
UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

LEONARDO RIBEIRO DA CUNHA

COMPLEXIDADE DE PADRÕES
ESTATÍSTICOS EM JOGOS *ONLINE* E NO
COMPORTAMENTO COLETIVO DE PEIXES

Maringá, Setembro de 2024.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

LEONARDO RIBEIRO DA CUNHA

COMPLEXIDADE DE PADRÕES
ESTATÍSTICOS EM JOGOS *ONLINE* E NO
COMPORTAMENTO COLETIVO DE PEIXES

*Trabalho apresentado ao Programa de Pós-Graduação
em Física da Universidade Estadual de Maringá como
tese de doutorado.*

Orientador: Prof. Dr. Renio dos Santos Mendes

Maringá, Setembro de 2024.

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

C972c

Cunha, Leonardo Ribeiro da

Complexidade de padrões estatísticos em jogos online e no comportamento coletivo de peixes / Leonardo Ribeiro da Cunha. -- Maringá, PR, 2024.
121 f. : color., figs., tabs.

Orientador: Prof. Dr. Renio dos Santos Mendes.

Tese (doutorado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Física, Programa de Pós-Graduação em Física, 2024.

1. Sistemas complexos. 2. Dinâmica de popularidade. 3. Videogames. 4. Padrão comportamental - Peixes. 5. Conclusão de tarefas. I. Mendes, Renio dos Santos, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Física. Programa de Pós-Graduação em Física. III. Título.

CDD 23.ed. 530.12

LEONARDO RIBEIRO DA CUNHA

**COMPLEXIDADE DE PADRÕES ESTATÍSTICOS EM JOGOS *ONLINE* E NO
COMPORTAMENTO COLETIVO DE PEIXES**

Tese apresentada à Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de doutor.

Aprovado em: Maringá, 20 de setembro de 2024.

BANCA EXAMINADORA

Prof. Dr. Renio dos Santos Mendes
Universidade Estadual de Maringá – UEM

Prof. Dr. José Roberto Iglesias
Universidade Federal do Rio Grande do Sul - UFRGS

Prof. Dr. Sílvio da Costa Ferreira Júnior
Universidade Federal de Viçosa - UFV

Prof. Dr. Luis Carlos Malacarne
Universidade Estadual de Maringá – UEM

Prof. Dr. Sérgio de Picoli Júnior
Universidade Estadual de Maringá – UEM

Agradecimentos

Ao escrever esta seção, sou tomado por um misto de gratidão e nostalgia. Ao concluir este ciclo de formação acadêmica, reconheço o privilégio que foi vivenciar cada etapa dessa jornada.

Primeiramente, agradeço à minha mãe, Luzia (ou Cida, como é carinhosamente conhecida), e ao meu pai, Daniel. Vocês foram exemplos de honestidade e resiliência, sempre presentes, sempre sendo minha fortaleza e oferecendo apoio incondicional. Agradeço também a todos os familiares que torceram por mim, tias, tios, primas, primos, e tantos outros. A família é grande, e o apoio foi imenso.

Minha gratidão se estende a todos os membros e ex-membros do Complex Lab. Em especial, gostaria de agradecer ao Denner e ao Higor, que me receberam calorosamente no laboratório, me ensinaram tanto e proporcionaram boas conversas. Ao André e ao Arthur, minha gratidão é ainda maior: vocês foram mais do que colegas de trabalho, foram quase irmãos. Agradeço também ao Haroldo, uma das mentes mais brilhantes que já conheci, por cativar a todos com sua proatividade e por ser um líder nato. Ao professor Renio, meu orientador, expresso um agradecimento especial, pois foi ele que tornou possível todo esse processo de doutoramento. Embora não possa citar todos nominalmente, quero que saibam que cada um de vocês no Complex Lab contribuiu para criar um ambiente acolhedor e inspirador. Vivi momentos inesquecíveis, experiências que levarei para toda a vida.

Aos amigos, cujas contribuições foram fundamentais tanto para minha formação acadêmica quanto pessoal, meu profundo agradecimento. Para evitar a injustiça de esquecer alguém, menciono vocês pelos nomes dos grupos de WhatsApp. Aos amigos do grupo “APS”, amizades raras e verdadeiras que começaram no ensino médio e continuam até hoje, meu muito obrigado. Aos membros do “Fakes Masters”, obrigado por sempre proporcionarem discussões úteis e risadas garantidas. Aos integrantes do “Fut Info UEM”, agradeço por me ajudarem a vencer o sedentarismo com nosso futsal descontraído. Ao pessoal do “Salinha VIP”, obrigado por dividir comigo a atmosfera da pós-graduação em Física. E um agradecimento especial ao Igor e ao Emanuel (ou “Mané”), amigos de longa data que sempre garantiram conversas hilárias.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao programa de Pós-graduação em Física da UEM (PFI) pelo apoio financeiro por meio das bolsas de estudo, bem como pela infraestrutura disponibilizada para o desenvolvimento desta pesquisa.

Resumo

A análise de sistemas complexos dispõe de ferramentas físicas, matemáticas e estatísticas para investigar fenômenos nos mais variados contextos. O atual trabalho emprega algumas dessas ferramentas no estudo de um sistema social (jogos *online*) e um biológico (cardumes de carpas prateadas). Objetivando construir um arcabouço teórico para o desenvolvimento de pesquisa de sistemas complexos, no capítulo 1, abordamos definições e conceitos da teoria de probabilidade. Após definir essas bases teóricas, no capítulo 2, investigamos a dinâmica de conclusão de tarefas no contexto de jogos *online*. Nessa direção, aproveitamos o dado de uma grande plataforma *online*, chamada *Steam*, para analisar características quantitativas dos jogos e da persistência dos seus jogadores. Sobretudo, o principal resultado foi a detecção de um decaimento exponencial médio ao considerar a proporção de jogadores em função do número de *achievements*. No capítulo 3, investigamos padrões na dinâmica de popularidade dos jogos. Para esse fim, utilizamos outro banco de dados disponibilizado pela *Steam*. Nessa investigação, detectamos cinco padrões de forma de popularidade, sendo que um deles, o padrão decrescente, representa a dinâmica de quase metade dos jogos. Além disso, comparamos os primeiros meses das séries de popularidade dos jogos às séries completas e encontramos que a grande maioria dos jogos tendem a manter seu padrão inicial de popularidade com o passar do tempo. No capítulo 4, analisamos a diferença no padrão de comportamento de carpas prateadas (*Hypophthalmichthys molitrix*) sob uma estrutura de sombra e fora da sombra. Os dados empregados nessa pesquisa foram obtidos via experimentos conduzidos em lagoas artificiais na Columbia Environmental Research Center, em Columbia no Missouri, Estados Unidos. Dois resultados foram destacados. Primeiro, verificamos que a sombra encoraja a residência das carpas prateadas durante o dia. Segundo, fora da estrutura de sombra, as carpas exibem estados de maior polarização, movendo-se alinhadas umas com as outras, o que resulta em um comportamento típico de cardume.

Palavras-chave: Sistemas complexos. Videogames. Conclusão de tarefas. Dinâmica de popularidade. Padrão comportamental dos peixes.

Abstract

The analysis of complex systems include physical, mathematical and statistical tools to investigate phenomena in several contexts. The current work shows some of these main tools and applications in social (online games) and biological (schools of silver carp) systems. In order to build a theoretical framework for development of complex systems research, in Chapter 1, we discuss definitions and concepts of probability theory. After defining these theoretical foundations, in Chapter 2, we investigate the dynamics of task completion in the context of online games. In this direction, we take advantage of data from a large online platform, called Steam, to analyze quantitative characteristics of games and the performance of their players. Especially, the main result was the detection of an average exponential decay when considering the proportion of players as a function of the number of achievements. In Chapter 3, we investigate patterns in the dynamics of game popularity. To this end, we used another dataset provided by Steam. In this investigation, we detect five shape patterns of popularity, one of which, the decreasing pattern, represents the dynamics of almost half of the games. Additionally, we compared the first few months of the games' popularity series with the entire series and found that the vast majority of games tend to maintain their popularity pattern over time. In Chapter 4, we analyze differences in behavioral patterns of silver carp (*Hypophthalmichthys molitrix*) under a shadow structure and outside the shadow. The data employed in this research were obtained through experiments conducted in artificial ponds at the Columbia Environmental Research Center, in Columbia, Missouri, United States. Two results were highlighted. First, we found that shade encourages the residence of silver carp during the day. Second, outside the shade structure, the carp exhibit higher polarization states, moving in alignment with each other, resulting in a typical schooling behavior.

