer, e ainda, para cada valor de N_s, varia-se N_p dentro do intervalo de 1 a 25.

5. Análise Exploratória

Nessa seção, serão feitas análises para melhor compreender as características do conjunto de dados com foco nas relações que surgem a partir do uso de um super-sniffer.

5.1. Razão de completude e fração de captura

Para poder avaliar a capacidade de detecção de super-sniffers de diferentes tamanhos, define-se a razão de completude. A razão de completude é expressa pela quantidade de pacotes capturados por um super-sniffer de tamanho x em relação ao total de pacotes capturados por todos os sniffers (ou seja, o super-sniffer de tamanho 10). Para determinar os valores de completude, calcula-se a média entre todas as combinações possíveis de sniffers, a fim de gerar um super-sniffer do tamanho desejado.

O resultado de razão de completude está representado na Figura 3. Os resultados apontam que diferentes subgrupos do super-sniffer original capturam apenas uma fração do total de pacotes, resultando em períodos mais longos de silêncio, e consequentemente, em maior latência na detecção da posição do dispositivo. Observa-se também que, à medida que o número de sniffers é incrementado, os ganhos de completude tornam-se cada vez menores, como pode ser constatado na Figura 3, sugerindo uma tendência de estabilização. Pode-se observar que, para receber mais de 90% dos pacotes transmitidos em qualquer uma das distâncias observadas, deve-se ter pelo menos 4 sniffers em atuação no super-sniffer. Já a partir de 7 sniffers, o ganho de completude é menor em qualquer um dos cenários.

Para o super-sniffer de tamanho 10, quer-se avaliar quantos pacotes foram capturados por um número fixo de sniffers. Assim, é possível ver a disponibilidade de medidas de um mesmo pacote e como ela varia para cada distância. Para isso, define-se a fração

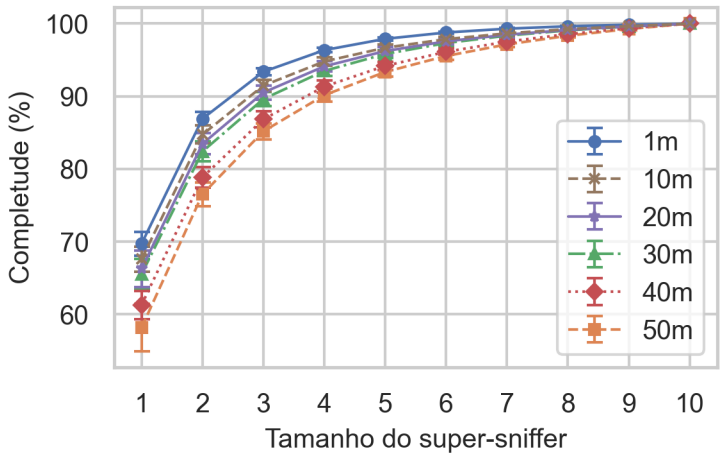


Figura 3. Completude por tamanho do super-sniffer.

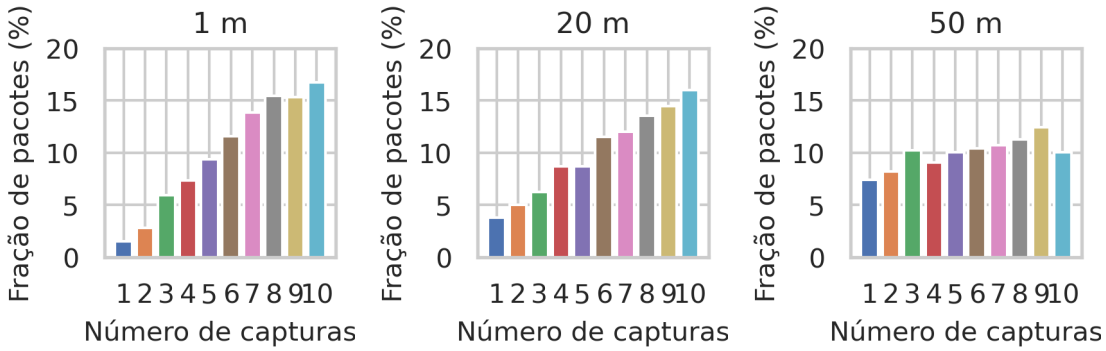


Figura 4. Fração de pacotes capturados pelo super-sniffer por distância. O número de capturas representa quantos sniffers individuais detectaram determinado conjunto de pacotes.

de captura, que é a razão entre a quantidade de pacotes detectados por um número x de sniffers sobre a quantidade total de pacotes detectados.

