FERNANDO FRANCISCO DRUSZCZ

DISPARIDADE DA PRODUÇÃO CIENTÍFICA ENTRE SUB-ÁREAS DA CIÊNCIA DA COMPUTAÇÃO BRASILEIRA

(*versão pré-defesa, compilada em 23 de fevereiro de 2023*)

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: André Luís Vignatti.

CURITIBA PR

2023

RESUMO

Este trabalho apresenta uma análise e comparação da qualidade de trabalhos científicos na área da ciência da computação. Em meio a um crescimento exponencial da quantidade de publicações, essa análise tem se tornado cada vez mais relevante. No decorrer do trabalho são introduzidas alternativas de como são feitas as análises atualmente. Também, são apresentados os problemas que surgem ao fazer uso das alternativas introduzidas. Ainda, é exposta uma proposta encontrada na literatura que soluciona os problemas encontrados. Ao final do trabalho é apresentada uma análise feita em cima da produção científica da ciência da computação brasileira. Nessa análise é mostrado como se comportam as citações das diferentes áreas da ciência da computação. Além disso, é aplicada uma solução com o objetivo de normalizar os dados entre as sub-áreas da ciência da computação brasileira, o que pode tornar a comparação da produção científica mais justa.

Palavras-chave: Análise de citações. Ciência da Computação Brasileira. Normalização.

ABSTRACT

This work presents an analysis and comparison of the quality between works within the field of Computer Science. Amid an exponential growth in the number of publications, this analysis has become increasingly relevant. In the course of this work alternatives of how the analysis in currently done will be introduced. Also, the problems that arise using these alternatives will be presented. Furthermore, a proposal found in the literature to solve these problems will be exhibited. At the end of this work an analysis upon the scientific production of the brazilian computer science community will be presented. In this analysis it is shown how the number of citations between different sub-fields of computer science behaves. Also, a solution with the goal of normalizing the data between the sub-fields of the brazilian computer science community is applied, which can make the comparison of the scientific production fairer.

Keywords: Citation Analysis. Brazilian Computer Science. Normalization

SUMÁRIO

1 INTRODUÇÃO 5
2 CONCEITOS PRELIMINARES 6
2.1 DEFININDO A IMPORTÂNCIA DE ARTIGOS E PESQUISADORES. . . . 6
2.1.1 Métricas 6
2.1.2 Modelos da Produção Científica 7
2.1.3 Formas Alternativas 8
2.2 DISPARIDADE ENTRE ÁREAS 9
2.2.1 Distribuição Lei de Potência 9
2.2.2 Analisando Diferentes Áreas 10
2.3 OBSERVAÇÕES FINAIS DO CAPÍTULO 12
3 METODOLOGIA 13
3.1 OS DADOS E FONTES 13
3.2 TRABALHANDO OS DADOS 14
3.3 OBSERVAÇÕES FINAIS DO CAPÍTULO 15
4 RESULTADO E ANÁLISE 17
4.1 DISPARIDADE ENTRE AS ÁREAS DA CIÊNCIA DA COMPUTAÇÃO BRASILEIRA. 18
5 CONCLUSÃO 21
REFERÊNCIAS 22

1 INTRODUÇÃO

Este trabalho tem como objetivo analisar as diferentes formas de avaliar trabalhos e produções científicas. Mais especificamente, é analisada como a quantidade de citações recebidas por um trabalho – uma métrica bibliográfica que se tornou referência na avaliação de trabalhos científicos – é usada. Ainda, iremos expor os problemas que surgem ao usar essa métrica, para então propor uma solução que contorne os problemas apresentados.

Será visto que os problemas relacionados ao uso da quantidade de citações encontrados na literatura são observados em grandes áreas da ciência, como a biologia, física, as engenharias, dentre outras (Noorden et al., 2014). Porém, comparar grandes áreas da ciência pode não ser tão relevante, visto que existem comitês específicos para cada uma dessas grandes áreas. O desafio ocorre ao comparar diferentes subáreas de uma mesma grande área, visto que frequentemente elas são submetidas a uma métrica comum da grande área, o que pode causar comparações equivocadas. Após uma análise experimental feita sobre os dados de publicações, será visto como esses problemas se comportam em subáreas de uma área específica da ciência, onde frequentemente trabalhos de sub-áreas diferentes são comparados: as sub-áreas da ciência da computação brasileira.

No Capítulo 2 será feita uma introdução ao tema, onde serão discutidas as alternativas para se fazer essa análise e quais os seus problemas. Ainda no Capítulo 2 será vista uma proposta para solucionar os problemas encontrados. No Capítulo 3 serão apresentados os dados que serão utilizados no experimento que será analisado no trabalho, e como esses dados foram trabalhados. Finalmente, no Capítulo 4 serão apresentados os resultados do experimento realizado, demonstrando como funciona a solução proposta.

2 CONCEITOS PRELIMINARES

Nesse capítulo serão vistos alguns conceitos necessários para o entendimento do trabalho. Aqui serão apresentadas as métricas e modelos mais comumente usadas e aceitas pela comunidade para definir a relevância de artigos científicos e seus autores. Paralelamente, será visto que essas ferramentas têm problemas que podem levar a conclusões erradas.

2.1 DEFININDO A IMPORTÂNCIA DE ARTIGOS E PESQUISADORES

O crescimento da produção científica tem sido exponencial nas últimas décadas (Wang e Barabási, 2021). Devido a isso, cresce a necessidade de diferenciar a qualidade desses trabalhos, visto que é cada vez mais difícil analisá-los individualmente. Esse tipo de análise é importante para a comunidade científica definir critérios de comparação. Em especial, ela é valiosa para diferenciar quais são os veículos, pesquisadores, instituições, etc. mais interessantes e importantes. Como será visto neste capítulo, essa é uma tarefa difícil, pois, basicamente, há apenas dados quantitativos sobre a produção científica, e a partir deles é que devem ser obtidas conclusões qualitativas. Por isso surgem divergências, como as que serão apresentadas aqui, na interpretação dos dados, que devem ser discutidas para melhorar as conclusões dessa análise.

Algo que tem sido considerado crucial para fazer essa análise é a informatização dos dados referentes a publicações científicas. Esse recente processo facilita que pesquisadores processem e tirem conclusões de dados de bases como a *Web of Science*, que possui mais de 155 milhões de publicações indexadas (of Science, 2023). Com base nesses dados é possível calcular métricas comumente usadas para tirar conclusões sobre a qualidade de artigos, pesquisadores e da produção científica de maneira geral.