Keywords: *Complex systems. Videogames. Task completion. Popularity dynamics. Behavioral pattern of fish.*

Introdução	10
1 Fundamentação teórica	13
1.1 Introdução a modelos probabilísticos	13
1.2 Medidas de centralidade	17
1.3 Momentos da distribuição e medida de dispersão	18
1.4 Assimetria e curtose	19
1.5 Modelos probabilísticos para múltiplas variáveis	21
1.6 Exemplos de distribuições	23
1.6.1 Distribuição uniforme	23
1.6.2 Distribuição normal	24
1.6.3 Distribuição log-normal	25
1.6.4 Distribuição exponencial	26
1.7 Lei dos grandes números	27
1.8 Teorema central do limite	28
1.9 Estimador de máxima verossimilhança	29
1.10 <i>Bootstrap</i>	31
1.11 Método dos mínimos quadrados	33
1.12 Coeficiente de determinação	34
2 Complexidade da popularidade e dinâmica das conquistas dentro dos jogos online	35
2.1 Introdução	36
2.2 Conjunto de dados	38
2.3 Resultados	40

2.4	Discussão	47
3	Padrões da dinâmica de popularidade dos jogos <i>online</i>	52
3.1	Introdução	52
3.2	Conjunto de dados	54
3.3	Métodos	61
3.4	Resultados	63
3.5	Discussão	70
4	Efeitos de ambientes sombreados no comportamento de peixes	74
4.1	Introdução	74
4.2	Métodos	76
4.3	Resultados	80
4.4	Discussão	88
	Considerações finais e perspectivas	90
A	Demonstração do Teorema central do limite	92
B	Dados gerados aleatoriamente	97
C	Figuras complementares do capítulo 3	98
D	Figuras complementares do capítulo 4	102
	Referências bibliográficas	107

A conceituação de sistemas complexos é bem abrangente. Assim, uma listagem precisa de todas as características que definem exatamente um sistema complexo se torna uma tarefa inviável. Apesar disso, certas propriedades são comumente associadas a sistemas complexos. Talvez, a principal propriedade de sistemas complexos seja a observação de que eles são geralmente constituídos de um grande número de componentes interagentes, fator determinante na representação canônica de sistemas complexos como redes complexas [1, 2]. Nesse tipo de representação, os nós simbolizam os componentes e as interações são caracterizadas pelas ligações entre os nós. Essas interações, por vezes não locais ou não lineares, levam ainda a comportamentos emergentes e auto-organizados, sem qualquer expressão de um controle central, outra propriedade comum a sistemas complexos [3]. Dessa maneira, por mais que os componentes apresentem suas particularidades, sistemas complexos revelam regularidades (mesmo que apenas estatísticas) quando consideramos o sistema como um todo [4]. Nesse contexto, poderíamos imediatamente pensar em duas abordagens de estudo de sistemas complexos, a modelagem matemática via equações diferenciais ou modelos computacionais (uma abordagem “mecanística”) e uma abordagem com foco nas observações empíricas e ferramentas estatísticas (modelagem baseada em dados) [5]. Em especial, nesta tese, focamos na abordagem baseada em dados para analisar um sistema social (jogos *online*) e um sistema biológico (cardume de peixes).

Como já deve estar claro, muitos sistemas em diversos contextos podem ser abordados como sistemas complexos. Esses sistemas podem apresentar uma ampla gama de fenômenos, que vão desde os naturais e biológicos até os econômicos e sociais [6–22]. Um exemplo clássico de análise de sistemas complexos no contexto de fenômenos naturais é o estudo das ocorrências de terremotos [23]. De acordo com resultados bem documentados na literatura, correlações temporais de curto alcance entre os terremotos é descrita pela Lei de Omori [24, 25]. Essa lei sugere que, após um terremoto principal, ocorre uma sequência de

tremores subsequentes cuja frequência f diminui ao longo do tempo t , seguindo a relação $f = t^{-\alpha}$, em que $\alpha \approx 1$. Em biologia, por exemplo, espécies podem ser interpretadas como os componentes de um sistema complexo, com as interações entre esses componentes estabelecidas pelas relações ecológicas entre as espécies [26–28]. Em uma vertente de aplicação em fenômenos econômicos, pode-se investigar a valorização e a desvalorização de produtos e empresas no mercado financeiro [29–31]. Nos últimos anos, um destaque tem sido o crescente interesse nas criptomoedas. Nesse contexto, estudos foram realizados para testar a hipótese de eficiência de mercado usando a série temporal do valor de mercado das criptomoedas [32, 33]. Em contextos sociais, uma empresa pode ser vista como um dos muitos sistemas complexos possíveis. Nesse caso, os trabalhadores seriam os componentes do sistema e suas interações representariam as relações de trabalho entre os componentes [34–36]. Outra aplicação em fenômenos sociais está relacionada a leis alométricas que relacionam índices sociais de sistemas urbanos à sua população [37, 38]. Enfim, em meio a tantos exemplos, fica evidente a relevância dos estudos de sistemas complexos.

A interdisciplinaridade nos estudos de sistemas complexos tem sido um grande trunfo para seu sucesso, mas um elemento chave para esse tipo de produção científica é a qualidade dos dados utilizados. Trabalhos científicos são cada vez mais dependentes da grande quantidade e alta qualidade dos dados empregados e não deixam dúvidas da importância central que novas observações e medições têm na ciência [39]. Além disso, os avanços tecnológicos no setor de processamento e armazenamento de dados possibilitam a circulação livre e aberta de diversos bancos de dados na internet, ampliando ainda mais o potencial das pesquisas [40]. Inclusive, foi cunhado o termo *big data*, que se tornou popular nos últimos anos para denotar esses grandes volumes de dados [41]. A popularização do *big data* revelou a necessidade de ferramentas para manipular grandes quantidades de dados de forma rápida e eficiente. Via de regra, essas ferramentas computacionais são baseadas em linguagens de programação. Em especial, a linguagem *Python* tem se destacado como uma linguagem de programação cada vez mais popular na comunidade científica [42]. Nela, por meio de bibliotecas disponíveis abertamente, é possível extrair dados das mais variadas fontes, organizá-los em *data frames*, manipulá-los matematicamente e construir gráficos de alta qualidade [43, 44].

Aproveitando a disponibilidade de dados circulando pela internet, parte do presente trabalho é dedicada ao estudo de jogos eletrônicos (videogames) a partir de bancos de dados disponíveis abertamente. Atualmente, os jogos eletrônicos são uma forma difundida de entretenimento entre crianças, jovens e adultos [45]. A globalização facilita o acesso aos jogos, enquanto que os avanços gráficos e melhorias na jogabilidade conquistam cada vez mais jogadores ao redor do mundo [46]. Nesse contexto, é possível perceber que os jogos eletrônicos já se tornaram parte do cotidiano das pessoas e que, portanto, estudar o comportamento humano dentro desses ambientes digitais significa, em certa medida, estudar o comportamento social humano. Em outras palavras, dados de jogos eletrônicos podem oferecer um

entendimento sobre padrões de comportamento humano e suas mudanças ao longo do tempo. Nesse sentido, nossa pesquisa foi dividida em duas partes: primeiramente, exploramos a persistência dos jogadores em concluir tarefas dentro dos jogos [47]; em seguida, investigamos padrões na dinâmica de popularidade dos jogos [48].

A outra parte deste trabalho consiste no estudo da dinâmica de movimento de uma espécie de peixe, conhecida como carpa prateada (*Hypophthalmichthys molitrix*). No campo da biologia tem havido um crescente interesse no estudo de espécies invasoras [49]. Por definição, espécies invasoras são seres vivos que de alguma forma foram transportados de uma região para outra e nessa nova região se proliferam de maneira descontrolada, ameaçando o equilíbrio do ecossistema [50]. Por conta disso, investigar as interações da espécie invasora com o meio e as outras espécies é de fundamental importância para o controle dos impactos ecológicos. Por exemplo, a carpa prateada (*Hypophthalmichthys molitrix*) é uma espécie nativa da China e foi importada para os Estados Unidos no início da década de 1970 para melhorar a qualidade da água em lagoas de retenção [51]. No entanto, por conta de fortes inundações, indivíduos da espécie escaparam e sua população têm crescido descontroladamente. Atualmente, ela é considerada uma espécie invasora no rio Mississippi, nos Estados Unidos [52]. Nesse contexto, decorrente de uma parceria com a Columbia Environmental Research Center, em Columbia no Missouri, tivemos a oportunidade de acesso aos dados de um experimento conduzido em lagoas artificiais, que possibilitou o desenvolvimento de um trabalho sobre padrões de movimentação da carpa prateada [53].

Esta tese está dividida em quatro capítulos e as considerações finais. O primeiro capítulo foca em apresentar as ferramentas básicas de probabilidade e estatística utilizadas no decorrer deste trabalho e por esse motivo denominamos o capítulo 1 de “Fundamentação teórica”. Nele introduzimos o conceito de probabilidade, os conceitos mais importantes da teoria de probabilidade, exemplos de distribuições de probabilidade, formas de estimar parâmetros, métodos para quantificar a qualidade de uma estimativa e noções de correlação estatística. No capítulo 2, intitulado “Complexidade da popularidade e dinâmica das conquistas dentro dos jogos *online*”, apresentamos uma investigação sobre a dinâmica de persistência das pessoas em concluir tarefas em jogos *online*. De forma mais específica, analisamos a taxa de sucesso dos jogadores ao longo de uma sequência de tarefas, denominadas *achievements*, de diferentes jogos [54]. No capítulo 3, chamado “Padrões da dinâmica de popularidade dos jogos *online*”, evidenciamos a existência de padrões de forma na dinâmica de popularidade dos jogos *online*. Obtemos esses padrões empregando uma abordagem de análise de *clustering* nos dados da média mensal de jogadores dos jogos da Steam [48]. No capítulo 4, denominado “Efeitos de ambientes sombreados no comportamento de peixes”, mostramos uma análise sobre os padrões de movimentação das carpas prateadas. Sobretudo, analisamos a diferença no comportamento das carpas em ambientes com e sem uma estrutura de sombra [47]. Nas “Considerações finais”, resumizamos todos os resultados obtidos.