É possível verificar a fração de captura para diferentes distâncias na Figura 4. No gráfico de 1 metro, observa-se que mais de 15% dos pacotes foram detectados por todos os 10 sniffers, enquanto menos de 5% foram capturados por apenas 1 ou 2 sniffers. Esses números tendem a diminuir e a crescer, respectivamente, dado o aumento da distância. Isso mostra como a distância contribui para uma maior perda de pacotes, porém a redundância de vários sniffers permite que esses mesmos pacotes permaneçam sendo detectados por uma quantidade menor de sniffers. Nota-se que como pacotes são perdidos por boa parte dos sniffers, aumentar o tamanho do super-sniffer pode nem sempre ser positivo. A saturação dos ganhos de completude, induz um aumento da fração de pacotes capturados em um número menor de sniffers, introduzindo mais ruído nos vetores, o que pode levar a erros durante a análise.

5.2. Correlação e distribuição de valores RSSI

Para determinar as características de independência entre as medidas feitas por diferentes sniffers para um mesmo pacote, é feita uma matriz de correlação com os dados coletados.

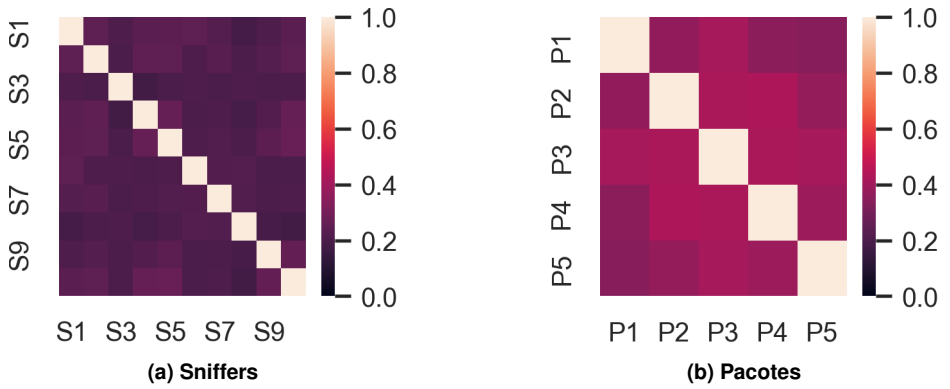


Figura 5. Matrizes de correlação de RSSI.

Essa matriz está representada na Figura 5a. Essa correlação é feita apenas nos casos onde todos os 10 sniffers capturaram um pacote. A análise da correlação entre sniffers revela uma correlação baixa ao analisar um mesmo pacote recebido pelos diferentes sniffers para qualquer distância. Esse fator indica independência entre as medidas dos diferentes sniffers, assim quanto mais sniffers melhor é a representação do resultado.

Para determinar a correlação entre os valores de RSSI de uma sequência temporal de pacotes, também é montada uma matriz de correlação que relaciona as medidas de RSSI entre uma sequência de pacotes para um mesmo sniffer. Essa matriz está representada na Figura 5b. São observadas sequências de 4 pacotes e não são consideradas as sequências na qual houve a perda de pelo menos um pacote. Nesse caso, observa-se que, para pacotes consecutivos, a correlação é maior, o que pode indicar menor contribuição da sequência de pacotes para a devida representação da medida. Comparando os dois gráficos da Figura 5, pode-se observar que a contribuição de uma quantidade maior de sniffers para capturar um mesmo pacote é mais significativa do que reter uma sequência de pacotes para realizar a mesma estimativa. Isso implica que naturalmente é necessário mais pacotes para compensar o desempenho de um super sniffer de determinado tamanho.

A análise da distribuição dos valores RSSI também é importante para determinar a viabilidade do processo de classificação. Mede-se a distribuição a partir das medidas do super-sniffer, excluindo os pacotes perdidos. As distribuições podem ser observadas na Figura 6. Para o conjunto de dados coletados, as distribuições apresentam um padrão discernível entre as diferentes distâncias. O valor médio tende a diminuir e o desvio padrão tende a aumentar. Há uma dificuldade de discernimento visual entre as distâncias de 20, 30 e 40 metros, o que é dado pelas condições do ambiente no momento do teste.