2.1.1 Métricas

Nessa seção serão vistas métricas comumente usadas, e relativamente fáceis de serem obtidas a partir dos dados da produção científica. Porém, como será visto no decorrer deste trabalho, essas métricas sofrem de diversas limitações quando o objetivo é revelar a qualidade de uma publicação ou de um pesquisador.

1. Produtividade

   A produtividade pode ser medida pela quantidade de trabalhos publicados por um pesquisador. Ela é interessante pois pode representar a capacidade de um autor concluir seus trabalhos. Porém ela não é significativa na avaliação da

qualidade desses trabalhos. Por exemplo, um autor poderia ter diversos artigos publicados, mas todos em veículos considerados de baixa qualidade.

2. Citações

   A partir de bases de dados de publicações científicas, hoje é possível contar quantas vezes um artigo foi referenciado por outros artigos, ou seja, quantas vezes ele foi citado. Essa métrica se tornou muito relevante para a discussão da importância de um trabalho, pois, como será visto mais à frente, pode representar a visão da comunidade sobre o trabalho.

   Um problema é que a quantidade de citações sozinha não é muito representativa. É possível imaginar que um pesquisador que publique frequentemente acumulará mais citações, o que não representa a qualidade de seu trabalho. E mesmo que façamos uma média, basta um trabalho excepcional que a média deixará de representar a qualidade da produção.

   Mais adiante será visto que existe um outro motivo que mostra que as métricas estatísticas clássicas, como a média, mediana, variância, etc., em cima das citações não são bons representantes da qualidade de um artigo ou pesquisador.

3. Índice-*h*

   O índice-*h* também é uma métrica baseada em citações. Porém ele tem uma definição diferente das métricas clássicas. O índice-*h* de um pesquisador é *h* se *h* de suas publicações têm pelo menos *h* citações (Hirsch, 2005). Por exemplo, se um pesquisador tem um índice-*h* igual a 15, significa que ele tem 15 publicações com mais que 15 citações, e todos os seus outros trabalhos têm menos que 15 citações.

   A grande vantagem dessa métrica é que ela não sofre variações devido publicações extraordinárias ou de baixa notoriedade. Por isso ela se tornou muito popular no meio acadêmico. Porém, além do problema que será visto na seção seguinte, o índice-*h* peca em avaliar pesquisadores que produzem pouco. Um exemplo é o caso de Peter Higgs, ganhador do prêmio Nobel de física em 2013, que tem um índice-*h* igual a 8, que é considerado baixo.

2.1.2 Modelos da Produção Científica

Existem modelos que tentam explicar matematicamente características da produção científica. Por exemplo, os modelos propostos por Wang e Barabási tentam explicar a quantidade de citações recebidas por trabalhos científicos através de modelos como o "fator *Q*". Esse modelo se propõe a calcular a quantidade de citações *c* que um trabalho feito em cima de uma ideia aleatória receberia de acordo com a habilidade do

pesquisador de desenvolvê-la. De uma maneira simplificada, $c = rQ$, onde r é um valor que representa o quão boa é a ideia, e Q é um valor que representa a habilidade do pesquisador em desenvolver as ideias. Quanto maiores forem r e Q, maior a quantidade de citações que a publicação receberá. Uma das conclusões que se pode tirar desse modelo é que autores que publicam mais frequentemente têm mais chances de encontrar um bom r e, então, produzir um trabalho de grande impacto (bastante citado).

Os autores desse modelo o desenvolvem mais, incluindo características da produção científica, como o crescimento exponencial da ciência, o envelhecimento do artigo, o *preferential attachment* (conceito de que novas publicações tendem a citar publicações com bastante citações), e a *fitness* (o quão capaz de resistir a outras características, como o envelhecimento).

Observação 1 (Foco na quantidade de citações) *Como foi visto nessa seção e na anterior, a discussão de como se avalia uma publicação científica é, a grosso modo, feita em cima da quantidade de citações que uma publicação recebeu.*

2.1.3 Formas Alternativas

Existem outras métricas que tentam avaliar melhor a qualidade das produções científicas. Uma das ideias propostas nesse sentido é a de avaliar apenas as publicações significativas, que teriam mais que um valor mínimo de citações (Wang e Barabási, 2021). Porém esse é um caminho subjetivo, onde não há concordância de qual é esse valor mínimo.

O objetivo dessas métricas em discussão é definir a qualidade de um trabalho, de forma que se possa comparar diferentes trabalhos. Para isso, seria possível fazer uma análise direta do conteúdo dessas produções, graças à recente disponibilidade de dados dessa natureza em diversas bases de dados. Mas essa abordagem é complicada, pois definir o que deve ser procurado no conteúdo do trabalho é subjetivo, e por isso há grandes divergências em relação a seus resultados. Outra dificuldade encontrada ao tentar fazer esse tipo de análise é que ela requer, relativamente, bastante recurso computacional. Por exemplo, bases de dados como o *DBLP*, que não armazenam o conteúdo dos artigos, já ocupam alguns gigabytes de armazenamento, não é difícil imaginar que extrapolando o tamanho do texto de um artigo para uma base como essa poderia ser necessário algumas centenas de gigabytes de armazenamento. Ainda, para processar essa quantidade de texto, dependendo do método utilizado, seria necessário tempo e recursos que podem não estar disponíveis para os pesquisadores.

Um ponto fundamental da avaliação da produção científica é que ela seja, além de mensurável, aceita pela comunidade científica. De nada adianta criar uma métrica que extraia a qualidade intrínseca de uma publicação, se ela não for aceita pelos membros da comunidade. Por isso é interessante olhar para a quantidade de

citações. Apesar de ela ser um dado externo ao trabalho, ela é uma representação do quão importante o trabalho é para a comunidade. Assim, fazer uma avaliação em cima da quantidade de citações é interessante, acima de tudo, pois se baseia na opinião da comunidade. E por isso tornamos evidente e explícita a Observação 1, pois, apesar de ser possível considerar métricas alternativas, iremos partir do pressuposto neste trabalho que, a grosso modo, a quantidade de citações reflete a qualidade de um trabalho.