Como já foi dito na Introdução, ferramentas de estatística e probabilidade são imprescindíveis para a análise de sistemas complexos. Neste capítulo, abordamos uma breve revisão de noções de probabilidade, em que expomos a definição de probabilidade e comentamos alguns conceitos indispensáveis da teoria de probabilidade, como média, desvio, coeficiente de assimetria e curtose. Aproveitamos e mostramos exemplos de distribuições de probabilidades, como uniforme, normal, log-normal e exponencial. Expomos dois teoremas fundamentais da teoria de probabilidade “Lei dos grandes números” e “Teorema central do limite”. Nas seções finais discutimos formas de estimar parâmetros e quantificar a qualidade das estimativas, como estimador de máxima verossimilhança, *bootstrap*, método dos mínimos quadrados e o coeficiente de determinação. O objetivo deste capítulo é fornecer um arcabouço teórico para facilitar a compreensão dos demais capítulos. Portanto, o leitor já familiarizado com os conceitos citados pode omitir a leitura deste capítulo e avançar para o capítulo seguinte.

1.1 Introdução a modelos probabilísticos

À primeira vista, os fenômenos aleatórios não parecem ter padrões e, portanto, não haveria uma forma sistemática de analisá-los. Porém, após um olhar mais cuidadoso, é possível evidenciar regularidades nos fenômenos aleatórios [55]. No lançamento de uma moeda, por exemplo, é sabido que existe a metade de chance dela cair com a face “cara” para cima. Esse resultado pode ser verificado após vários lançamentos e anotando a frequência com que ocorrem as caras. Então, o que é possível é prever, dentro de um intervalo de confiança, os resultados e suas respectivas probabilidades para determinados fenômenos [56].

Os modelos nos quais se mensuram as chances de cada resultado são chamados de modelos

probabilísticos ou modelos de probabilidade. O conceito de probabilidade pode ser construído pelo quociente entre a frequência de ocorrência de um resultado específico e o número total de ensaios. Outra forma de visualizar o conceito de probabilidade é associar números não negativos às diferentes possibilidades, em que a soma de todos esses números resulte em um. Esse processo chama-se normalização. Logo, seja l um possível resultado de um fenômeno aleatório associado à probabilidade p_l ,

$$p_l \geq 0, \quad (1.1)$$

cujas normalização é dada por:

$$\sum_l p_l = 1. \quad (1.2)$$

Utilizando a linguagem probabilística, l é chamado de variável aleatória e p_l é a massa de probabilidade da variável l .

O conjunto de todas as combinações possíveis para l pode ser chamado de espaço amostral, comumente representado pela letra $\mathcal{S}$. Por exemplo, seja l a variável aleatória que represente a soma de dois dados comuns de 6 lados. Nesse caso, a variável l pode assumir qualquer valor do seguinte conjunto $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, e leva ao espaço amostral

$$\mathcal{S} = \left\{ \begin{array}{l} (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) \\ (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) \\ (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) \\ (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) \\ (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) \\ (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \end{array} \right\}. \quad (1.3)$$

Já as probabilidades associadas aos valores de l são, então

$$\begin{aligned} p_2 &= \frac{1}{36}, & \{(1, 1)\}, \\ p_3 &= \frac{2}{36}, & \{(1, 2), (2, 1)\}, \\ p_4 &= \frac{3}{36}, & \{(1, 3), (2, 2), (3, 1)\}, \\ p_5 &= \frac{4}{36}, & \{(1, 4), (2, 3), (3, 2), (4, 1)\}, \\ p_6 &= \frac{5}{36}, & \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}, \\ p_7 &= \frac{6}{36}, & \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}, \\ p_8 &= \frac{5}{36}, & \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}, \\ p_9 &= \frac{4}{36}, & \{(3, 6), (4, 5), (5, 4), (6, 3)\}, \\ p_{10} &= \frac{3}{36}, & \{(4, 6), (5, 5), (6, 4)\}, \\ p_{11} &= \frac{2}{36}, & \{(5, 6), (6, 5)\}, \\ p_{12} &= \frac{1}{36}, & \{(6, 6)\}. \end{aligned} \quad (1.4)$$

Na Figura 1.1 representamos graficamente a distribuição de probabilidade desse exemplo.

Nem sempre os problemas têm um conjunto discreto de resultados e nesses casos as possibilidades preenchem um intervalo contínuo de valores. Naturalmente, surge o conceito de variável aleatória contínua e de densidade de probabilidade. A densidade de probabilidade é uma função da variável aleatória, assume valores não negativos e sua integração sob todos os valores resulta em 1 (normalização no caso contínuo). Sendo x uma variável aleatória contínua e $\rho(x)$ sua densidade de probabilidade, a definição fica:

$$\rho(x) \geq 0, \quad (1.5)$$

e a normalização é dada por

$$\int_{-\infty}^{\infty} \rho(x) dx = 1. \quad (1.6)$$

Existe uma diferença sutil na construção do conceito de probabilidade no caso discreto e no caso contínuo. Para variáveis discretas, cada evento l tem uma probabilidade associada p_l , então, caso surja o interesse em mais de um resultado, basta somar as probabilidades associadas. Desse modo, é possível considerar $l = 1, l = 2, \dots, l = n$ resultados de interesse, sabendo que l tem N possíveis resultados ($N > n$). Então, a probabilidade de

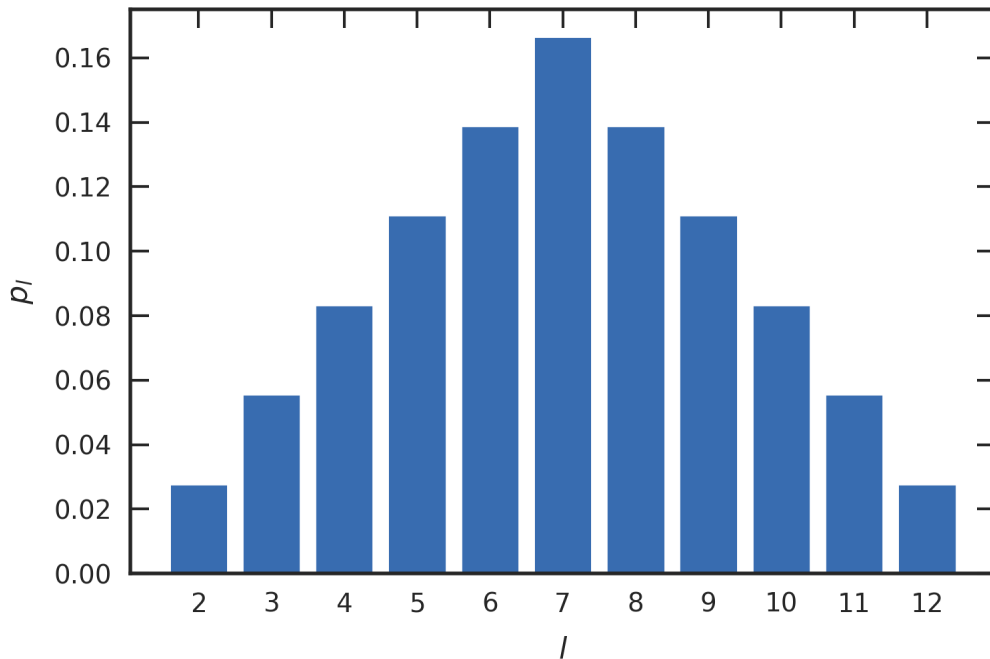


Figura 1.1: Massa de probabilidade da soma de dois dados comuns de 6 lados. A variável aleatória l representa a soma de dois dados comuns de 6 lados e p_l suas respectivas probabilidades. As barras em azul foram construídas baseadas na Equação (1.4).

obter n resultados de interesse é:

$$P(l = 1, 2, 3, \dots, n) = \sum_{l=1}^n p_l. \quad (1.7)$$

No caso de uma variável aleatória contínua, em vez de perguntar a probabilidade de um valor ocorrer, é mais adequado questionar a probabilidade de certa vizinhança em torno de um valor ocorrer [56]. Isso está associado à troca do somatório pela integral. Sendo x uma variável aleatória contínua $x \in \mathbb{R}$, a probabilidade de x ter seu resultado dentro de um intervalo $[a, b)$ de possibilidades é:

$$P(a \leq x < b) = \int_a^b \rho(x) dx. \quad (1.8)$$

Por exemplo, suponha que a variável x represente o tempo em anos da vida útil de um componente eletrônico. Não seria natural um aparelho eletrônico apresentar defeito nos primeiros dias de uso, assim como faz pouco sentido ele ficar operante por muitos e muitos anos. Uma distribuição de probabilidade candidata a modelar esse tipo de problema é a log-normal [57] (iremos abordar detalhadamente a distribuição log-normal neste capítulo mais adiante). A Figura 1.2 mostra a distribuição log-normal de parâmetros $\mu = 2$ e $\sigma =$