Outro fator relevante a ser considerado diz respeito à distribuição dos pacotes perdidos, ou seja, o ruído no sistema. A Figura 7a é a correlação entre a perda de pacotes entre sniffers, que se observa baixa entre todos os sniffers. Da mesma forma, a Figura 7b mostra a correlação do surgimento de zeros entre sequências de pacote, que é praticamente nula. Finalmente, a Figura 7c indica a taxa de perda de pacotes por sniffer, que coincide com a primeira coluna do gráfico da Figura 3 e mostra que a perda é constante entre os sniffers. Com isso, pode-se observar que o ruído nos dados ocasionado pela perda é não espacial, estacionário e uniforme entre os sniffers.

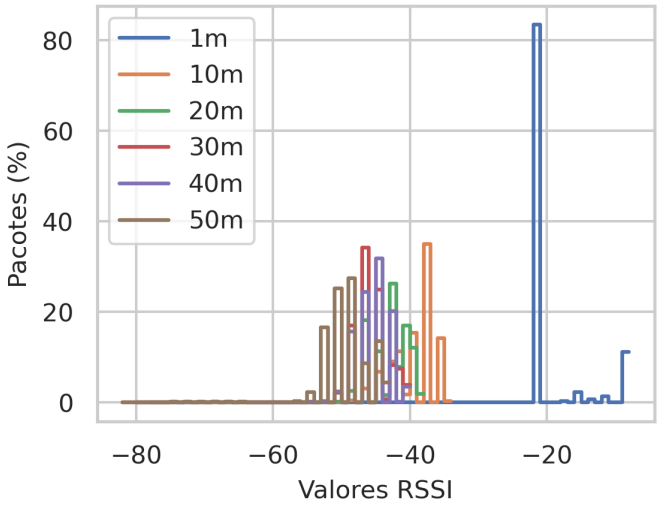
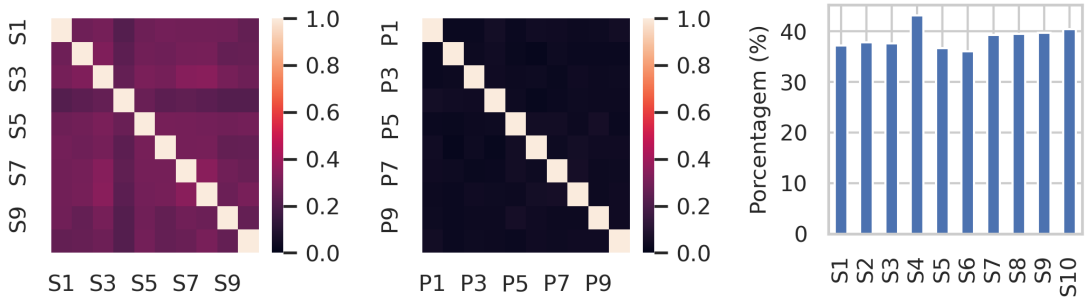


Figura 6. Distribuição de valores RSSI para diferentes distâncias.



(a) Sniffers. **(b) Pacotes.** **(c) Razão de zeros entre sniffers.**

Figura 7. Matrizes de correlação de pacotes perdidos entre sniffers (a), pacotes (b). Razão da distribuição de pacotes perdidos entre sniffers (c).

6. Análise dos Modelos de Classificação

A seção discute os resultados agregados dos dois modelos: o modelo LDPL e o modelo k-NN. Para ambos, o objetivo é encontrar quais os parâmetros ideais para classificação e determinar a acurácia de classificação de cada um dos modelos.

6.1. Modelo LDPL

Para a caracterização do η, pode-se observar que o z-score, que varia entre 0,5 e 3, faz o valor de η variar entre 1,59 e 1,73. Esses valores foram utilizados para determinar os melhores parâmetros para a classificação de distância, cujo resultado está na Tabela 2. Nessa tabela, foi separado o melhor resultado de acurácia dos resultados para valores fixos de N_s e N_p, que estão em negrito. O melhor resultado ficou para $N_s = 9$ e $N_p = 10$, o que era esperado, já que quanto maior a quantidade de dados melhor é a representatividade da distribuição e portanto mais acurada é a estimativa. Pode-se também comparar a vantagem entre usar sniffers redundantes contra usar múltiplos pacotes para o caso sem redundância. Nota-se um ganho da acurácia de 8% com o uso de vários sniffers para um pacote ao contrário de uma inesperada redução de 3% com o uso de vários pacotes para um sniffer.