Observação 2 (As citações representam a opinião da comunidade) *A comunidade expressa seu interesse por um trabalho citando ele. Logo quanto mais importante um tema for para ela, mais citado ele será.*

2.2 DISPARIDADE ENTRE ÁREAS

Há uma percepção empírica de que algumas áreas recebem mais citações que outras. Por exemplo, examinando os 100 trabalhos mais citados até 2014, é possível ver que produções em "técnicas de laboratório de biologia", além de serem os seis primeiros mais citados, aparecem com muito mais frequência que produções de outras áreas (Noorden et al., 2014). Nessa seção será discutida a veracidade dessa percepção.

Ao observar a quantidade de citações recebidas pela produção científica, verifica-se que quase metade de toda a produção sequer foi citada uma única vez. E descobre-se que menos de 0,1% de todas as publicações atingiram a marca de 1000 citações (Noorden et al., 2014). Isso coloca em dúvida a natureza da distribuição das citações desses trabalhos. Ao traçarmos o gráfico dessa distribuição fica claro que não se trata de uma distribuição normal (comumente encontrada em situações naturais), mas sim de uma distribuição lei de potência.

2.2.1 Distribuição Lei de Potência

Uma distribuição lei de potência é definida pela função: $f(x) = ax^{-c}$, onde $a > 0$ é uma constante, e c também é uma constante, tipicamente $2 \leq c \leq 3$ (Broido e Clauset, 2019). Essa definição matemática da distribuição pode tornar difícil a diferenciação dela a outras distribuições bem conhecidas, por exemplo a distribuição normal, como pode ser visto no gráfico (a) da Figura 2.1. Porém uma maneira fácil de diferenciá-la é traçando o gráfico da distribuição com ambos os eixos logarítmicos. Ao fazer isso é obtido um gráfico com uma reta, característica que diferencia a distribuição lei de potência de outras distribuições similares, como pode ser visto no gráfico (b) da Figura 2.1.

Uma das principais características de uma distribuição lei de potência é a "cauda longa". Essa característica tem a ver com o fato de que os valores altos de x são altamente representativos. Um caso particular da distribuição lei de potência

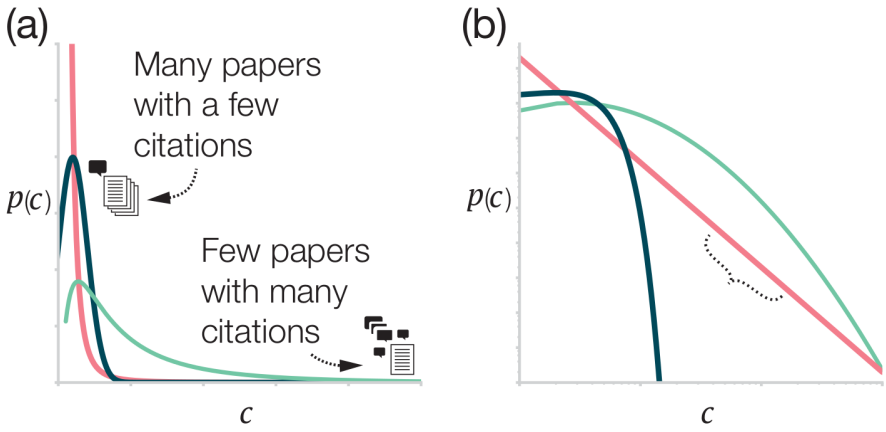

Figura 2.1: Exemplo das distribuições.
O gráfico (a) ilustra as distribuições lei de potência (rosa), normal (azul) e log-normal (verde) em uma plotagem linear-linear. Já o gráfico (b) mostra essas mesmas distribuições em uma plotagem log-log.

é a distribuição de Pareto, que é uma distribuição lei de potência, onde por meio de observações empíricas foram verificados vários casos em que 20% dos objetos é responsável por 80% do produto. Ou seja, a "cauda" da distribuição é extremamente relevante para o resultado geral do que está sendo observado. Outra característica notável das distribuições lei de potência é que elas são distribuições livres de escala, o que significa que, simplificadamente, qualquer recorte da distribuição ainda configura uma distribuição lei de potência (Broido e Clauset, 2019).

Observação 3 (Ausência de Média Representativa) *Devido às características da distribuição lei de potência, ela não tem uma média representativa (Gladwell, 2006).*

Um dos efeitos dessa característica de "cauda longa" é em relação a média e métricas como variância e desvio padrão. Nesse tipo de distribuição, uma quantidade exponencialmente pequena dos objetos da distribuição corresponde a uma quantidade exponencialmente grande do produto dela, e uma quantidade exponencialmente grande dos objetos corresponde a uma quantidade exponencialmente pequena do produto. Como essas métricas não conseguem extrair essa característica das distribuições lei de potência, elas não são significantes para analisar esse tipo de distribuição.

2.2.2 Analisando Diferentes Áreas

Ao traçar o gráfico da distribuição de citações é possível observar as características de uma distribuição lei de potência. Isso já indica que a análise desse tipo de dado deve ser feita com mais cuidado. A partir da previamente mencionada percepção de haver áreas mais citadas que outras, pesquisadores analisaram a distribuição das citações de diversas áreas. O que foi encontrado é que existem áreas onde publicações com 100 citações são até 50 vezes mais comuns que em outras, que é o caso de "biologia

evolucionária" comparada a "engenharia aeroespacial", como pode ser visto na Figura 2.2 (Radicchia et al., 2008).

