Reprodutibilidade de Experimentos em Redes de Computadores através do Catálogo de Dados RNP

Vitor Fontana Zanotelli[1], Nilson Luís Damasceno[2], Arthur Almeida Vianna[2], Gustavo Araujo[3], Michael Prieto Hernandez[3], Giovanni Comarela[1], Magnos Martinello[1], Antonio A. de A. Rocha[2]

[1]Departamento de Informática (DI) – Universidade Federal do Espírito Santo (UFES)

[2]Instituto de Computação (IC) – Universidade Federal Fluminense (UFF)

[3]Rede Nacional de Ensino e Pesquisa (RNP)

fz.vitor@gmail.com, nilson.uff.br@gmail.com, arthurvianna@id.uff.br

{gustavo.araujo, michael.hernandez}@rnp.br

{gc, magnos}@inf.ufes.br, arocha@ic.uff.br

Abstract. *The importance of experiment reproducibility is widely discussed in the scientific community. Proposals of both challenges and solutions are found in the literature. A common obstacle is related to data availability. The National Research and Education Network (RNP) collects and stores data related to its services, and to facilitate access to it, the Data Catalog project was created. This work presents the Catalog in two steps: i) first, by describing the project and its databases, and ii) then, a case of replicating work from the literature concerning the use of machine learning models for RTT prediction. For the replication, recurrent neural networks (RNNs, GRUs, and LSTMs) are used, achieving results close to the original ones.*

Resumo. *A importância da reprodução de experimentos é amplamente discutida na comunidade científica. São encontrados tanto desafios quanto propostas de soluções na literatura. Um entrave comum está relacionado à disponibilidade de dados. A RNP coleta e armazena dados relacionados aos seus serviços e para facilitar seu acesso, o projeto Catálogo de Dados foi criado. Esse trabalho apresenta o Catálogo em duas etapas: i) primeiro a partir da descrição do projeto e de suas bases de dados e, ii) em seguida, um caso de replicação de trabalho da literatura referente ao uso de modelos de aprendizado de máquina para predição de RTT. Para a replicação, são utilizadas redes neurais recorrentes (RNNs, GRUs e LSTMs), alcançando resultados próximos aos originais.*

1. Introdução

A reprodução de experimentos é essencial para garantir a confiança nos resultados apresentados em trabalhos na literatura e como forma de validação da pesquisa científica. Para um estudo ser reprodutível, as metodologias, ferramentas e dados devem ser compartilhados de forma clara e precisa à comunidade científica. Embora seja cada vez mais comum a exigência do compartilhamento do código e do conjunto de dados utilizados na

escrita e publicação de trabalhos científicos, esses nem sempre são disponibilizados adequadamente. São diversos os fatores que dificultam o processo, uma lista não exaustiva é: (1) documentação inadequada da metodologia, código e/ou dados; (2) restrições totais ou parciais ao código e dados, quando proprietários, sigilosos ou privativos; (3) falta de padronização e (4) falta de incentivos para justificar o trabalho extra dos autores no período de submissão.

A RNP[1] fornece serviços de conectividade em todo o território nacional, atendendo universidades, hospitais e institutos de pesquisa. Dessa forma, um grande volume de dados de monitoramento e desempenho é gerado e armazenado. Para promover o uso desses dados em pesquisas científicas e trazer transparência aos serviços prestados, a RNP iniciou o projeto Catálogo de Dados, tendo como um dos seus principais objetivos facilitar a identificação e a localização de dados de rede coletados sobre a infraestrutura do backbone nacional. O projeto deve atender, de forma ágil e eficiente, às demandas de pesquisadores relacionadas ao acesso dos dados de rede coletados e processados pela instituição.

Atualmente existe uma grande demanda de dados para a utilização de técnicas de aprendizado de máquina. A reprodução de experimentos que utilizam tais técnicas depende da disponibilidade de um conjunto de dados adequado, enfrentando também as dificuldades apresentadas anteriormente, podendo impossibilitar a sua realização. Este trabalho ilustra como trabalhos científicos que utilizam dados relacionados a monitoramento de redes de computadores podem ser replicados utilizando dados coletados a partir da infraestrutura de redes da RNP. É objetivo do trabalho apresentar à comunidade científica o projeto Catálogo de Dados e os tipos de dados que podem ser encontrados na RNP, associando os dados utilizados em trabalhos científicos aos dados disponíveis e exemplificando como realizar a replicação dos experimentos presentes na literatura. Dessa forma, espera-se contribuir para um ambiente mais favorável a reprodução de experimentos na comunidade científica nacional.

A organização desse trabalho é descrita a seguir. A Seção 2 apresenta os trabalhos relacionados na literatura. A Seção 3 apresenta o Projeto Catálogo de Dados. A Seção 4 apresenta os tipos de dados disponíveis e suas características no domínio de dados da RNP. A Seção 5 discute a reprodutibilidade de experimentos de aprendizado de máquina em redes e apresenta exemplos de casos de uso dos dados disponíveis. A Seção 6 ilustra um exemplo de replicação de experimento com dados RNP e, por fim, a Seção 7 discute os resultados obtidos e oportunidades para trabalhos futuros.

2. Trabalhos Relacionados

O objetivo desta seção é posicionar este trabalho na literatura, principalmente no que relaciona o uso de fontes de dados, experimentação e reprodução de trabalhos científicos em redes de computadores.

A definição dos critérios para a reprodução correta de um experimento pode variar conforme o contexto de utilização. Por exemplo, a Tabela 1 ilustra uma visão considerando as dimensões dos dados e da análise, conforme apresentado em [Arnold et al. 2019]. Nesse contexto, são definidos quatro possíveis casos a serem estudados, onde um experimento é: (a) Reprodutível quando se utilizam os mesmos dados

[1]Rede Nacional de Ensino e Pesquisa: `https://www.rnp.br/`

Tabela 1. Dimensões envolvidas na reprodução de um experimento [Arnold et al. 2019].