Tabela 2. Resultados de acurácia em função dos parâmetros N_s e N_p. Os valores em negrito foram forçados para filtrar a tabela e determinar o melhor resultado nessas condições.

N_p	N_s	η (z)	**Acurácia** [%]	**NMSE**
10	9	1,65 (1,9)	79,5	16,9
1	10	1,61 (1,3)	68,3	205,1
10	**1**	1,58 (0,8)	57,3	304,6
1	**1**	1,60 (1,0)	60,0	442,8

Tabela 3. Acurácia em função da remoção dos dados com baixa precisão.

Fração removida [%]	**0**	**10**	**20**	**30**	**40**	**50**	**60**	**70**
Acurácia [%]	78,6	81,6	83,7	86,5	90,2	94,7	97,7	97,6

O resultado da remoção dos valores com desvio padrão alto pode ser visto na Tabela 3. Pode-se observar nessa tabela que a remoção dos dados provê um aumento linear na acurácia, e pode levar até a valores de 97%. Porém, esse aumento tem o custo de descartar mais da metade das medidas feitas. Esse processo pode aumentar significativamente a latência das aplicações e, com variações no ambiente, esse descarte pode penalizar as classes de forma diferente.

Para o modelo com os parâmetros ideais $N_p = 9$, $N_s = 10$ e $\eta = 1,654$, tem-se a matriz de confusão da Tabela 4. A tabela, normalizada ao longo das colunas, mostra a distribuição de classificação para cada uma das classes. É possível ver que mesmo com todos os ajustes e com aumentos significativos de acurácia, nem todas as classes foram igualmente beneficiadas. A melhor classe tem acerto de 97,1% enquanto a pior de apenas 16,6%. Essas características evidenciam que, mesmo com todos os ajustes, o modelo ainda é penalizado por não ser capaz de representar a alta complexidade do ambiente.

Tabela 4. Matriz de confusão para a classificação de distância com o modelo LDPL, normalizada nas colunas.

Resultados (%)						
	Distância Real (m)					
Prevista (m)	**1**	**10**	**20**	**30**	**40**	**50**
1	95,5	0	0	0	0	0
10	4,5	85,7	0	0	0	0
20	0	13,5	92,2	82,2	0	0
30	0	0,8	7,8	16,6	0,8	0
40	0	0	0	0,8	85,2	2,9
50	0	0	0	0,4	13,9	97,1

6.2. Modelo k-NN

Na Figura 8, apresenta-se os resultados do k-NN para diferentes seleções de N_s e N_p. Os resultados exibidos na figura representam as médias obtidas dentre todas as combinações possíveis para formar o super-sniffer, de modo que todas as combinações tenham o mesmo peso. Como esperado, os resultados revelam que aumentar o valor de N_s e N_p não necessariamente resulta em uma melhoria nas métricas do modelo, dada a existência do