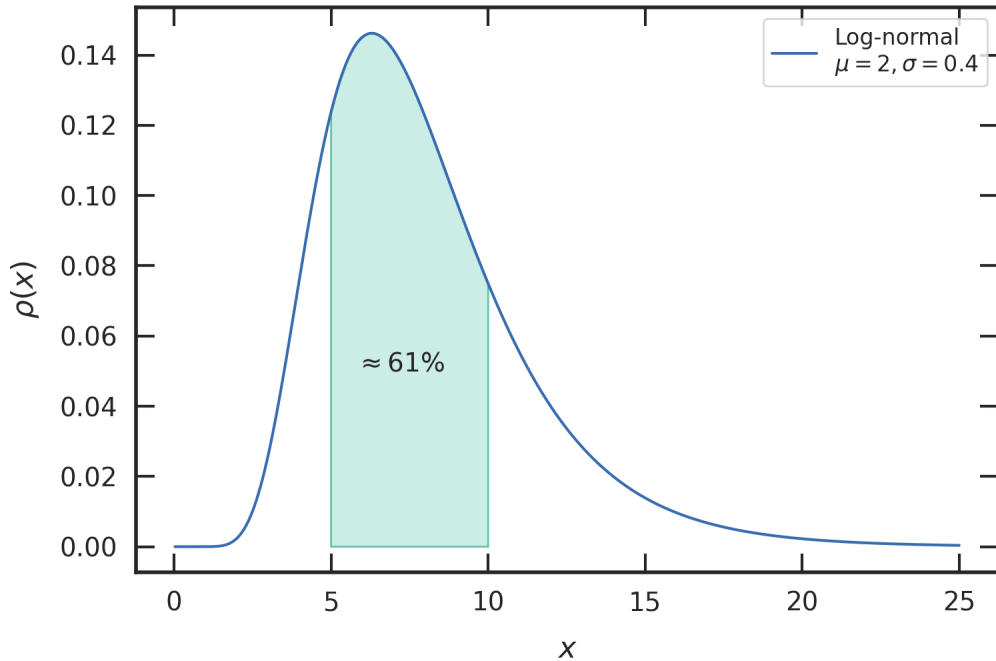


Figura 1.2: Função densidade de probabilidade de uma distribuição log-normal com parâmetros $\mu = 2$ e $\sigma = 0,4$. Consideramos a vida útil de um componente eletrônico (x , medida em anos) modelada por uma log-normal. A área sob a curva pintada em verde é a integral no intervalo $[5, 10]$ e representa a probabilidade do componente eletrônico apresentar defeito neste intervalo de tempo.

0,4. No caso de x , não é possível perguntar a probabilidade de um componente eletrônico estragar com exatamente 7 anos de uso, pois a integral de um ponto é sempre nula. Por isso, na Figura 1.2, baseado na área sob a curva (pintada em verde) percebemos que 61% dos componentes eletrônicos são danificados entre 5 e 10 anos de uso.

1.2 Medidas de centralidade

A distribuição de probabilidade é o objeto matemático com a maior informação sobre o sistema, e a partir dela várias outras grandezas podem ser obtidas. Nessa direção, as medidas de centralidade muitas vezes oferecem informações valiosas sobre a natureza do problema, sendo que as mais conhecidas são a média, a mediana e a moda. A média ($\langle l \rangle$ ou $\langle x \rangle$), também chamada de esperança matemática, é definida como

$$\langle l \rangle = \sum_l l p_l \quad (1.9)$$

para o caso discreto e

$$\langle x \rangle = \int_{-\infty}^{\infty} x \rho(x) dx \quad (1.10)$$

para o caso contínuo. A mediana ($l_{1/2}$ ou $x_{1/2}$) divide os dados em duas partes iguais, de modo que existe metade da probabilidade de obter um valor menor (maior) que a mediana. Matematicamente a mediana pode ser obtida calculando

$$\sum_l^{l_{1/2}} p_l = \frac{1}{2} \quad (1.11)$$

para o caso discreto e

$$\int_{-\infty}^{x_{1/2}} \rho(x) dx = \frac{1}{2} \quad (1.12)$$

no contínuo. A moda é o dado com maior frequência, facilmente evidenciado no máximo da função de probabilidade. Motivado pela moda pode-se classificar as distribuições de probabilidade, como: unimodal, com um máximo; bimodal, com dois máximos; e é comum utilizar multimodal para distribuições com várias modas. Considerando uma variável aleatória discreta, a moda da distribuição ou as modas da distribuição são os valores de l que apresentam o(s) maior(es) valor(es) de p_l . Já para variável aleatória contínua, a moda (x_{moda}) pode ser obtida resolvendo a seguinte equação:

$$\left(\frac{d\rho}{dx} \right)_{x=x_{\text{moda}}} = 0. \quad (1.13)$$

Na Figura 1.3, mostramos três exemplos de distribuições de probabilidade demarcando as medidas de centralidade citadas.

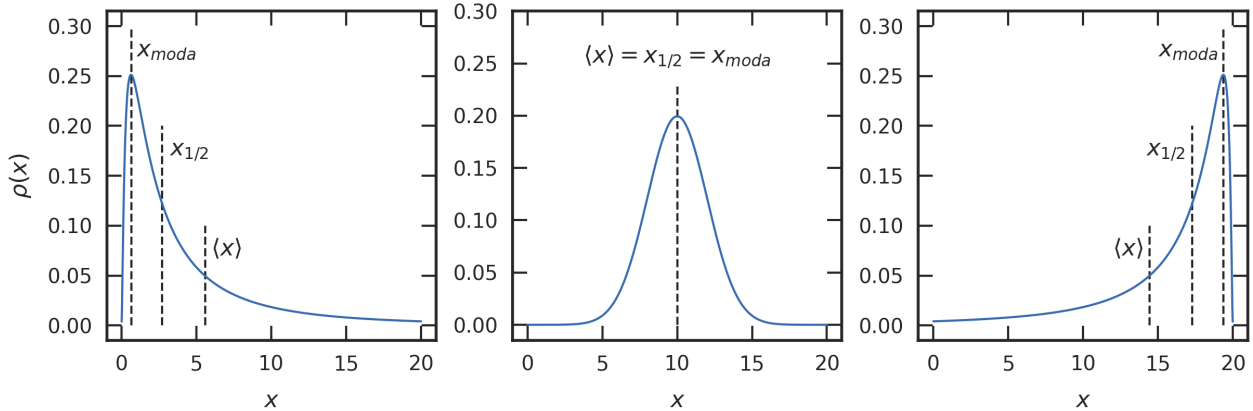


Figura 1.3: Exemplos de distribuições de probabilidades sinalizando as posições da média, mediana e moda. Consideramos três distribuições de probabilidade representadas pelas linhas contínuas azuis. Em cada painel marcamos as posições da média, mediana e moda por linhas pretas verticais e tracejadas.

1.3 Momentos da distribuição e medida de dispersão

Além da média ser uma medida de centralidade, ela é encarada também como um operador matemático dentro da teoria de probabilidade. Para entender melhor isso, precisamos definir a operação média. Seja $f(x)$ uma função qualquer da variável aleatória x , a média dessa função é dada por:

$$\langle f(x) \rangle = \int_{-\infty}^{\infty} f(x) \rho(x) dx. \quad (1.14)$$

Dessa maneira, se $f(x) = x$ recuperamos a definição da média de x , Equação (1.10). Observando essa possibilidade de generalizar a média, tem-se de forma natural na teoria de probabilidade a definição de momentos. Por definição, o n -ésimo momento de uma distribuição de probabilidade $\rho(x)$ é dado por:

$$\langle x^n \rangle = \int_{-\infty}^{\infty} x^n \rho(x) dx. \quad (1.15)$$

Assim, o n -ésimo momento pode ser obtido considerando a média da função $f(x) = x^n$. Fica claro que o primeiro momento ($n = 1$) de uma distribuição de probabilidade é a própria média. Todas as definições apresentadas nesse parágrafo podem ser consideradas para o caso discreto, basta trocar a integral pelo somatório e supor uma função, $f(l)$, dependente da variável aleatória discreta l .

Além das medidas de centralidade, diversas grandezas podem ser definidas para resumir a distribuição de probabilidade. Uma das mais conhecidas é o desvio padrão, comumente

representado pela letra σ . Conectado com o desvio padrão, outra medida bastante utilizada é a variância σ^2 , que nada mais é que o desvio padrão elevado ao quadrado. Aproveitando a operação de média, a variância pode ser definida como:

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2. \quad (1.16)$$

A variância representa uma medida do grau de dispersão médio dos dados em relação à média. Quanto mais disperso e distante da média, maior a variância. Por isso, a variância e o desvio padrão fazem parte do grupo de medidas de dispersão. Vale dizer que é mais comum utilizar o desvio padrão pois ele apresenta a mesma dimensão da variável aleatória, enquanto a variância tem a dimensão da variável ao quadrado.