		Dados	
		Iguais	Distintos
Análise	Igual	Reprodutível	Replicável
	Distinta	Robusto	Generalizável

e análise chegando a mesma conclusão, (b) Robusto quando os mesmos dados e análises distintas chegam a conclusões iguais ou equivalentes, (c) Replicável quando são utilizados dados distintos para uma mesma análise e por fim, (d) Generalizável quando o resultado é robusto e replicável.

Já para a *Association for Computing Machinery* (ACM) os critérios apresentados em [ACM 2020] consideram a reprodução em três níveis: (a) Repetível quando o mesmo time de pesquisadores consegue repetir sob as mesmas condições o mesmo resultado, (b) Reprodutível quando um time distinto sob as mesmas condições consegue resultado equivalente ao original e (c) Replicável quando um time distinto de forma independente consegue alcançar o mesmo resultado do trabalho original. A visão difere um pouco da visão anterior, mas as preocupações são as mesmas. Em outros contextos possivelmente encontram-se definições diferentes destas; por isso, é necessário deixar claro quais os critérios para definição de reprodução quando o assunto é discutido.

Embora a presença dos dados, do código e da possibilidade de reprodução sejam desejados, e às vezes até exigidos em conferências e periódicos, estudos encontrados na literatura apontam uma realidade distante do ideal. Nas publicações da IEEE[2] em processamento de sinais a presença de *datasets* acompanhando artigos era baixa, presente em apenas um terço, a presença do código menor ainda, em cerca de um décimo [Vandewalle et al. 2009]. Já [Collberg and Proebsting 2016] mostra que considerando trabalhos publicados na ACM, é encontrada a replicabilidade fraca, onde é apenas verificado se o trabalho é acompanhado do código disponível e compilável, em também apenas aproximadamente um terço dos trabalhos verificados.

A comunidade científica apresenta uma preocupação em relação à reprodução e validação de experimentos e de trabalhos publicados. Na literatura encontram-se trabalhos publicados enumerando boas práticas para a área de computação em geral [Sandve et al. 2013] e promovendo a aprendizagem de estudantes por meio de competições [Canini and Crowcroft 2017]. Em [Scheitle et al. 2017] são levantadas as dificuldades enfrentadas tanto pelos autores quanto por aqueles responsáveis por reproduzir os trabalhos, sendo propostas formas de construir um ecossistema favorável para atender a necessidade de ambos. É também discutida a necessidade de competições e desafios como base para familiarizar a comunidade científica com os processos envolvidos e também a necessidade de promover incentivos aos autores e até o uso de *badges* para identificar trabalhos que passem certos critérios. O uso da ferramenta Popper é proposto em [David et al. 2019] para lidar com os problemas inerentes aos experimentos de redes de computadores em casos como: simulações, *testbeds* e medições em situações reais. Ao

[2]Institute of Electrical and Electronics Engineers: `https://www.ieee.org/`

se considerar a disponibilidade dos dados, outro desafio existente apontado pela literatura é a necessidade de anonimizar e ofuscar partes dos dados e metadados durante o processo de revisão, o que pode impactar diretamente no entendimento e no processo reprodução [Bajpai et al. 2017].

Iniciativas para a criação de repositórios de dados de rede não são uma ideia nova, sendo possível encontrar conjuntos bem documentados e disponíveis de medições [Shannon et al. 2005, Yeo et al. 2006] e também na área de segurança de redes [Garcia et al. 2014, Valeros and Garcia 2022]. Não é objetivo do Catálogo sanar todos os problemas apresentados ou implementar todas as propostas, mas sim somar aos esforços já existentes na comunidade científica brasileira na formação de um ambiente favorável a uma cultura forte de experimentação e reprodução na pesquisa nacional em redes de computadores. Os dados gerados no funcionamento da rede da RNP apresentam uma perspectiva nacional de uma rede universitária de grande escala, possuindo características singulares. Sendo um dos objetivos do Catálogo facilitar o acesso a esses dados.

Uma forma de utilização do Catálogo é permitir, quando consideramos a divisão em [Arnold et al. 2019], o uso de dados distintos para a replicação ou generalização de experimentos, ou quando consideramos as definições da [ACM 2020] na replicação destes. Mesmo quando os dados originais de um experimento estão disponíveis, as fontes de dados da RNP ainda se mostram úteis para proporcionar ao pesquisador esse acesso na dimensão dos dados, que quando distintos, contribuem para a validação de trabalhos já existentes. O Catálogo também apresenta a vantagem de apresentar dados que são abertos ao público. Assim, um trabalho baseado em dados sigilosos poderia ser publicamente validado utilizando, como substitutos, dados de fontes da RNP.

3. Catálogo de Dados da RNP

Ao prover serviços de conectividade em todo território nacional, a RNP coleta grandes volumes de dados de monitoramento e desempenho. O acervo resultante contém dados com potencial de serem utilizados em pesquisas científicas envolvendo a criação de modelos teóricos com finalidades diversas, como detecção de anomalias, predição de eventos, etc. Entretanto, a descrição tradicional de um acervo de dados por vezes não oferece informações suficientes aos pesquisadores para que alguns dos dados úteis disponíveis sejam identificados ou considerados como relevantes para suas pesquisas. Visando superar essa limitação, a RNP criou o projeto Catálogo de Dados.