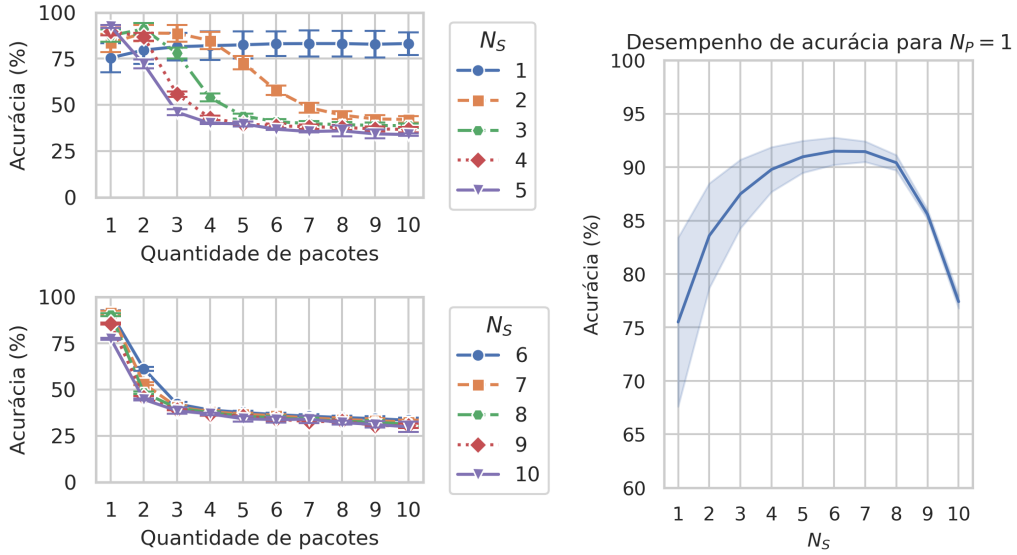


Figura 8. Acurácia do k-NN para diferentes tamanhos de super-sniffer.

problema de dimensionalidade e de ruído. Nessa perspectiva, o melhor modelo obtido não é aquele com $N_s = 10$, mas sim um intermediário, $N_s = 7$, que consegue capturar informações suficientes sem introduzir excesso de ruído. Essa característica é evidenciada ainda mais pelos resultados para $N_s = 10$, no qual o vetor cresce o suficiente para que mesmo o melhor modelo possua métricas longe de um grau satisfatório.

Paralelamente, aumentar N_p produz um efeito semelhante, mas que não é uniforme para todos os valores de N_s. Valores menores de N_s, dentro do intervalo de 1 a 4, conseguem produzir vetores suficientemente pequenos mesmo para N_p mais elevados, evitando o problema. No entanto, diminuir N_s também traz o risco de perder parte da representatividade do tráfego, além de apresentar instabilidade ao convergir em algumas combinações específicas de sniffers, como indicado pelo alto desvio padrão. Por exemplo, para as combinações com $N_s = 2$, e $N_p = 1$ o par de sniffers (S_2 e S_4) possui uma acurácia média de 92,97%, enquanto o par (S_1 e S_6) possui apenas 72,10%.

Observando as Tabelas 4 e 5, pode-se comparar os modelos em configurações que resultam nos seus melhores resultados. Nota-se que ambos os modelos enfrentam desafios com a perda de acurácia em distâncias que apresentam distribuições RSSI semelhantes, evidenciando a complexidade do ambiente de teste. No entanto, o modelo k-NN se destaca ao analisar todas as distâncias, apresentando uma proporção maior de acertos na maioria delas. Assim, o k-NN demonstra uma clara vantagem, fornece resultados mais consistentes em diversas categorias de distância, com a particularidade de exigir apenas a informação relativa a um único pacote. Essa característica é importante porque é capaz de reduzir drasticamente a latência na detecção de dispositivos, tanto devido ao número menor de pacotes, quanto da falta de necessidade de descartar amostras.

7. Conclusão e Trabalhos Futuros

Este trabalho propôs um novo método de determinação de distância radial baseado no uso de super-sniffers. Analisou-se como diferentes configurações de super-sniffer interferem nos parâmetros dos modelos, a fim de determinar como a redundância introduzida impacta

Tabela 5. Matriz de confusão média dos resultados do k-NN por categoria de distância com parâmetros $N_s = 7$ e $N_p = 1$, normalizada nas colunas.

Resultados (%)						
	Distância Real (m)					
Prevista (m)	**1**	**10**	**20**	**30**	**40**	**50**
1	98,96	0,46	0,09	0,08	0,08	0,34
10	0,04	94,75	1,88	1,05	1,02	1,26
20	0,04	0,84	90,17	3,10	4,55	1,29
30	0,07	0,40	2,83	88,16	4,72	3,82
40	0,07	0,54	4,91	6,30	85,08	3,11
50	0,08	0,44	0,49	2,53	1,32	95,13

os resultados. Observou-se que os modelos trazem resultados satisfatórios, especialmente o k-NN, que se prova melhor que o LDPL, com altas taxas de acurácia para diversos parâmetros de tamanho de super-sniffers inclusive para quantidade pequenas de pacotes por entrada, chegando a uma taxa de acurácia de 91,73%. Portanto, concluiu-se que super-sniffers se apresentam como um meio robusto de melhorar tanto a qualidade quanto a quantidade de medidas de RSSI, sendo capazes de melhorar a acurácia de métodos baseados em k-NN e LDPL em relação a sniffers individuais. Trabalhos futuros devem ser realizados de modo a determinar como o número de sniffers ideal varia em diferentes conjuntos de dados, além de outros cenários de coleta.