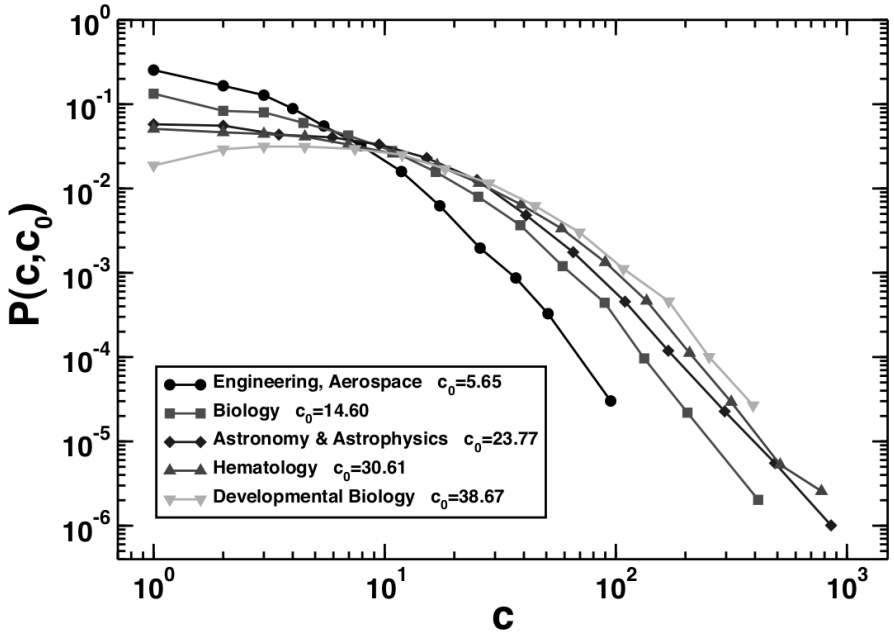

Figura 2.2: Distribuição de citações por área.
Nessa figura é possível observar que a probabilidade de uma publicação ter c citações (P(c,c_0)) varia bastante entre as áreas apresentadas. Também é apontada a média de citações de cada área do ano plotado c_0, que mostra uma diferença de 5,65 em "engenharia, aeroespaço" para 38,67 em "biologia evolucionária".

Isso se deve ao fato do padrão de citações em cada área ser diferente. Então Radicchia et al. propuseram uma maneira de normalizar o "sucesso" (quantidade de citações) de um trabalho. Eles perceberam que, apesar da média não ser uma métrica significativa para analisar a quantidade de citações, conforme evidenciado na Observação 3, ela poderia ser usada para capturar a diferença entre áreas distintas. Assim, a normalização proposta por Radicchia et al. pode ser descrita pela Equação 2.1, onde a quantidade de citações normalizada $\tilde{c}$ do trabalho i é a quantidade de citações dele c_i dividida pela média de citações μ na área a e ano de publicação y do trabalho i. Traçando o gráfico da distribuição de $\tilde{c}_i$ (Figura 2.3), é possível observar que a disparidade de citações entre as áreas desaparece, o que permite compará-la entre diferentes áreas.

$$\tilde{c}_i = \frac{c_i}{\mu_{a,y}} \tag{2.1}$$

Vale notar que os gráficos de distribuições lei de potência frequentemente são plotados a partir de uma distribuição relativa ao total de publicações, ou, ainda, a partir de uma distribuição cumulativa dos dados. Isso ocorre pois facilita a compreensão da característica de cauda longa. Por exemplo, no caso das citações da Figura 2.2 fica

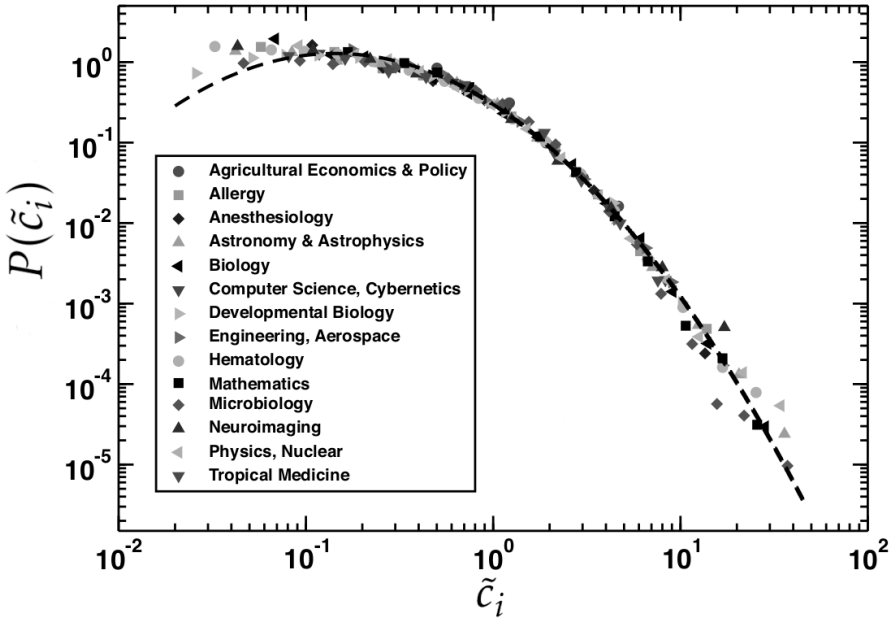

Figura 2.3: Distribuição de citações por área (normalizada).
Probabilidade $P(\tilde{c}_i)$ de se encontrar uma publicação com $\tilde{c}_i$ citações normalizadas separado por área.

claro que exponencialmente poucas publicações (10^{-6}) têm exponencialmente muitas citações (10^3).

2.3 OBSERVAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi visto que há uma necessidade de comparar publicações científicas e seus autores. Utilizando as recentes bases de dados sobre a produção científica é possível calcular métricas e testar modelos que ajudam a analisar e entender a qualidade de artigos e autores. Porém, como evidenciado na Observação 1, elas são fortemente baseadas na quantidade de citações que uma publicação recebe. Isso se torna um problema, visto que há uma disparidade na distribuição de citações entre áreas distintas.

Para resolver o problema dessas métricas e modelos, Radicchia et al. propuseram uma maneira de normalizar a quantidade de citações, dividindo a quantidade de citações recebidas por uma publicação pela média de citações na área e ano da publicação, conforme descrito na Equação 2.1. Essa normalização diminui a disparidade de citações entre áreas distintas, permitindo a comparação entre elas.

3 METODOLOGIA

Nesse capítulo serão descritos quais os dados obtidos para o experimento descrito no Capítulo 4 e quais as suas fontes. Ainda, será explicado como eles foram obtidos, armazenados e trabalhados.

3.1 OS DADOS E FONTES

1. *CSIndexbr*

   O objetivo do trabalho é analisar a disparidade de citações entre sub-áreas da ciência da computação brasileira. Para isso é preciso conhecer os autores brasileiros dessa área. Felizmente isso já é feito pelo *CSIndexbr*, um projeto brasileiro com objetivo de prover informações relevantes, abertas e transparentes sobre a produção científica brasileira de ciência da computação (Valente e Paixao, 2018).