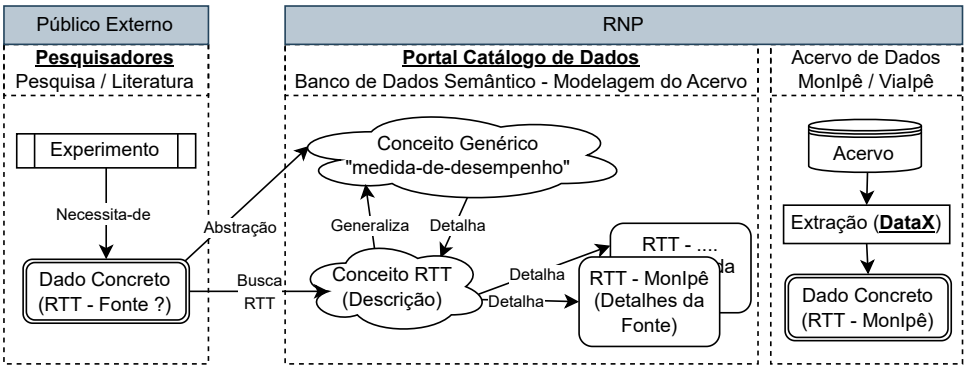

Figura 1. Diagrama Esquemático do Catálogo de Dados RNP.

O projeto Catálogo de Dados compreende dois sistemas computacionais: o "Portal Catálogo de Dados", no qual os pesquisadores podem explorar descrições detalhadas do acervo de dados disponibilizado pela RNP e selecionar conjuntos de dados relevantes desse acervo; e a ferramenta DataX, desenvolvida para facilitar a exportação dos conjuntos de dados selecionas para o pesquisador requerente. A Figura 1 ilustra o funcionamento do Catálogo de Dados RNP. Os pesquisadores podem acessar o Portal quando precisam de conjuntos de dados concretos (*datasets*) para realização de novos experimentos (Pesquisa) ou para reprodução de experimentos (Literatura). O Portal permite que o pesquisador busque o dado desejado diretamente, caso mais comum em reprodução de experimentos, ou realize uma exploração mais ampla dos dados disponíveis no acervo, possivelmente abstraindo os dados concretos desejados. Em qualquer caso, o Portal oferece mecanismos para o pesquisador navegar por termos que representam desde conceitos mais genéricos e abstratos de dados até chegar aos termos que descrevem detalhes das fontes de dados do acervo da RNP nas quais os dados concretos podem ser obtidos. Uma vez identificada a fonte de dados concretos, o pesquisador deve requisitar o *dataset* à RNP, que poderá usar os recursos do DataX para extraí-los do acervo.

O "Portal Catálogo de Dados" é uma aplicação Web que permite a exploração do acervo de dados disponibilizado pela RNP. Todas as informações exibidas pelo Portal, inclusive a configuração da sua própria interface, são armazenadas em um banco de dados semântico e dinamicamente atualizável, baseado no padrão "Resource Description Framework"(RDF) [Pan 2009]. O RDF representa informações como declarações (frases) de três partes (triplas) no formato "Sujeito Predicado Objeto", ou mais usualmente, na forma "v_1:Sujeito v_2:Predicado v_3:Objeto", onde v_i é um prefixo que identifica o vocabulário (domínio de conhecimento) referente ao nome que o sucede. O uso de vocabulários permite a reutilização de nomes sem criar ambiguidades. Por exemplo, a declaração "monipe:RTT rnpk:Concept rnp:RTT", utiliza o nome "RTT" duas vezes, mas com prefixos/vocabulários diferentes, o que permite que um mesmo nome tenha significados diferentes quando usado com vocabulários diferentes. Essa forma de representar informações auxilia na representação de conhecimento e na criação de taxonomias (hierarquias de classificação) e ontologias [Powers 2003]. Graças ao RDF, o banco de dados semântico flexível (sem um *schema* rígido) utilizado pelo Portal pode conter descrições detalhadas e extensíveis de termos interconectados através de predicados. As declarações nesses bancos de dados podem ser representadas como grafos direcionados nos quais, para cada declaração, o Sujeito é o vértice origem, o Predicado é a aresta direcionada e o Objeto, sempre que não for um valor literal (número, texto, etc.), é o vértice destino.

Em resumo, o banco de dados do Portal é capaz de descrever conjuntos de dados organizados de forma tradicional, como os *datasets* organizados como sequências de registros contendo uma ou mais colunas (campos) com valores atômicos, como tabelas de um modelo relacional com dados estatísticos ou como séries temporais com múltiplos valores numéricos. Além dessas organizações tradicionais, o banco de dados do Portal também é capaz de descrever dados organizados de forma mais complexa, como dados multi-valorados (*arrays*), hierarquizados (árvores) ou mesmo organizados como grafos. Ele também permite a declaração de regras para validação de valores dos dados e para validação das estruturas desses dados. Finalmente, o banco de dados também fornece mecanismos de suporte semântico, que permitem descrever o inter-relacionamento de qualquer dos dados descritos, seja qual for sua fonte, ao mesmo tempo que pode enri-

quecê-los com múltiplas descrições, potencialmente multi-linguísticas, apropriadas para os diferentes vocabulários dos usuários que consumem os dados.

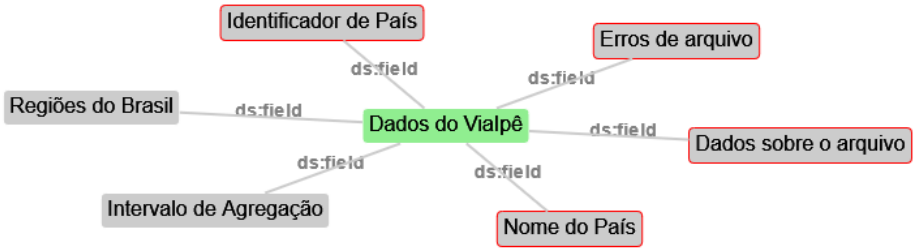

Figura 2. Exemplo de Grafo Interativo do Portal para o Sujeito "Dados do Viaipê".