1.4 Assimetria e curtose

Em relação à forma da dispersão, outras duas medidas podem ajudar a caracterizar uma distribuição, a assimetria (S) e a curtose (K). O coeficiente de assimetria, como o próprio nome diz, fornece uma medida da assimetria da distribuição de probabilidade e ele é definido como:

$$S = \frac{\langle (x - \langle x \rangle)^3 \rangle}{\sigma^3}. \quad (1.17)$$

Pensando graficamente, para $S > 0$, a cauda direita é mais “longa” que a cauda esquerda; para $S < 0$, o contrário é verdadeiro e a cauda esquerda é mais “longa” que a cauda direita; e para $S = 0$ a distribuição de probabilidade é simétrica em relação à média. A Figura 1.4 mostra exemplos de distribuições com $S > 0$, $S = 0$ e $S < 0$. Resgatando a Figura 1.3, percebemos que, no caso unimodal, para $S < 0$, temos $x_{\text{moda}} \geq x_{1/2} \geq \langle x \rangle$, para $S > 0$, $x_{\text{moda}} \leq x_{1/2} \leq \langle x \rangle$ e para $S = 0$, temos $\langle x \rangle = x_{1/2} = x_{\text{moda}}$. Entretanto, é importante ressaltar que embora a relação entre o coeficiente de assimetria e as medidas de centralidade pareça óbvia, sua veracidade está condicionada à distribuição de probabilidade satisfazer a desigualdade de Van Zwet [58], isto é, $\rho(x_{1/2} - x) + \rho(x_{1/2} + x) \geq 1$, para todo x .

Assim como a assimetria, a curtose descreve um aspecto particular de uma distribuição de probabilidade. Em teoria de probabilidade e estatística, a curtose é uma medida da extensão das caudas da distribuição de probabilidade de uma variável aleatória real. A medida padrão da curtose (K), proposta por Karl Pearson, é uma versão padronizada¹ do quarto momento da distribuição, isto é,

$$K = \frac{\langle (x - \langle x \rangle)^4 \rangle}{\sigma^4}. \quad (1.18)$$

A curtose está associada à probabilidade localizada ao centro (região distante da média

¹De maneira mais geral, o n -ésimo momento padronizado de uma distribuição é definido como $\left\langle \left(\frac{x - \langle x \rangle}{\sigma} \right)^n \right\rangle$.

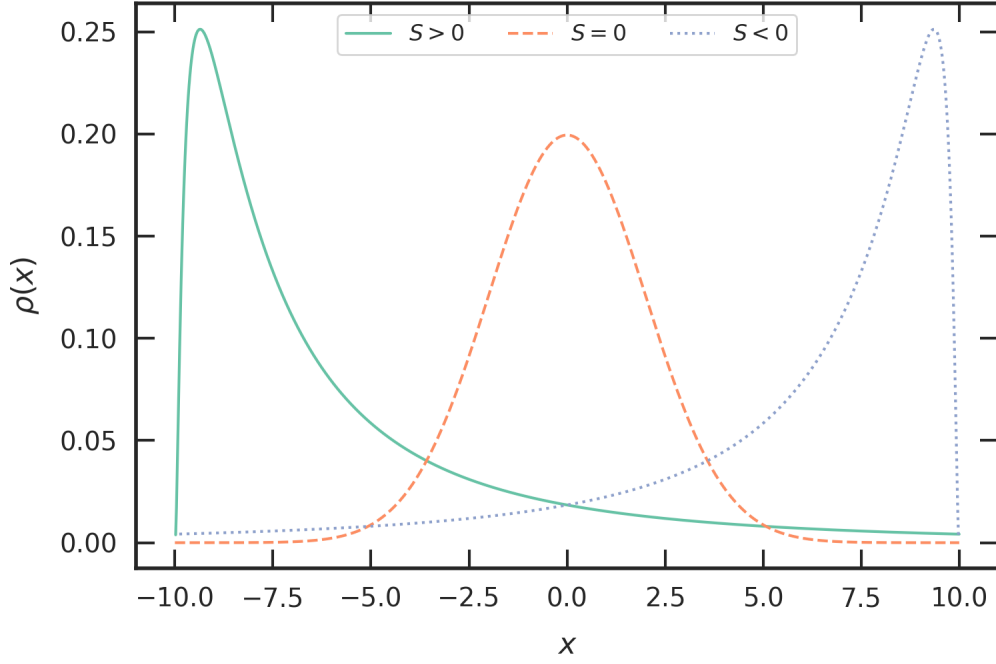


Figura 1.4: Exemplos de distribuições de probabilidades com diferentes coeficientes de assimetria. Representamos três distribuições de probabilidades, uma com $S > 0$ (verde), outra com $S = 0$ (laranja) e a última com $S < 0$ (azul).

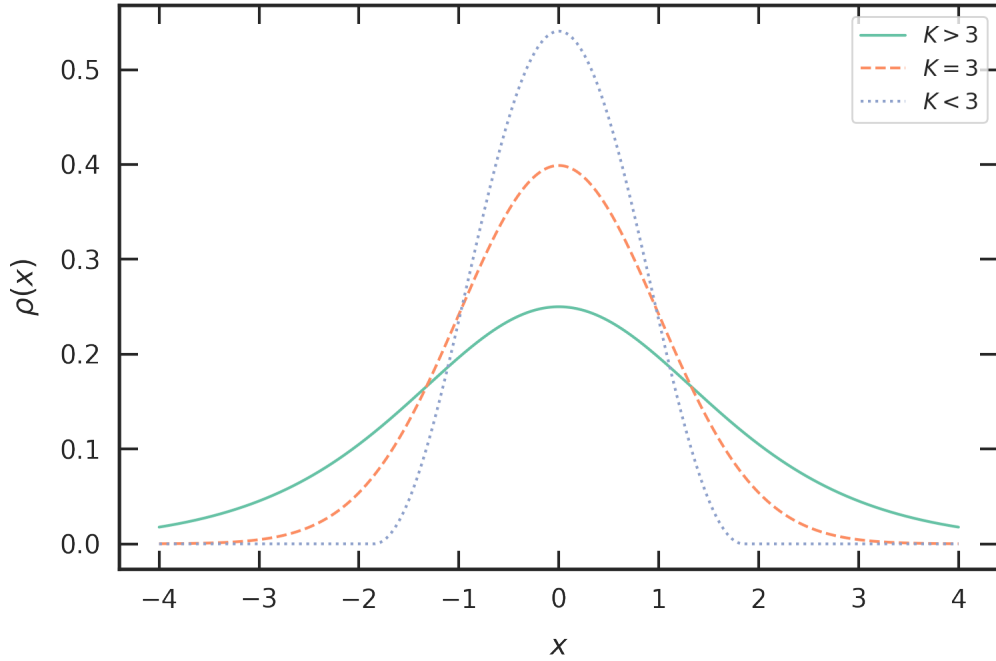


Figura 1.5: Uma amostra de distribuições de probabilidade com diferentes valores de curtose. Representamos uma distribuição leptocúrtica (verde), mesocúrtica (laranja) e platicúrtica (azul).

até um desvio padrão) e às caudas (localizadas a mais de três desvios padrões da média) de uma distribuição [59]. Mais claramente, desvios pequenos da média, próximos ao centro de uma distribuição, contribuem muito pouco para a curtose, enquanto valores nas caudas

contribuem significativamente para essa estatística. Existem três formas de classificar uma distribuição usando a curtose: platicúrtica ($K < 3$), mesocúrtica ($K = 3$) e leptocúrtica ($K > 3$). A distribuição normal é o principal exemplo de distribuição mesocúrtica, de modo que distribuições platicúrticas têm caudas mais curtas comparadas a uma gaussiana, enquanto distribuições leptocúrticas tem caudas mais longas. Sendo assim, distribuições platicúrticas (leptocúrticas), produzem menos (mais) *outliers*² que a distribuição normal. Na Figura 1.5, exemplificamos cada caso.

1.5 Modelos probabilísticos para múltiplas variáveis

Todos os conceitos apresentados até aqui referem-se a distribuições de probabilidade de uma única variável aleatória, porém é possível generalizar todo o formalismo probabilístico para o caso de duas ou mais variáveis. Nessa direção, a distribuição conjunta é uma extensão do conceito de probabilidade para várias variáveis. Sendo x e y variáveis aleatórias, então a função densidade de probabilidade $\rho(x, y)$ é graficamente uma superfície. Além disso, para os intervalos $[a, b]$ em x e $[c, d]$ em y , tem-se a probabilidade decorrente da integração:

$$P(a \leq x < b, c \leq y < d) = \int_c^d \int_a^b \rho(x, y) dx dy. \quad (1.19)$$

A função $\rho(x, y)$ é chamada densidade de probabilidade conjunta e deve ser não negativa e normalizável, ou seja,

$$\rho(x, y) \geq 0, \quad (1.20)$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(x, y) dx dy = 1. \quad (1.21)$$

A função de probabilidade conjunta é o objeto de maior informação do problema, pois ela acumula todas as informações estatísticas de x e y . Note que para calcular a probabilidade, a integração requer uma vizinhança em torno do par ordenado (x, y) . Porém, existe uma forma de analisar separadamente o comportamento de cada variável. Nesse caso, $\rho_X(x)$ corresponde à probabilidade da variável x independente do resultado de y , e $\rho_Y(y)$ à probabilidade da variável y independente do resultado de x :

$$\rho_X(x) = \int_{-\infty}^{\infty} \rho(x, y) dy, \quad (1.22)$$

$$\rho_Y(y) = \int_{-\infty}^{\infty} \rho(x, y) dx. \quad (1.23)$$

As funções $\rho_X(x)$ e $\rho_Y(y)$ são denominadas densidade de probabilidade reduzida.