   O *CSIndexbr* mantém e disponibiliza uma lista contendo todos os pesquisadores de ciência da computação brasileiros vinculados a instituições de ensino superior. Ele também mantém e disponibiliza uma lista relacionando veículos de publicação a áreas da ciência da computação. Esses dados serão importantes para obter informações sobre a produção brasileira e relacionar cada publicação a uma área.

2. *DBLP Computer Science Bibliography*

   O *DBLP* é um projeto alemão que tem como objetivo indexar a produção científica de ciência da computação para dar suporte a pesquisadores através dessa plataforma aberta e gratuita. Em 2019 o DBLP já havia indexado mais de 4,4 milhões de publicações de mais de 2,2 milhões de pesquisadores (Bibliography, 2022).

   Utilizando a lista de autores brasileiros obtida do *CSIndexbr*, o DBLP forncerá a lista de todas as publicações desses autores. O DBLP também traz a informação sobre em qual veículo a publicação foi feita, e, com as informações do *CSIndexbr*, é possível saber a qual sub-área da ciência da computação aquela publicação pertence.

3. *OpenCitations*

   O *OpenCitations* é uma organização italiana sem fins lucrativos ligada à Universidade de Bologna. O objetivo dela é dar acesso aberto a dados bibliográficos e

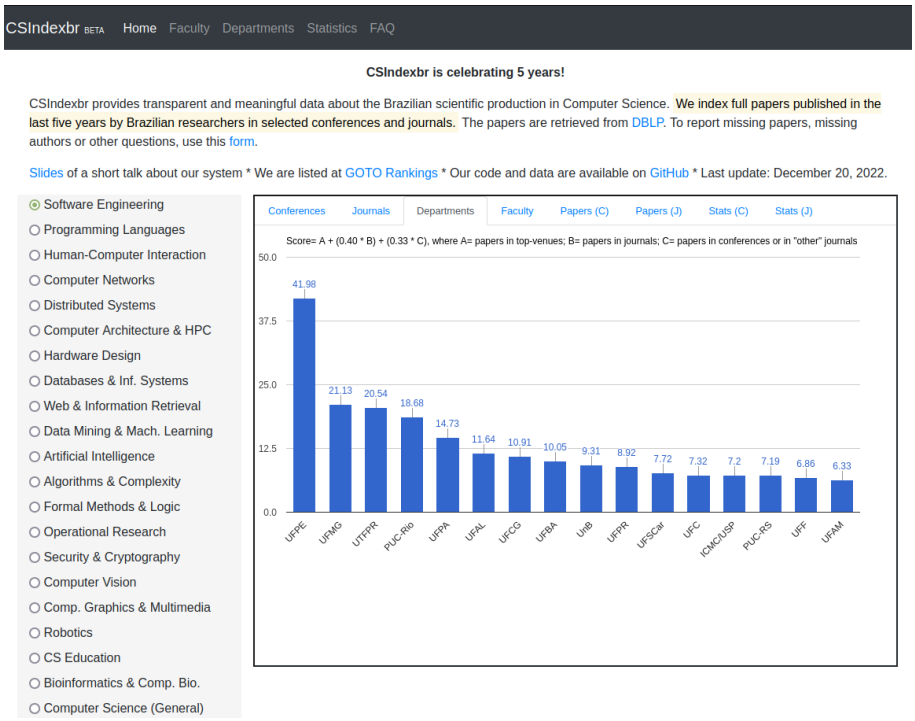

Figura 3.1: Site do *CSIndexbr*.

de citações de publicações (Waltman et al., 2020). Como o *DBLP* não fornece informações sobre as citações dos artigos, a lista de publicações obtidas dele será usada para obter a quantidade de citações recebidas por cada publicação a partir do *OpenCitations*.

3.2 TRABALHANDO OS DADOS

Para armazenar os dados obtidos foi criado um banco de dados no *PostgreSQL* com o diagrama apresentado da Figura 3.2.

Para popular todas as tabelas, inicialmente foi necessário obter a relação dos pesquisadores brasileiros da ciência da computação. Essa relação está disponível no *GitHub* do *CSIndexbr*. Dela foram pegos o nome, a instituição de ensio e o pid (identificador do pesquisador no *DBLP*) e inseridos na tabela "pesquisadores". Em seguida, para popular a tabela "publicacoes" foram feitas chamadas a API do *DBLP* usando o pid de cada pesquisador. Assim foram obtidos todos os dados das colunas da tabela, exceto a quantidade de citações recebidas. Simultaneamente foi se relacionando a autoria das publicações pela tabela "publicacoes_pesquisadores".

As sub-áreas da ciência da computação foram obtidas do *CSIndexbr* e inseridas na tabela "areas". Elas podem ser observadas na Figura 3.1. Para distinguir a qual área pertence uma publicação, é seguido, novamente, os passos do *CSIndexbr*, que relaciona veículos de publicação ("venues") às áreas (relação também disponível no *GitHub* do projeto). Assim, cada publicação está vinculada a um veículo, que está vinculado às

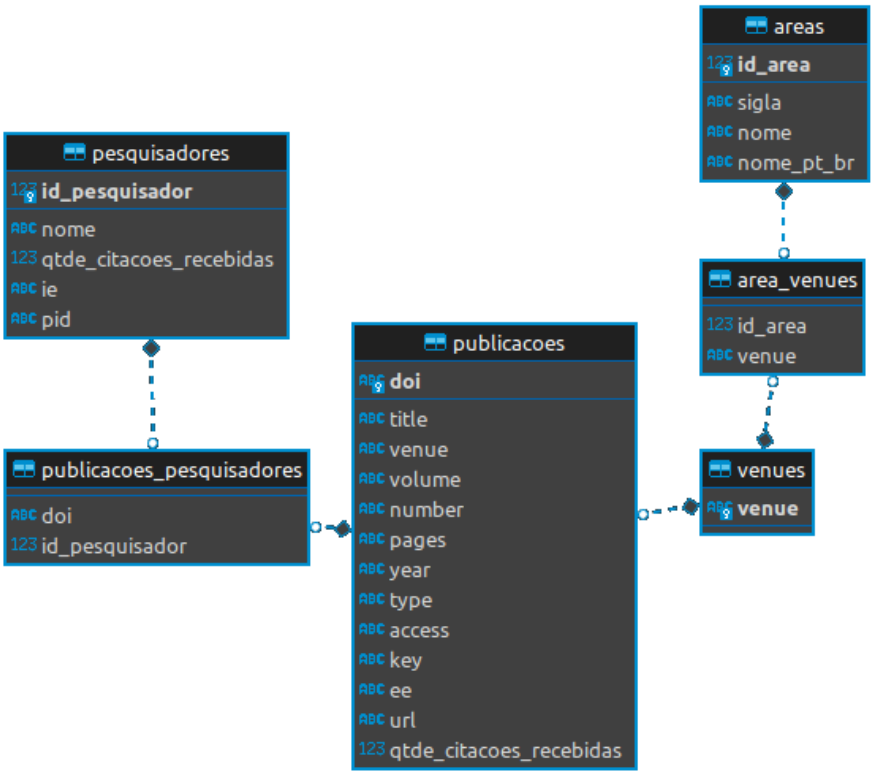

Figura 3.2: Diagrama do Banco de Dados.