O Portal utiliza uma interface baseada em grafos interativos criados a partir das declarações presentes no banco de dados. A Figura 2 exibe um grafo interativo, no qual os vértices são representados como retângulos e as arestas como linhas nomeadas que conectam esses retângulos. A cada interação do usuário, a interface de exploração apresenta o vértice selecionado (retângulo verde) no centro da área de visualização, descobre as declarações no banco de dados nas quais este vértice é Sujeito e exibe as ligações/predicados como linhas que conectam o retângulo verde (vértice-sujeito) aos retângulos cinza que representam os vértices-objeto. Quando um retângulo cinza sem borda vermelha (vértice-objeto selecionável) é clicado, ele se torna o novo vértice selecionado (retângulo verde) e uma nova interação se inicia. Detalhes adicionais sobre o vértice selecionado que não aparecem no grafo interativo são exibidas em áreas de visualização auxiliares.

A versão atual do Portal Catálogo de Dados contém o resultado da modelagem realizada a partir dos seguintes sistemas da RNP: (1) ViaIpê[3], que contém medições entre PoPs e as instituições usuárias locais (universidades, institutos federais e centros de pesquisa) e (2) MonIpê [4], um serviço que realiza o monitoramento dos links entre os PoPs da RNP. O processo de modelagem seguiu a estratégia *bottom-up*, iniciando na descrição dos dados concretos presentes nas fontes de dados dos sistemas escolhidos e acrescentando os conceitos gerais identificados ao longo da análise dos dados, além de termos e conceitos adicionais que pudessem aprimorar o entendimento das descrições essenciais. O Banco de Dados do Portal, baseado no padrão RDF, não só se mostrou adequado para registrar as informações referentes aos dados coletados pela RNP como também mostra-se capaz de interligar esses dados com informações referentes à governança de dados, modelagem de processos, administração de papéis, organização administrativa, etc.

Uma vez que o pesquisador tenha explorado o acervo de dados da RNP através do Portal Catálogo de Dados, pode-se então solicitar os *datasets* selecionados para que sejam enviados para um repositório local. O sistema DataX atua como um facilitador na exportação dos conjuntos de dados do acervo RNP para o repositório do pesquisador solicitante. O sistema foi implementado através do framework de código aberto *Spring Cloud Data Flow* que implementa formas de operar e programar fluxos de dados e *pipelines* a partir da integração de diversas fontes de dados de rede, e a aplicação de políticas

[3]https://viaipe.rnp.br/
[4]https://www.rnp.br/servicos/experimentos-avancados/eciencia/monipe

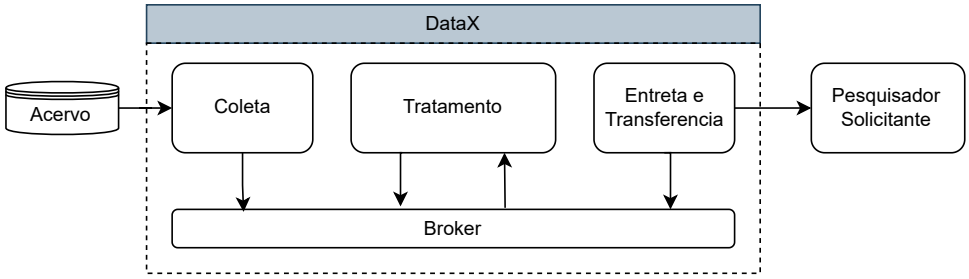

Figura 3. Diagrama Esquemático do DataX.

e representação de processos. A Figura 3 ilustra como é realizada a exportação dos *datasets* através do DataX. Primeiramente, é informado para o sistema qual subconjunto de dados deve ser exportado e o repositório para onde os dados devem ser enviados. Em seguida, os *datasets* são coletados do acervo RNP e passam por um barramento (*broker*) responsável por trafegar e implementar um sequenciamento de tratamentos, conforme o nível de acesso do usuário e sensibilidade de cada informação. Esses tratamentos podem ser, sanitização dos dados, remoção de campos sensíveis, transformação de estrutura ou/e qualquer política corporativa da RNP necessária para o compartilhamento. Por fim, o dado é entregue e armazenado no repositório informado pelo pesquisador solicitante.

4. Fontes de Dados do Catálogo

Esta seção apresenta a Rede Ipê e um subconjunto dos dados coletados e armazenados pela RNP. Os dados catalogados pelo projeto são detalhados, evidenciando suas características e especificidades.

A RNP provê o serviço "Rede Ipê"[5] para garantir conectividade em todo território nacional. A Rede Ipê estende-se por todo o Brasil, com um Ponto de Presença (PoP) em cada um dos 26 estados e no Distrito Federal. Cada PoP é responsável pelo serviço de conectividade em sua região, focando primariamente em atender instituições acadêmicas como universidades mas também atendendo centros de pesquisa e hospitais. A topologia nacional da rede possui conexões entre os PoPs nacionais e também conexões internacionais com redes externas à RNP. Internacionalmente, a Rede Ipê conecta-se a outros países por meio da RedCLARA (América Latina e Europa), Monet (Estados Unidos) e SACS-WACS (África).

A RNP também oferece formas de realizar o monitoramento da Rede Ipê por meio de duas ferramentas: (1) Via Ipê, um aplicativo Web que permite que os usuários visualizem, em tempo real, estatísticas e métricas referentes a qualidade do serviço e desempenho dos PoPs para as instituições usuárias locais, e (2) MonIpê que monitora os parâmetros de rede entre os PoPs. Embora seja possível aos usuários verificar o estado da conexão de suas instituições ou do serviço em tempo real, dados históricos não se encontram disponíveis na ferramenta, sendo necessário que o pesquisador realize a coleta caso seja necessário. O resultado dessa coleta contínua forma um grande acervo de dados, recorrentemente explorado e analisado por times de analistas da própria RNP. Dessa forma, os dados de monitoramento podem ser divididos em duas fontes de dados sob óticas distintas.

[5]Rede Ipê: `https://www.rnp.br/sistema-rnp/rede-ipe`

Tabela 2. Descrição das variáveis presentes conjunto de dados das medições entre PoPs e instituições usuárias de seus serviços.