8. Agradecimentos

Este trabalho foi realizado com recursos do CNPq, FAPERJ e RNP. Além disso, o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e pelo projeto ANR Mitik (PRC AAPG2019).

Referências

Barai, S., Biswas, D., and Sau, B. (2017). Estimate distance measurement using nodemcu esp8266 based on rssi technique. In *2017 IEEE Conference on Antenna Measurements & Applications (CAMA)*, page 170–173.

Chuku, N. and Nasipuri, A. (2021). RSSI-based localization schemes for wireless sensor networks using outlier detection. *Journal of Sensor and Actuator Networks*, 10.

Dong, Q. and Dargie, W. (2012). Evaluation of the reliability of RSSI for indoor localization. In *2012 International Conference on Wireless Communications in Underground and Confined Areas*.

Freudiger, J. (2015). How talkative is your mobile device?: an experimental study of wi-fi probe requests. In *Proceedings of the 8th ACM Conference on Security and Privacy in Wireless and Mobile Networks*.

Gupta, V., Beyah, R., and Corbett, C. (2007). A Characterization of Wireless NIC Active Scanning Algorithms. In *2007 IEEE Wireless Communications and Networking Conference*, pages 2385–2390.

Heurtefeux, K. and Valois, F. (2012). Is RSSI a good choice for localization in wireless sensor network? In *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*.

Jaisinghani, D., Naik, V., Kaul, S. K., and Roy, S. (2017). Sniffer-based inference of the causes of active scanning in WiFi networks. In *2017 Twenty-third National Conference on Communications (NCC)*, pages 1–6, Chennai, India. IEEE.

Jose, A. A., Rishikesh, P. H., and Shaju, S. (2023). Mitigation of RSSI variations using frequency analysis and kalman filtering. In *Proceedings of the International Conference on Cognitive and Intelligent Computing*.

Li, G., Geng, E., Ye, Z., Xu, Y., Lin, J., and Pang, Y. (2018). Indoor positioning algorithm based on the improved RSSI distance model. *Sensors, vol. 18, no. 9, p. 2820, Aug. 2018*.

Mishra, A. K., Viana, A. C., and Achir, N. (2023). Do wifi probe-requests reveal your trajectory? In *2023 IEEE Wireless Communications and Networking Conference (WCNC)*.

R, V., Mittal, V., and Tammana, H. (2021). Indoor localization in BLE using mean and median filtered RSSI values. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*.

Rappaport, T. (2002). *Wireless Communications Principles and Practice. 2nd Edition*. Prentice Hall.

Saha, S., Chaudhuri, K., Sanghi, D., and Bhagwat, P. (2003). Location determination of a mobile device using IEEE 802.11b access point signals. In *2003 IEEE Wireless Communications and Networking, 2003. WCNC 2003.*

Syed, M. I., Fladenmuller, A., and Amorim, M. D. D. (2022a). RSSI: Lost and alone, a case for redundancy. In *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*.

Syed, M. I., Fladenmuller, A., and Dias de Amorim, M. (2022b). How much can sniffer redundancy improve Wi-Fi traffic? In *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, pages 1–5.

Syed, M. I., Flandenmuller, A., and Dias de Amorim, M. (2022c). PyPal: Wi-Fi Trace Synchronization and Merging Python Tool. Technical report, LIP6 UMR 7606, UPMC Sorbonne Université, France.

Verma, V. and Singh, A. (2019). Indoor location determination using radio signal strength model for distance estimation. In *2019 International Conference on Computer Communication and Informatics (ICCCI)*, page 1–4.

Wu, Z., Jedari, E., Muscedere, R., and Rashidzadeh, R. (2016). Improved particle filter based on WLAN RSSI fingerprinting and smart sensors for indoor localization. *Computer Communications Volume 83, June 2016, Pages 64-71*.

Yiu, S., Dashti, M., Claussen, H., and Perez-Cruz, F. (2017). Wireless RSSI fingerprinting localization. *Signal Processing Volume 131, February 2017, Pages 235-244*.