sub-áreas. Vale notar que um veículo de publicação pode estar vinculado a mais de uma sub-área, e, portanto, uma publicação também.

Finalmente a quantidade de citações que uma publicação recebeu é obtida através de chamadas a API do *OpenCitations*, utilizando o DOI de cada publicação e atualizando a entrada correta na tabela "publicacoes". Assim foram obtidos todos os dados necessários para realizar o experimento do capítulo seguinte.

Nesse processo foram registrados mais de 1100 pesquisadores brasileiros da ciência da computação. Também foram encontrados quase 46 mil publicações. Porém apenas cerca de 11 mil puderam ser relacionadas a alguma sub-área da ciência da computação e serão objetos do experimento. Para processar, gerar os gráficos e analisar os dados foram utilizadas as bibliotecas *matplotlib* e *psycopg2* do *Python*. A distribuição dessas publicações por sub-área pode ser encontrada na Tabela 3.1.

3.3 OBSERVAÇÕES FINAIS DO CAPÍTULO

Obtendo os dados descritos nesse capítulo será possível fazer uma análise semelhante à descrita na Seção 2.2.2 sobre a ciência da computação brasileira. Além de observar as diferenças entre as áreas da ciência da computação, será possível, também, estabelecer os parâmetros de normalização necessários para diminuir a disparidade de citações de diferentes áreas.

Tabela 3.1: Quantidade de publicações por sub-área

Sub-área	#Publicações
Algoritmos e Complexidade	731
Arquitetura de Computadores	976
Base de Dados e Sistemas de Informação	522
Bioinformática	197
Ciência da Computação para Educação	48
Design de Hardware	264
Engenharia de Software	1458
Gráficos e Multimídia	601
Inteligência Artificial	1358
Interação Humano-Computador	188
Linguagens de Programação	251
Mineração de Dados e Apredizagem de Máquina	393
Métodos Formais e Lógica	367
Pesquisa Operacional	746
Redes de Computadores	1097
Robótica	275
Segurança e Criptografia	90
Sistemas Distribuídos	619
Visão Computacional	462
Web e Recuperação de Dados	423

4 RESULTADO E ANÁLISE

Obtendo os dados descritos no capítulo anterior, foram analisados mais de 11 mil publicações brasileiras da ciência da computação. Neste capítulo será visto como esses dados se encaixam na discussão feita até aqui. Também será apresentada uma análise dos parâmetros obtidos para comparar distintas áreas da ciência da computação brasileira.

Como vimos anteriormente, a distribuição de citações tem uma natureza peculiar. E a distribuição de citações na ciência da computação brasileira não é diferente. Existem áreas que são mais citadas que outras, por exemplo, ao observar as 100 publicações mais citadas (Tabela 4.1), é encontrado que áreas como "visão computacional" aparecem muito mais frequentemente que áreas como "robótica", e áreas como "interação humano-computador" sequer entram nessa lista.

Tabela 4.1: Quantidade de publicações por área entre as 100 mais citadas

	#Publicações no top 100
Visão Computacional	15
Redes de Computadores	13
Pesquisa Operacional	12
Gráficos e Multimídia	8
Inteligência Artificial	7
Base de Dados e Sistemas de Informação	7
Mineração de Dados e Aprendizagem de Máquina	7
Engenharia de Software	6
Bioinformática	4
Arquitetura de Computadores	4
Web e Recuperação de Dados	4
Algoritmos e Complexidade	3
Sistemas Distribuídos	3
Linguagens de Programação	2
Robótica	2
Segurança e Criptografia	2
Métodos Formais e Lógica	1
Ciência da Computação para Educação	0
Design de Hardware	0
Interação Humano-Computador	0

Outra característica é que 18% de todas as publicações, cerca de 1800 trabalhos, nunca foram citados (Figura 4.1). Também pode ser visto que a distribuição dos dados obtidos é uma distribuição lei de potência. E com isso o seu comportamento de "cauda longa", onde, nesse caso, 20% das publicações mais citadas representam 60% de todas

as citações desses trabalhos. Ao analisar esses dados, a forma mais direta de sumarizar resultados seria através de uma média. Porém, conforme a Observação 3, essa métrica não é significativa para esse tipo de dado, uma vez que seguem a distribuição lei de potência.

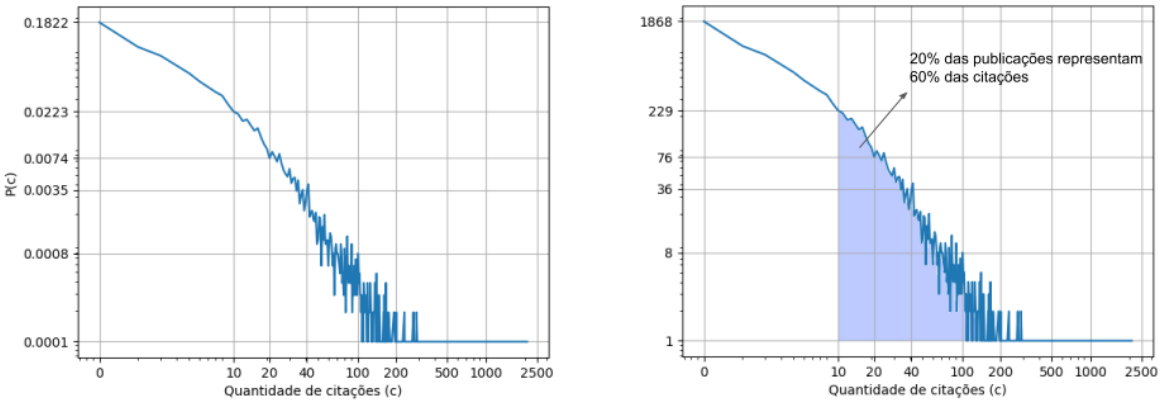

Figura 4.1: Distribuição de citações.
No gráfico à esquerda foi plotada a probabilidade $P(c)$ de se encontrar uma publicação com c citações. No gráfico à direita a mesma distribuição foi plotada em escalar onde a região hachurada representa 20% de todas as publicações, que tem 60% das citações do todo.