Variável	Descrição
Tempo	Data e hora da medição.
Estado	Sigla representando o estado onde a medição ocorreu.
Instituição	Local usuário do serviço do PoP a que a medição se refere.
Interface	Identificador da interface de rede do local.
Client Side	Indica se a medição foi realizada do lado do *client* ou PoP.
Packet Loss	Valor de perda de pacotes: *min, avg e max* em percentual.
RTT	*Round Trip Time*: *min, avg e max* em milissegundos (ms).
Download	Taxa de download: *min, avg e max* em bits por segundo (bps).
Upload	Taxa de upload: *min, avg e max* em bits por segundo (bps).

Tabela 3. Descrição dos conjuntos de dados coletados, considerando as medições entre PoPs.

Conjunto	Descrição
throughput	Quantidade observada de dados enviados (bps).
packet-retransmits	Número de pacotes retransmitidos em uma transferência TCP.
histogram-rtt	Histograma de tempos de RTT (ms).
histogram-ttl-reverse	Histograma reverso de tempos de RTT (ms).
histogram-owdelay	Histograma de atrasos unilaterais (ms).
histogram-ttl	Histograma do número de saltos durante a transferência de pacotes.
packet-trace	Trajeto do pacote retornado por *traceroute* ou *tracepath*.

Medições entre PoPs e instituições locais

Essa fonte de dados é relativa às medições de rede entre um PoP e instituições usuárias de seus serviços em sua região (e.g. universidades e centros de pesquisa). Como discutida anteriormente, possui uma interface visual denominada Via Ipê. Os dados disponíveis e uma breve descrição se encontram na Tabela 2. O conjunto de dados é composto por dezessete variáveis: cinco delas individualizam uma observação (tempo, unidade federativa, local, interface e *client side*) e doze são variáveis numéricas que quantificam a observação, apresentando o valor médio, mínimo e máximo em um intervalo de tempo para as seguintes métricas de rede: *packet loss*, rtt, download e upload.

Medições entre PoPs

Esse conjunto de dados apresenta a medições de métricas de rede entre um par de PoPs (e.g. PoP Espírito Santo e PoP Minas Gerais). Os dados são coletados através da ferramenta MonIPÊ[6], uma plataforma de monitoramento da qualidade e desempenho da rede, em conjunto com as ferramentas do perfSONAR[7], coleção de software *open source* para realização de medições em rede. Os dados disponíveis e sua descrição se encontram na Tabela 3, são utilizadas ferramentas distintas para medir as métricas de rede que ocorrem em intervalos pré-definidos no tempo. Nos casos dos histogramas são retornados os valores *min, avg,* e *max* da sua respectiva variável e no *packet-trace* os valores *min, avg,* e *max* do RTT e do TTL.

[6]MonIPÊ: `https://www.rnp.br/servicos/experimentos-avancados/eciencia/monipe`

[7]perfSONAR: `https://www.perfsonar.net/`

5. Reprodução de Experimentos em Redes de Computadores

Esta seção apresenta um exemplo de aplicação de algoritmo de aprendizado de máquina para solução de um problema de previsão de métricas de rede. Foram selecionados artigos da literatura relacionados ao problema de previsão de latência, que utilizam conjuntos de dados semelhantes aos dados armazenados pela RNP.

5.1. Predição de RTT com técnicas de machine learning

Um dos problemas estudados na literatura de redes é a previsão de métricas de desempenho. Entre as métricas usuais, um dos casos é referente à predição do *Round-Trip Time* (RTT), o tempo de "ida e volta" de uma mensagem. O valor do RTT é uma forma de medir a latência, ou atraso, da rede. É uma métrica importante para a qualidade de serviço, e a ocorrência de valores fora dos usuais (e.g. aumentos repentinos) pode ter impacto negativo em diversos tipos de aplicação, como em jogos e chamadas de vídeo. Os trabalhos descritos a seguir tratam do problema em contextos distintos.

Em [Dong et al. 2019], o problema de predição de valores de RTT ao longo do tempo é tratado através do uso de redes neurais recorrentes, mais especificamente através da utilização de redes compostas por *Minimal Gated Units* (MGUs), uma variante das redes formadas por *Gated Recurrent Units* (GRUs). O conjunto de dados foi obtido através da medição de RTT entre dois nós de rede distantes (da ordem de centenas de quilômetros) durante um mês inteiro, possuindo aproximadamente duzentas mil observações. Os resultados do modelo proposto são comparados com outros modelos de redes recorrentes clássicas como RNN (*Recurrent Neural Networks*), GRU e *long short-term memory* (LSTM). A comparação utiliza as métricas de erro médio absoluto (MAE), erro quadrático médio (RMSE) e o coeficiente de determinação (R^2) e o melhor resultado em todos os casos é do modelo proposto utilizando MGUs.

Já o trabalho apresentado em [Li and Zhang 2021] trata do mesmo problema utilizando redes neurais recorrentes do tipo *Transformers* [Vaswani et al. 2017]. A entrada do modelo é uma série temporal de valores, em intervalos variáveis de tempo e o modelo prevê os próximos valores numa janela de tamanho pré-definido. O *dataset* utilizado foi construído pelos próprios autores utilizando o *Wireshark* para capturar e analisar pacotes TCP, registrando o instante da medição e o valor de RTT, totalizando aproximadamente 6000 amostras. As métricas de avaliação utilizadas são o MAE e a perplexidade (PPL)[8]. O modelo proposto é comparado com um modelo utilizando redes LSTM e com o uso de uma fórmula empírica[9] para predição, obtendo resultados superiores em relação às métricas utilizadas.

Em [Wassermann et al. 2017] é proposto o NETPerfTrace, um sistema para rastreio de rota capaz de predizer mudança de caminhos e alterações na latência dos mesmos. O sistema utiliza árvores de decisão para prever o tempo restante de vida de um caminho, o número de mudanças de caminho num determinado período e a latência de uma rota. As *features* utilizadas são estatísticas derivadas de distribuições empíricas (e.g. média, mínimo, máximo e percentis) relacionadas a três variáveis relacionadas a rotas: tempo

[8] A perplexidade pode ser definida como a exponencial da perda: $ppl = exp(loss)$.