² *Outliers* são valores muito extremos de um conjunto de dados. Os outliers podem indicar dados válidos que são raros ou exceções, ou podem sinalizar erros que precisam ser investigados e possivelmente corrigidos.

Em certos casos, a distribuição reduzida permite entender o comportamento de uma variável independentemente da outra, mas muitas vezes as variáveis são dependentes entre si de alguma forma, ou seja, o resultado de uma variável afeta o resultado da outra variável, ou até mesmo pode haver uma interação mútua. Nesse contexto, surge naturalmente a medida de dependência entre as variáveis x e y , denominada covariância (C), definida por:

$$C = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle. \quad (1.24)$$

Matematicamente, a covariância mede a relação linear entre as duas variáveis e pode variar de $(-\infty, \infty)$. Valores de covariância positivos indicam que valores acima da média de uma variável estão associados a valores acima da média da outra variável. Valores de covariância negativos indicam que valores acima da média de uma variável estão associados com valores abaixo da média da outra variável. Covariância nula significa tipicamente que as variáveis são independentes uma da outra. A Figura 1.6 ilustra os três casos de covariância expostos.

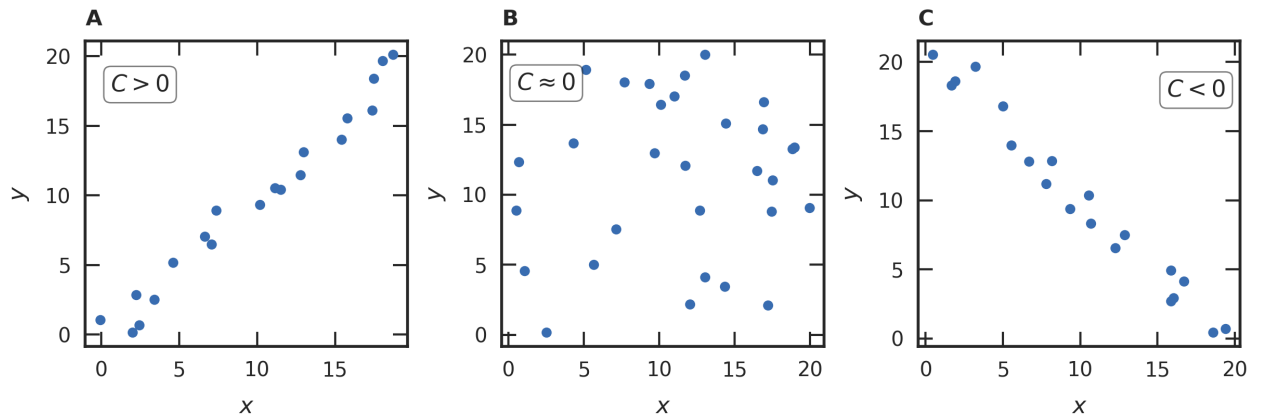


Figura 1.6: Exemplos de dados com diferentes valores de covariância. (A) Representamos o caso de covariância positiva, (B) covariância nula e (C) covariância negativa.

Voltando à distribuição conjunta de duas variáveis, caso seja necessário uma mudança de variáveis, esse procedimento deve obedecer as regras de cálculo de integração. Sendo u e v funções dependentes de x e y , dada as transformações $u = f_1(x, y)$ e $v = f_2(x, y)$, a distribuição de probabilidade conjunta de u e v é dada por:

$$\rho_{U,V}(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(u - f_1(x, y)) \delta(v - f_2(x, y)) \rho_{X,Y}(x, y) dx dy. \quad (1.25)$$

Seguindo as regras de cálculo, se a mudança das variáveis x e y para as variáveis u e v for unívoca, tem-se:

$$\rho_{U,V}(u, v) = \rho_{X,Y}(x(u, v), y(u, v)) |J|, \quad (1.26)$$

em que $|J|$ é o determinante da matriz jacobiana da transformação $(x, y) \rightarrow (u, v)$. Vale dizer que, para mais de duas variáveis, basta usar a extensão natural das regras de cálculo.

A densidade de probabilidade conjunta terá que continuar sendo não negativa e satisfazer a condição de normalização.

No contexto de análise de múltiplas variáveis aleatórias, o caso das variáveis independentes merece destaque. No mundo real, alguns sistemas são bem representados por uma infinidade de componentes independentes. Mesmo sendo independentes, a partir do resultado de cada uma das componentes pode surgir um padrão global. Saber analisar cada componente separadamente para obter o comportamento do sistema todo é de grande importância [2]. Como já foi dito, variáveis aleatórias são ditas independentes quando o resultado de uma não altera diretamente o resultado das demais. Por exemplo, jogar um dado e girar uma moeda são eventos independentes e o valor que sair no dado não alterará o giro da moeda e vice-versa. Seja $x_1, x_2, x_3, \dots, x_n$ variáveis aleatórias independentes com distribuições $\rho_1(x_1), \rho_2(x_2), \rho_3(x_3), \dots, \rho_n(x_n)$, respectivamente. A independência das variáveis implica:

$$\rho(x_1, x_2, x_3, \dots, x_n) = \rho_1(x_1)\rho_2(x_2)\rho_3(x_3) \dots \rho_n(x_n), \quad (1.27)$$

em que $\rho(x_1, x_2, x_3, \dots, x_n)$ é a distribuição conjunta das variáveis $x_1, x_2, x_3, \dots, x_n$. Por exemplo, isso leva ao fato de que o momento do produto das variáveis é igual ao produto dos momentos de cada variável:

$$\langle x_1 x_2 x_3 \dots x_n \rangle = \langle x_1 \rangle \langle x_2 \rangle \langle x_3 \rangle \dots \langle x_n \rangle. \quad (1.28)$$

Esse tipo de resultado mostra que a covariância é nula para variáveis independentes, pois como já foi dito na definição de covariância, ela fornece uma medida do grau de dependência entre as variáveis.

1.6 Exemplos de distribuições

Como deve estar claro, distribuições de probabilidade são recorrentes na natureza e são úteis para descrever fenômenos. Portanto, a familiaridade com a origem e a forma matemática delas é importante para o estudo de sistemas complexos. Nesta seção, apresentamos algumas distribuições recorrentes na natureza.

1.6.1 Distribuição uniforme

A distribuição uniforme modela eventos com mesma probabilidade de ocorrer. Intuitivamente, essa distribuição vem à mente quando pensamos nos exemplos mais simples de probabilidade, como o lançamento de um dado de 6 lados ou de uma moeda. A distribuição uniforme pode ser discreta ou contínua, e além disso, ela pode variar dentro de qualquer intervalo finito dos números reais. Formalmente, a distribuição uniforme depende de dois

parâmetros: o começo do intervalo (a) e o fim do intervalo (b) de eventos possíveis da variável aleatória. Dado o intervalo $[a, b]$, a distribuição uniforme é dada por

$$\rho(x) = \begin{cases} \frac{1}{b-a}, & \text{se } a \leq x \leq b \\ 0, & \text{caso contrário} \end{cases}, \quad (1.29)$$

e, portanto, a média é $(a + b)/2$ e a variância é $(b - a)^2/12$. Na Figura 1.7, mostramos exemplos de distribuições uniformes com diferentes parâmetros. Note que quanto maior for o intervalo $[a, b]$ menor é a amplitude da densidade de probabilidade.

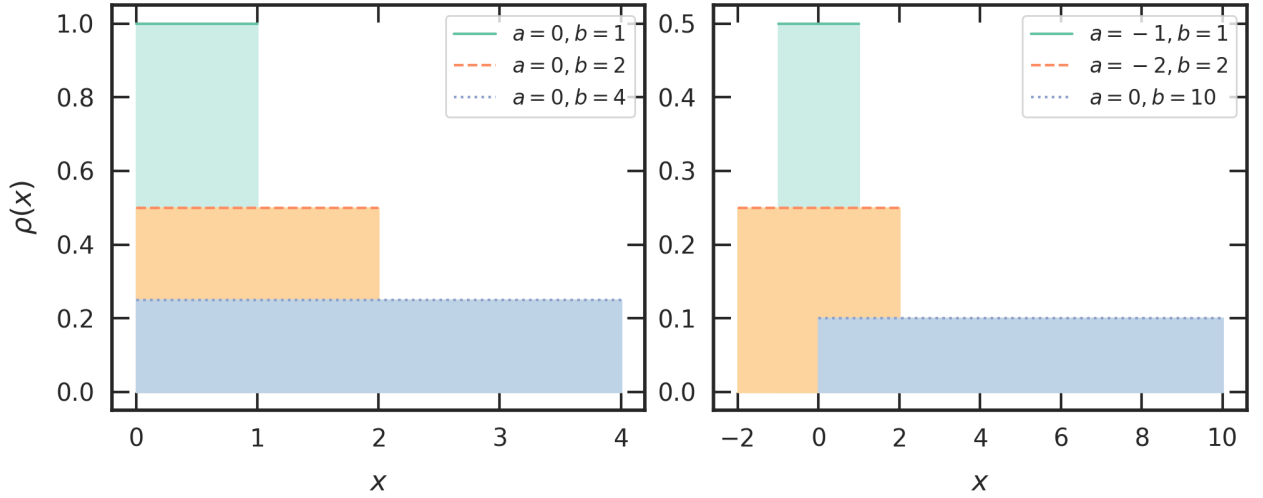


Figura 1.7: Distribuição uniforme. Apresentamos algumas distribuições uniformes variando os parâmetros a e b .