4.1 DISPARIDADE ENTRE AS ÁREAS DA CIÊNCIA DA COMPUTAÇÃO BRASILEIRA

Assim como foi observado em grandes áreas da ciência, na ciência da computação brasileira também há uma disparidade de citações entre suas sub-áreas. Nos dados analisados foi encontrado que publicações na área de "visão computacional" com pelo menos 100 citações são 11 vezes mais comuns que na área de "arquitetura de computadores". Como é possível observar na Figura 4.2 e na Tabela 4.2, as diferenças entre as sub-áreas da ciência da computação brasileira são bastante consideráveis. Como os dados estão divididos entre 20 sub-áreas diferentes, serão apresentadas apenas 6 sub-áreas nas figuras e tabelas como exemplo.

Na Figura 4.2 e na Tabela 4.2 pode se observar, por exemplo, que a sub-área de "Métodos Formais e Lógica" tem cerca de 10% de suas publicações com 20 ou mais citações. Já a sub-área de "Segurança e Criptografia" tem cerca de 31% de suas publicações com 20 ou mais citações. Ou seja, encontrar um trabalho com 20 ou mais citações é três vezes mais fácil em "Segurança e Criptografia" que em "Métodos Formais e Lógica". Vale notar que os dados apresentados na Figura 4.2 e na Tabela 4.2 são os casos mais extremos. As outras 14 sub-áreas se encaixam no meio desse gráfico, algumas mais ao extremo inferior, outras mais ao superior.

Dada a disparidade na distribuição de citações apresentada na análise dos dados, é necessário ajustar essa distribuição para mitigar a diferença entre as áreas.

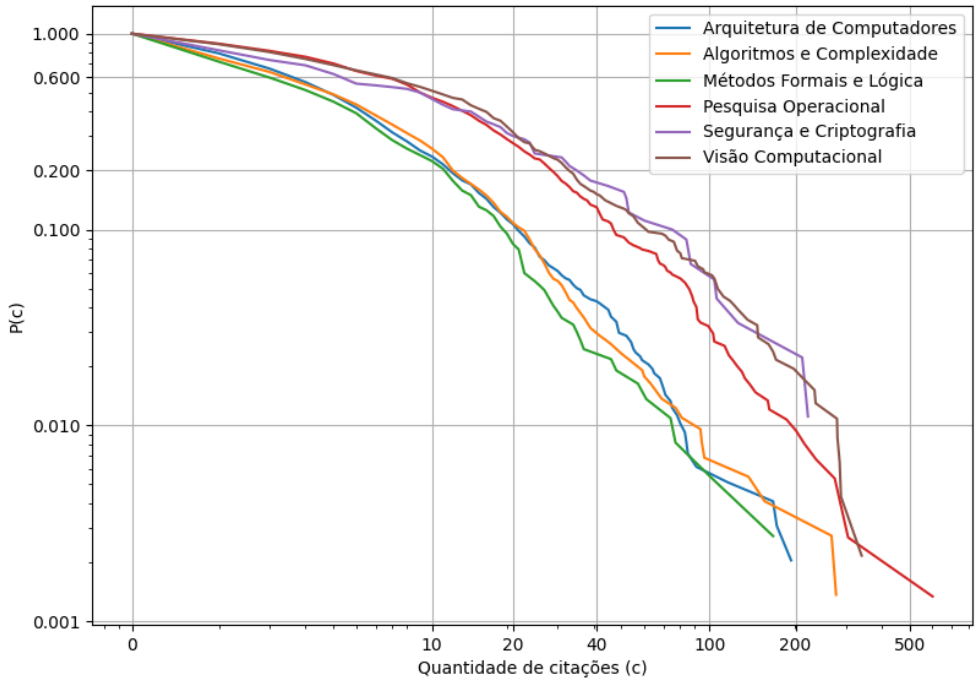

Figura 4.2: Distribuição de citações por área.
Probabilidade $P(c)$ de encontrar publicações com c ou mais citações por área.

Tabela 4.2: Probabilidade $P(c)$

Sub-Área	c=20	c=40	c=100
Visão Computacional	0,335	0,155	0,058
Segurança e Criptografia	0,311	0,166	0,055
Pesquisa Operacional	0,289	0,130	0,032
Métodos Formais e Lógica	0,095	0,021	0,005
Algoritmos e Complexidade	0,116	0,028	0,005
Arquitetura de Computadores	0,112	0,043	0,005

Como descrito em trabalhos anteriores será utilizada a Equação 2.1 para normalizar a distribuição de citações entre as sub-áreas (Radicchia et al., 2008). Com esse objetivo, para cada sub-área é calculada a média de citações para cada ano $\mu_{a,y}$, então a quantidade de citações de uma publicação c_i é dividida pela $\mu_{a,y}$ correspondente, obtendo $\tilde{c}_i$ (veja detalhes na Seção 2.2.2, Equação 2.1). Assim, é obtida uma distribuição mais representativa da importância de cada trabalho na ciência da computação. Como pode se observar na Figura 4.3 e na Tabela 4.3, com essa distribuição ajustada agora é possível comparar trabalhos de diferentes sub-áreas da ciência da computação, visto que as distribuições estão mais similares. Vale notar que essa distribuição ajustada continua sendo uma distribuição lei de potência, portanto a Observação 3 continua relevante.

Observando a Figura 4.3 e a Tabela 4.3 é possível perceber que a diferença previamente mencionada entre "Métodos Formais e Lógica" e "Segurança e Criptografia" é bastante mitigada. E isso se aplica a todas as sub-áreas analisadas. Nessa distribuição $P(\tilde{c})$ tem uma diferença negligenciável entre as sub-áreas. Assim, $\tilde{c}$ pode ser utilizado para comparar a produção científica entre elas, empregando métricas como o índice-h e modelos como o discutido na Seção 2.1.2.