[9] A fórmula empírica utilizada é: $SRTT_i = \alpha RTT_{i-1} + (1 - \alpha)SRTT_{i-1}$. Onde $SRTT_i$ é o RTT estimado no tempo i, RTT_j o valor real de RTT no instante j e α uma constante calculada empiricamente, mas com o valor usual de $\frac{1}{8}$.

de vida, mudanças e latência (RTT). O conjunto de dados é composto de uma semana de medições obtidas pelo *traceroute* da cidade de Paris, realizadas pelo M-Lab[10]. O sistema é comparado com o DTRACK [Cunha et al. 2014], apresentando um desempenho superior com um percentual de erros de predição menor para a maioria dos casos apresentados.

5.2. Formalização do problema

Os três trabalhos apresentados realizam a predição de valores de RTT, os dois primeiros num contexto mais próximo e limitado apenas a predição de valores futuros e o último como parte de uma proposta de solução mais ampla. Os dados utilizados foram gerados pelos pesquisadores nos dois primeiros casos e uma base de dados já existente no terceiro caso. Nos três casos, o tipo de dado utilizado é conceitualmente o mesmo: séries temporais de valores de RTT.

A tarefa de predição de valores futuros de RTT realizada nos três trabalhos pode ser generalizada como um problema de aprendizado de máquinas de regressão. Dado um conjunto $\mathcal{D} = \{(\mathbf{x_1}, \mathbf{y_1}), ..., (\mathbf{x_n}, \mathbf{y_n})\}$ onde $\mathbf{x_t} = [r_{t-w+1}, ..., r_{t-1}, r_t]^T$ é um vetor representando uma série temporal de valores de RTT r em uma janela de w medições e $\mathbf{y_t} = [r_{t+1}, ..., r_{t+k}]^T$ um vetor das k medições seguintes, a tarefa a ser realizada é treinar um modelo $\mathcal{M}(\cdot)$ utilizando um subconjunto de $\mathcal{D}$ chamado de treino, de forma que para os dados não utilizados no treino, $\mathcal{M}(\mathbf{x_i}) \approx \mathbf{y_i}$ sob alguma métrica de desempenho. Para construir um conjunto como $\mathcal{D}$, basta acesso a um conjunto $\mathcal{U} = \{r_1, r_2, ..., r_n\}$ com medições sequenciais de RTT, então $\mathcal{D}$ pode ser criado por uma janela deslizante de tamanho $w + k$. Como evidenciado pelos artigos, diferentes fontes de dados podem ser utilizadas para construir $\mathcal{U}$, e consequentemente $\mathcal{D}$. Considerando o caso de utilização de dados disponíveis na RNP, para encontrar os conjuntos adequados para a realização do experimento é preciso buscar por fontes que contém o conceito, ou tipo de dado, RTT, atrelado ao tempo em que ocorreu a medição. Facilitar o processo de encontrar as fontes de dados mais adequadas no domínio de dados da RNP através dos conceitos utilizados no experimento é a proposta do projeto Catálogo de Dados, detalhado na Seção 3.

6. Caso concreto de uso

Nesta seção é realizada a replicação de um dos artigos descritos anteriormente para exemplificar um caso concreto de uso dos dados disponíveis e do Catálogo. O artigo selecionado, [Dong et al. 2019], descreve o algoritmo proposto e a forma de coleta dos dados, mas o código e o conjunto de dados não é disponibilizado, um problema comum como já discutido anteriormente. Nesse cenário, a extensão e variabilidade presente nas fontes de dados da RNP permite tanto a seleção de um subconjunto de dados com perfil próximo ao utilizado pelo trabalho que se planeja reproduzir, quanto um perfil distinto para testar a replicabilidade e generalização das soluções propostas.

6.1. Seleção dos dados

Os dados gerados e armazenados pela RNP são provenientes de uma rede extensa e diversa, como já discutido anteriormente. Dessa forma, após a escolha do tipo de dado a ser utilizado (nesse caso o RTT) é possível selecionar diferentes subconjuntos de dados de latência. O trabalho citado utiliza um mês de medições entre dois nós de uma rede

[10] *M-Lab open Internet measurement initiative*: `https://www.measurementlab.net/`.

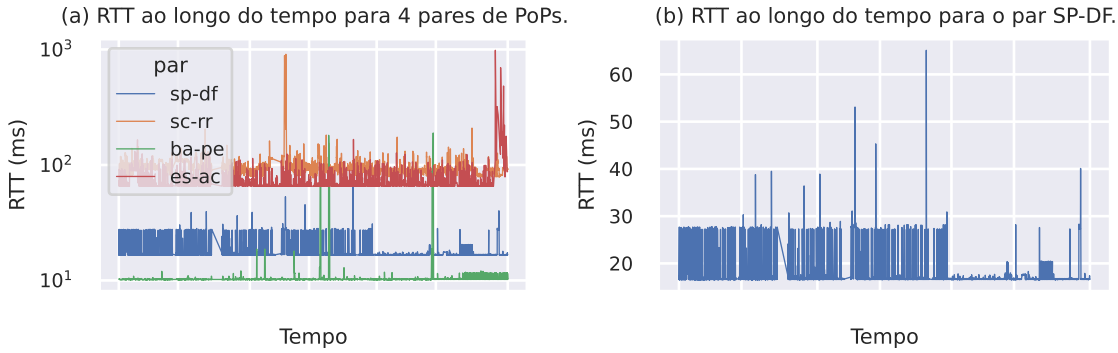

Figura 4. RTT ao longo do período de outubro a dezembro de 2021 para: (a) Pares de PoPs e (b) Par São Paulo - Distrito Federal.

com uma distância da ordem de milhares de quilômetros. Estatísticas sobre os dados não são fornecidas, sendo possível apenas aferir pela descrição e imagens disponíveis certas propriedades como: ordem de grandeza de valores, periodicidade e uma noção da variabilidade.