1.6.2 Distribuição normal

Se pudéssemos eleger a distribuição mais famosa e importante para a teoria de probabilidade, a distribuição normal com certeza seria escolhida. A distribuição normal, também conhecida como distribuição gaussiana, é uma distribuição de probabilidade contínua e simétrica em relação à média ($S = 0$). Sua forma gráfica apresenta a forma de um sino, mostrando que os eventos próximos da média são mais frequentes. Matematicamente, dada a média (μ) e a variância (σ^2), a distribuição normal é definida como:

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \quad (1.30)$$

Lembramos que a classificação das distribuições pela curtose definida na Equação (1.18) é em referência à normal padronizada ($K = 3$ quando $\mu = 0$ e $\sigma^2 = 1$). Além disso, como já mostrado na Figura 1.3, a distribuição normal apresenta a média, mediana e moda iguais ($\langle x \rangle = x_{1/2} = x_{\text{moda}}$). Na Figura 1.8, apresentamos alguns exemplos de distribuição normal

com diferentes parâmetros. Vale dizer que, se uma variável x é distribuída normalmente com média μ e variância σ , a variável $z = (x - \mu)/\sigma$ terá uma distribuição normal de média 0 e variância 1, também conhecida como normal padronizada.

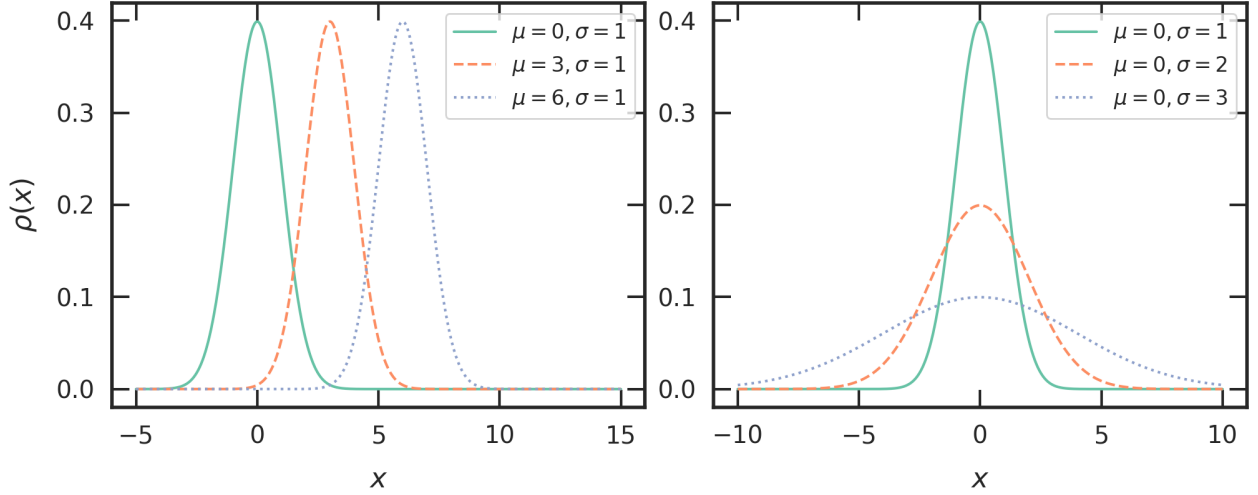


Figura 1.8: Distribuição normal. Ilustramos distribuições normais com diferentes valores de μ e σ .

1.6.3 Distribuição log-normal

A log-normal é uma distribuição de probabilidade contínua cuja variável assume apenas valores positivos. Além disso, essa distribuição é frequentemente usada para modelar variáveis que apresentam crescimento multiplicativo, como preços de ações, tamanhos de partículas, níveis de poluição e tempos de sobrevivência. Vale dizer que, se uma variável é descrita por uma distribuição normal, então a exponencial dessa variável é descrita por uma log-normal. Mais adiante o Teorema central do limite deixará em evidência a grande importância da distribuição normal, e como consequência direta disso, a importância da log-normal. Considerando os parâmetros μ e σ , a distribuição log-normal é

$$\rho(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}, \quad (1.31)$$

sua média é $e^{\mu + \frac{\sigma^2}{2}}$ e sua variância é $(e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$. Na Figura 1.9, mostramos exemplos de distribuições log-normais para diferentes valores de μ e σ . Na teoria de probabilidade, a log-normal é classificada como uma distribuição de cauda pesada, essa nomenclatura devendo-se ao fato de sua cauda convergir à zero mais lentamente que outras (por exemplo, normal ou exponencial).

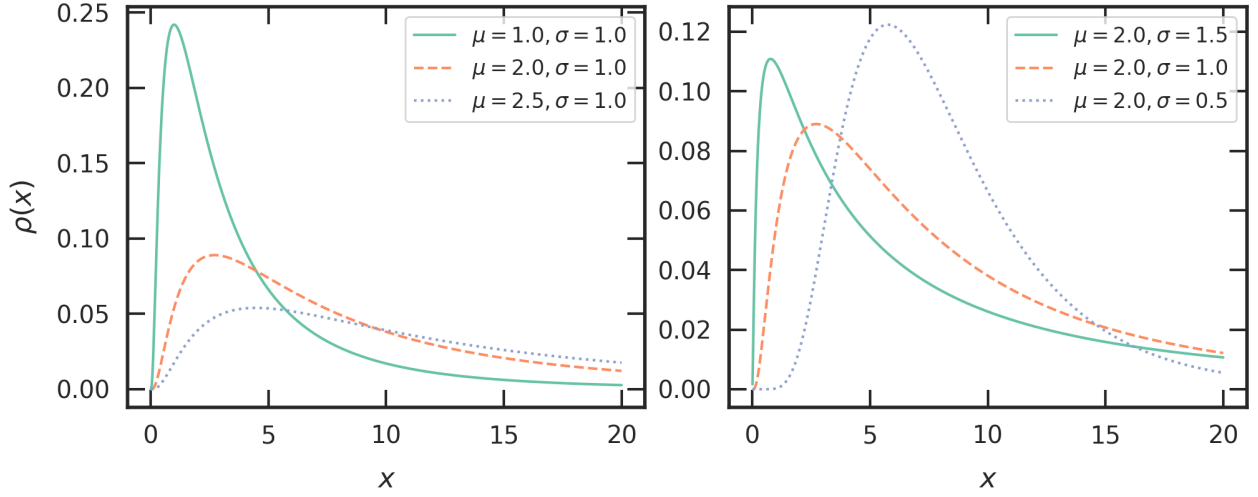


Figura 1.9: Distribuição log-normal. Alguns exemplos da distribuição log-normal variando seus parâmetros.

1.6.4 Distribuição exponencial

A distribuição exponencial é muito útil para modelar uma sequência de eventos de certa raridade e com mesma chance de sucesso/falha. Dito de outra forma, um processo no qual os eventos ocorrem de forma contínua e independente a uma taxa média constante. Para obter alguma intuição sobre essa interpretação da distribuição exponencial, imagine alguém esperando algum acontecimento. Por exemplo, imagine um vendedor esperando o próximo cliente. Em cada segundo, a probabilidade de um novo cliente entrar na loja é muito pequena. Porém, a cada segundo existe uma chance de alguém passar pela loja e querer entrar. Supondo que o interesse de cada pessoa querer entrar na loja é aproximadamente igual, então eventualmente ao passar várias pessoas por ali alguém vai entrar na loja. Pensando matematicamente, a distribuição exponencial é dada por

$$\rho(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{se } x \geq 0 \\ 0, & \text{se } x < 0 \end{cases}, \quad (1.32)$$

em que λ é positivo. A média da distribuição exponencial é $1/\lambda$ e a variância é $1/\lambda^2$. A propriedade chave da exponencial é o fato dela não ter memória. Para entender essa propriedade qualitativamente, imagine que cada pessoa toma sua decisão de entrar ou não na loja independentemente das outras. Então mesmo que muitas pessoas tenham optado por não entrar, a partir disso é como começar tudo de novo e a chance da pessoa seguinte entrar é a mesma. A Figura 1.10 mostra a distribuição exponencial para $\lambda = 5, 3, 2$ e 1 .