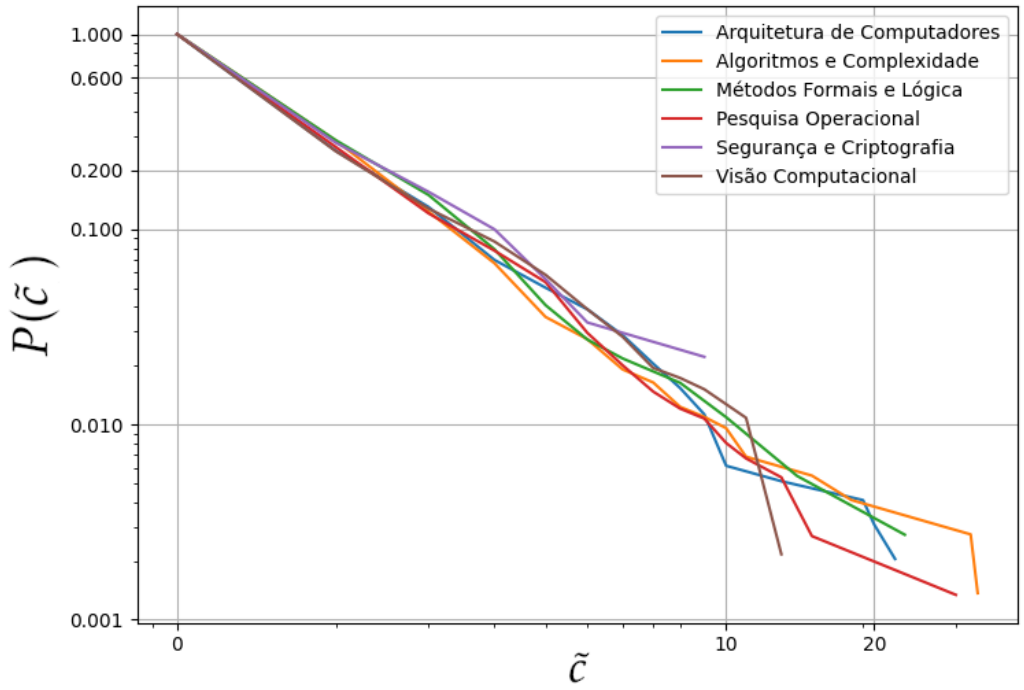

Figura 4.3: Distribuição de citações por área normalizado.
Probabilidade $P(\tilde{c})$ de encontrar publicações com $\tilde{c}$ ou mais citações normalizadas por área.

Tabela 4.3: Probabilidade $P(\tilde{c})$

Sub-Área	$\tilde{c}$=2	$\tilde{c}$=4	$\tilde{c}$=8
Visão Computacional	0,251	0,086	0,019
Segurança e Criptografia	0,277	0,100	0,022
Pesquisa Operacional	0,264	0,077	0,014
Métodos Formais e Lógica	0,286	0,079	0,016
Algoritmos e Complexidade	0,284	0,067	0,016
Arquitetura de Computadores	0,253	0,069	0,020

5 CONCLUSÃO

Como foi visto durante o trabalho, avaliar a produção científica é bastante importante. Para isso, há uma grande discussão de como fazer essa avaliação. A grande dificuldade é que os dados objetivos obtíveis dessa produção, em especial a quantidade de citações, não podem ser diretamente comparadas. Existem propostas para tentar avaliar essa produção sem usar a quantidade de citações, mas elas esbarram no problema da subjetividade. Outra característica que torna a quantidade de citações o ponto central dessa avaliação é o fato de ela, naturalmente, representar uma opinião da comunidade em relação ao trabalho (Observação 2). O que é muito importante para existir um consenso e uma aceitação do método avaliativo.

Assim como foi observado em outros trabalhos, na análise aqui feita foi mostrado que a distribuição das citações não é igual para diferentes áreas. Isso impede uma comparação justa entre trabalhos de áreas distintas, pois, de fato, existem áreas mais citadas que outras. Também, foi verificado que o fator de normalização proposto anteriormente para aproximar a diferença da distribuição entre áreas funciona. E isso possibilita uma análise mais justa da qualidade de todos os trabalhos da ciência da computação brasileira.

Porém vale lembrar que esses resultados devem ser considerados dentro de propostas de modelagem e métricas. Como diz a Observação 1, o foco dessas propostas está na quantidade de citações, porém todas elas pecam ao não considerarem as diferenças entre áreas. No caso dos modelos, apesar de os melhores modelos representarem bem a produção científica e conseguirem fazer predições precisas, eles podem acabar dando um viés errado em uma possível análise que os utilize, trazendo um resultado injustamente mais positivo para pesquisas de uma área que naturalmente é mais citada. E métricas como o índice-h também devem ser ajustadas levando essa discussão em consideração, que, se bem difundidas, podem levar a um entendimento mais justo do que é considerado um bom valor para a métrica.

REFERÊNCIAS

Bibliography, D. C. S. (2022). Monthly snapshot release of november 2022. https://dblp.org/xml/release/dblp-2022-11-01.xml.gz. Acessado em 30/11/2022.

Broido, A. D. e Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, 10(1017).

Gladwell, M. (2006). Million-dollar murray. *New Yorker, DEPT. OF SOCIAL SERVICES*, 82(1).

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Noorden, R. V., Maher, B. e Nuzzo, R. (2014). The top 100 papers. *Nature*, 514(7524):550–553.

of Science, W. (2023). Web of science platform. https://clarivate.com/webofsciencegroup/solutions/webofscience-platform/. Acessado em 30/01/2023.

Radicchia, F., Fortunatoa, S. e Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. Em *Proceedings of the National Academy of Sciences*, página 17268 –1727, Torino - Italy.

Valente, M. T. e Paixao, K. (2018). CSIndexbr: Exploring the Brazilian scientific production in Computer Science. *arXiv*, abs/1807.09266.

Waltman, L., Larivière, V., Milojević, S. e Sugimoto, C. R. (2020). Opening science: The rebirth of a scholarly journal. *Quantitative Science Studies*, 1(1):1–3.

Wang, D. e Barabási, A.-L. (2021). *The Science of Science*. Cambridge University Press.