É esperado que após a utilização do Catálogo por um pesquisador, seja possível estabelecer uma relação entre as *features* utilizadas em um trabalho da literatura e as fontes de dados da RNP atualmente modeladas. Considerando as bases já catalogadas, tanto o Via Ipê quanto o MonIPÊ são fontes viáveis para a replicação do trabalho em questão.

Para fins desta seção foram utilizados os dados do MonIPÊ, descritos na Seção 4, de medições entre PoPs. Pelo fato de cada PoP estar presente em uma unidade federativa distinta (com exceção da Paraíba no caso de João Pessoa e Campina Grande), é possível selecionar pares de PoPs com níveis de distâncias distintas, permitindo a escolha de perfis de latência variados. A Figura 4(a) mostra os dados referentes a quatro pares de PoPs no período de outubro até o final de dezembro de 2021: (1) São Paulo e Distrito Federal, (2) Santa Catarina e Roraima, (3) Bahia e Pernambuco e (4) Espírito Santo e Acre. Já a Figura 4(b) destaca o RTT do par São Paulo e Distrito Federal. Nota-se que as características dos dados mudam ao longo do tempo, portanto, outra escolha importante é da janela de tempo a ser utilizada. Além de diferirem entre si, um mesmo par de PoPs pode, ao longo do tempo, apresentar mudança em suas características, como pode ser visto tanto na Figura 4 onde o valor da latência é ilustrado ao longo do tempo, e na Figura 5 onde a distribuição de probabilidade dos valores é apresentada para cada mês, mostrando a diferença entre o mês de dezembro e os demais meses. Para a replicação foram utilizados os três meses referentes a latência entre os PoPs São Paulo e Distrito Federal.

6.2. Modelagem e Resultados

O trabalho selecionado [Dong et al. 2019] propõe o uso das redes do tipo MGUs e realiza comparações com redes tradicionais como RNNs, LSTMs e GRUs. O código utilizado para as MGUs não se encontra disponibilizado, limitando a análise aos modelos presentes em bibliotecas padrões de aprendizado de máquina. O trabalho também não fornece todos os valores de hiperparâmetros, valores coerentes com experimentos passados e com a literatura foram utilizados quando necessários. Para a replicação é utilizada a linguagem

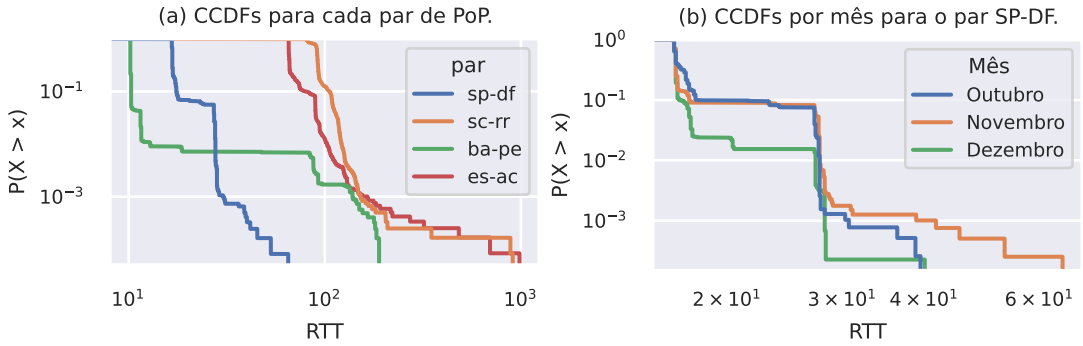

Figura 5. CCDFs do RTT durante o período de outubro a dezembro de 2021 para: (a) Pares de PoPs e (b) Par São Paulo - Distrito Federal.

Tabela 4. Comparação entre os resultados obtidos após reprodução do experimento e resultados originais.

	Replicação			Artigo		
Modelo	RMSE	MAE	R^2	RMSE	MAE	R^2
RNN	1.13 (0.55)	0.68 (0.67)	0.73 (0.34)	2.953	2.257	0.892
GRU	1.46 (0.84)	0.81 (0.58)	0.52 (0.48)	2.162	1.685	0.958
LSTM	0.83 (0.06)	0.40 (0.16)	0.88 (0.02)	2.538	1.927	0.927

de programação Python e as bibliotecas NumPy, Pandas, Scikit-Learn e Tensorflow[11].

Para a replicação foram utilizados três modelos recorrentes: RNN, GRU e LSTM. Os modelos criados possuem duas camadas recorrentes, seguidas de uma camada de *Dropout* e por fim uma densa. O otimizador e função de perda são o Adaptive Moment Estimation (Adam) e raiz quadrática média, respectivamente. Para a predição foram utilizados 20 valores consecutivos de RTT para prever o valor seguinte.

A Tabela 4 apresenta os resultados encontrados após a replicação, com a média e desvio padrão após dez execuções, e os resultados originais do trabalho escolhido. É importante observar que as métricas de RMSE e MAE podem ser comparadas entre os modelos de um mesmo conjunto de teste, mas não faz sentido comparar entre testes distintos, visto que a escala dos dados pode influenciar no valor. O valor de R^2 indica o ajuste do modelo aos dados, valores próximos de um indicam um bom ajuste. Essas métricas foram escolhidas por serem as utilizadas no trabalho a ser reproduzido. Em se tratando de séries temporais, uma análise mais aprofundada dos resíduos seria preferível para verificar o ajuste, mas encontra-se fora do escopo e objetivos deste trabalho.

No artigo, o modelo GRU obteve o melhor resultado, já na replicação o melhor resultado é a LSTM. O valor de R^2 indica que os modelos do trabalho se ajustaram melhor aos dados do que os obtidos na replicação do experimento, mas é apresentado apenas o resultado de uma execução. Na replicação, valores próximos são alcançados e é possível observar que apenas a LSTM é consistente em seus resultados, os outros apresentam valores altos de desvio padrão. As razões da diferença podem incluir as características distintas das fontes de dados. É possível verificar que o conjunto de dados utilizado

[11]O código e os dados utilizados estão disponibilizados em `https://github.com/VitorSpa/trab-catdados`

apresentou mudança em suas características ao longo do tempo, com o mês de dezembro distinto dos demais. Embora o mesmo resultado não tenha sido alcançado, os resultados são próximos. Dessa forma, o Catálogo e os dados RNP contribuem para contornar a ausência de código e disponibilização dos dados no artigo citado.

7. Conclusão

A reprodução correta de experimentos é necessária para atestar os resultados presentes na literatura científica. Discussões sobre os desafios envolvidos e possíveis soluções são recorrentes, no entanto, a plena reprodução de trabalhos científicos continua sendo um desafio. Parte do problema está relacionado a disponibilidade de dados bem documentados.

Este trabalho apresenta a reprodução de um experimento utilizando o Projeto Catálogo de Dados da RNP. O Projeto contempla um Portal que permite a exploração do acervo de dados coletados, armazenados e catalogados pela RNP. O Portal utiliza um banco de dados flexível, baseado em conceitos de Web Semântica, que possibilita descrições detalhadas dos dados e das relações entre eles e comunidades de usuários com demandas distintas.

O trabalho realiza a reprodução de um trabalho acadêmico no qual os dados originais não disponibilizados pelos autores, sendo substituídos por um conjunto de dados da RNP, localizados utilizando o Catálogo. Os resultados obtidos são próximos aos originais e *insights* obtidos ao longo do processo são apresentados.

Como oportunidades de trabalhos futuros e reconhecendo que Catálogo de Dados é um projeto em andamento, pode-se considerar o acréscimo de novas fontes de dados, a inclusão de aperfeiçoadas, a identificação mais precisa das necessidades da comunidade científica ao acessar o Portal e a detecção de inter-relações entre dados e pesquisadores. Espera-se, através deste trabalho, facilitar e promover a utilização dos dados presentes na RNP para a produção científica.

Referências

ACM (2020). Artifact review and badging version 1.1. `https://www.acm.org/publications/policies/artifact-review-and-badging-current`. Acessado: 03.01.2024.

Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., Whitaker, K., et al. (2019). The turing way: a handbook for reproducible data science. *Zenodo*.

Bajpai, V., Kühlewind, M., Ott, J., Schönwälder, J., Sperotto, A., and Trammell, B. (2017). Challenges with reproducibility. In *Proceedings of the Reproducibility Workshop*, Reproducibility '17, page 1–4, New York, NY, USA. Association for Computing Machinery.

Canini, M. and Crowcroft, J. (2017). Learning reproducibility with a yearly networking contest. In *Proceedings of the Reproducibility Workshop*, Reproducibility '17, page 9–13, New York, NY, USA. Association for Computing Machinery.

Collberg, C. and Proebsting, T. A. (2016). Repeatability in computer systems research. *Commun. ACM*, 59(3):62–69.

Cunha, I., Teixeira, R., Veitch, D., and Diot, C. (2014). Dtrack: A system to predict and track internet path changes. *IEEE/ACM Transactions on Networking*, 22(4):1025–1038.

David, A., Souppe, M., Jimenez, I., Obraczka, K., Mansfield, S., Veenstra, K., and Maltzahn, C. (2019). Reproducible computer network experiments: A case study using popper. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS '19, page 29–34, New York, NY, USA. Association for Computing Machinery.

Dong, A., Du, Z., and Yan, Z. (2019). Round Trip Time Prediction Using Recurrent Neural Networks With Minimal Gated Unit. *IEEE Communications Letters*, 23(4):584–587. Conference Name: IEEE Communications Letters.

Garcia, S., Grill, M., Stiborek, J., and Zunino, A. (2014). An empirical comparison of botnet detection methods. *computers & security*, 45:100–123.

Li, R. and Zhang, X. (2021). All You Need is Transformer: RTT Prediction for TCP based on Deep Learning Approach. In *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, pages 348–351.

Pan, J. Z. (2009). Resource description framework. In *Handbook on ontologies*, pages 71–90. Springer.

Powers, S. (2003). *Practical RDF: solving problems with the resource description framework*. "O'Reilly Media, Inc.".

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS computational biology*, 9(10):e1003285.

Scheitle, Q., Wählisch, M., Gasser, O., Schmidt, T. C., and Carle, G. (2017). Towards an ecosystem for reproducible research in computer networking. In *Proceedings of the Reproducibility Workshop*, Reproducibility '17, page 5–8, New York, NY, USA. Association for Computing Machinery.

Shannon, C., Moore, D., Keys, K., Fomenkov, M., Huffaker, B., and claffy, k. (2005). The internet measurement data catalog. *SIGCOMM Comput. Commun. Rev.*, 35(5):97–100.

Valeros, V. and Garcia, S. (2022). Hornet 40: Network dataset of geographically placed honeypots. *Data in Brief*, 40:107795.

Vandewalle, P., Kovacevic, J., and Vetterli, M. (2009). Reproducible research in signal processing. *IEEE Signal Processing Magazine*, 26(3):37–47.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wassermann, S., Casas, P., Cuvelier, T., and Donnet, B. (2017). NETPerfTrace: Predicting Internet Path Dynamics and Performance with Machine Learning. In *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*, Big-DAMA '17, pages 31–36, New York, NY, USA. Association for Computing Machinery.

Yeo, J., Kotz, D., and Henderson, T. (2006). Crawdad: A community resource for archiving wireless data at dartmouth. *SIGCOMM Comput. Commun. Rev.*, 36(2):21–22.