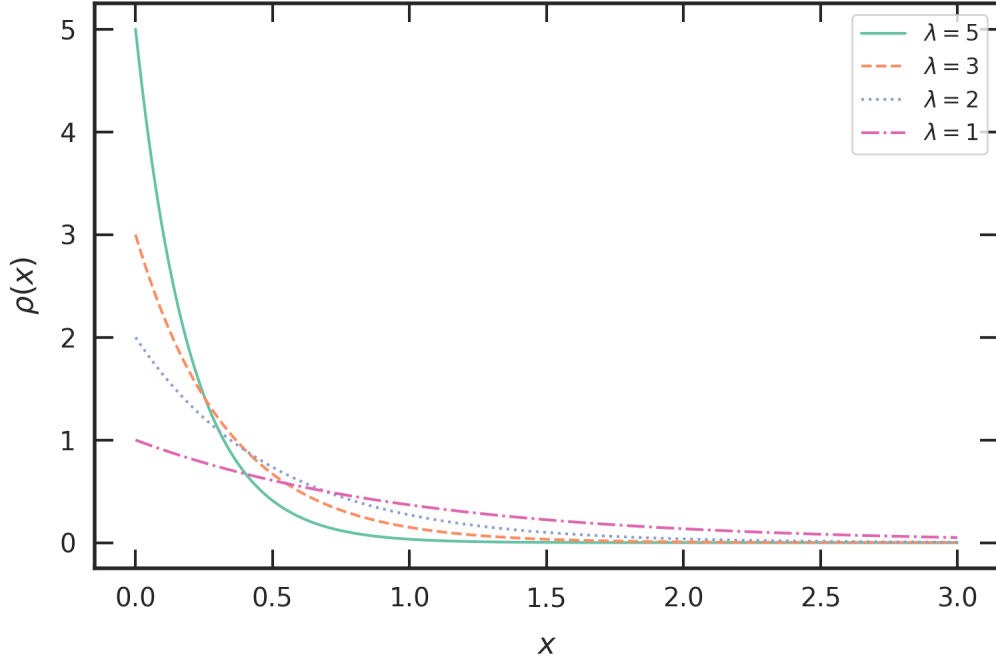


Figura 1.10: Distribuição exponencial. Representação da distribuição exponencial para os valores de $\lambda = 5, 3, 2$ e 1 .

1.7 Lei dos grandes números

A Lei dos grandes números formaliza matematicamente o núcleo da teoria de probabilidade apresentado até aqui. Desde o início deste capítulo apresentamos a probabilidade como um resultado a priori, ou seja, não consideramos nenhum problema em que houve experimentação real. No exemplo da soma de dois dados baseamos nossa interpretação de probabilidade em um grande número de observações e dessa forma assumimos a priori que cada face do dado tem a mesma chance de cair para cima. Por outro lado, em uma situação real na qual um dado é lançado um número finito de vezes, nada garante que todas as seis faces caiam para cima exatamente na mesma proporção. Da mesma forma, deve acontecer com o lançamento de uma moeda e se uma moeda for lançada 100 vezes, nada garante que ela cairá exatamente 50 vezes com a “cara” para cima.

De uma maneira geral, a Lei dos grandes números sustenta a interpretação frequentista de probabilidade, em que a probabilidade é uma idealização decorrente de um número infinito de observações [56]. Sejam $x_1, x_2, x_3, \dots, x_n$ variáveis aleatórias independentes e identicamente distribuídas. A partir disso, a Lei dos grandes números afirma que:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n x_j \rightarrow a, \quad (1.33)$$

sendo $a = \langle x_j \rangle$. A única condição para a validade desse teorema é que a integral $\int_{-\infty}^{\infty} |x| \rho(x) dx$ exista. Dessa forma, à medida que a amostra tenha um número reduzido de observações, as

flutuações e desvios podem aumentar. Vale dizer que na teoria de probabilidade e estatística, há uma distinção entre Lei fraca e Lei forte dos grandes números. A Lei forte oferece uma garantia mais forte, mais certa e mais rigorosa de convergência. Já para a Lei fraca, a média amostral converge para próximo do valor esperado, não diretamente para ele. Qualitativamente, a Lei fraca apresenta uma garantia mais fraca de que a média amostral se aproxima da média populacional. Não será discutido o rigor matemático da diferença entre essas duas Leis, pois o objetivo desta seção é apenas demonstrar que existe uma convergência (seja ela forte ou fraca) entre a média amostral e a média populacional quando o número de amostras é muito grande.

1.8 Teorema central do limite

O Teorema central do limite é um dos mais importantes resultados em probabilidade e estatística, sendo fundamental para a teoria de inferência estatística. Segundo esse teorema, à medida que aumenta o tamanho de uma amostra, a distribuição amostral da média tende a uma distribuição normal. Como a distribuição normal é unimodal e simétrica, então ao nos distanciarmos da média, a probabilidade de ocorrência diminui. Isso permite inferir a média e o desvio padrão amostral sem precisar realizar a difícil tarefa de amostrar a população toda [60]. De forma prática, podemos pensar em uma variável aleatória qualquer cuja distribuição de probabilidade é desconhecida. Neste caso, se fizermos múltiplas amostragens, com um número adequado de observações é possível estimar a média dela. Obviamente que quanto maior e mais amostras forem coletadas, melhor será a estimativa.

Deve estar claro o quão geral é o Teorema central do limite e como ele pode aparecer em diversos contextos. Matematicamente, sejam $x_1, x_2, x_3, \dots, x_n$ variáveis aleatórias independentes e identicamente distribuídas com média a e variância b . Suponha uma variável z desse conjunto de variáveis x_j , dada por

$$z = \frac{1}{\sqrt{nb}} \left(\sum_{j=1}^n x_j - na \right). \quad (1.34)$$

Assim, a distribuição de probabilidade da variável z é uma distribuição normal padronizada quando $n \rightarrow \infty$:

$$\rho(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \quad (1.35)$$

Para a validade desse teorema, é necessário garantir a existência da média a , da variância b e ausência de correlações entre as variáveis x_j . Note que em essência a distribuição normal surge em decorrência da soma de múltiplas variáveis, assim se pensarmos em um processo multiplicativo surge a log-normal. Suponha a variável $z = x_1 x_2 x_3 \dots x_n$, então $\ln z = \ln x_1 + \ln x_2 + \ln x_3 + \dots + \ln x_n$. Considerando $y_j = \ln x_j$, a variável $\ln z$ terá distri-

buição normal e a variável z terá distribuição log-normal. Uma demonstração matemática detalhada do Teorema central do limite está disponível no apêndice A. Além disso mostramos alguns exemplos do quão rápido algumas amostras convergem para a distribuição normal dependendo do seu tamanho n .

1.9 Estimador de máxima verossimilhança

Em estatística, um estimador é uma regra para calcular uma determinada quantidade baseada em dados observados. Existe todo um campo de estudo em estatística que aborda os mais diversos tipos de estimadores e dessa maneira seria muito extensivo estudar esse tema em sua completude [61]. Assim sendo, vamos nos restringir ao estimador mais popular para determinar os parâmetros de uma distribuição de probabilidade, o estimador de máxima verossimilhança. Antes de abordar diretamente o método de estimativa, devemos esclarecer como funciona na prática a inferência estatística. Grande parte das vezes, o pesquisador não sabe a distribuição de probabilidade de origem dos dados, pois sistemas reais podem ter um alto grau de complexidade. Nesses casos, não existe nenhum procedimento para escolher uma distribuição de probabilidade que possa ser testada. Então, o pesquisador precisa ter uma certa intuição para propor uma hipótese e testar se essa hipótese satisfaz as regras de significância estatística. Esses procedimentos para verificação da qualidade do ajuste serão comentados em seções adiante. Portanto, nesta seção vamos considerar que exista um dado com um modelo já proposto e apenas queremos estimar os parâmetros do modelo a partir dos dados.

A máxima verossimilhança estima os parâmetros de um modelo baseado na maximização da função de verossimilhança. Até aqui representamos a distribuição de probabilidade por ρ e ela sempre representou uma função matemática de uma dada variável aleatória, por exemplo x . Além disso, a distribuição ρ sempre dependia de um parâmetro ou um conjunto de parâmetros fixos. Nesse contexto, a função de verossimilhança é definida como a função da distribuição de probabilidade cujas variáveis independentes são os parâmetros, dado um conjunto particular de dados observados [62]. Explicando em outras palavras, suponha uma distribuição de probabilidade hipotética $\rho(x; \theta)$, em que x é a variável aleatória e θ é o parâmetro da distribuição. Considere ainda que exista um conjunto de m observações, $X = \{x_1, x_2, \dots, x_m\}$. Nesse caso, a função de verossimilhança será dada por

$$L(\theta; x) = \rho(x_1; \theta) \rho(x_2; \theta) \dots \rho(x_m; \theta) = \prod_{j=1}^m \rho(x_j; \theta), \quad (1.36)$$

na qual o parâmetro θ será tratado como a variável independente dado o conjunto de observações X . Em uma interpretação livre, a verossimilhança pode ser vista como a probabilidade não normalizada do valor de um parâmetro particular para um dado conjunto de dados [62].

Como a função de verossimilhança é definida como uma multiplicação de muitos termos, uma alternativa para os cálculos envolvendo $L(\theta; x)$ é tomar o logaritmo, ou seja, $\ln L(\theta; x)$. Por fim, o parâmetro θ é determinado resolvendo a equação

$$\frac{d \ln L}{d\theta} = 0, \quad (1.37)$$

que representa a maximização da função de verossimilhança. Vale dizer que para os casos de mais de um parâmetro, como um conjunto $\{\theta_1, \theta_2, \dots\}$, a única mudança no método de máxima verossimilhança é a Equação (1.37) se tornar um conjunto de derivadas parciais em relação às variáveis $\{\theta_1, \theta_2, \dots\}$.

Para fins ilustrativos, considere hipoteticamente um pesquisador interessado em estudar a altura média das pessoas. Suponha que foi coletada a altura de 200 pessoas e o pesquisador expôs a amostra na Tabela B.1 do apêndice B. Dessa maneira, vamos considerar que a variável x representa a altura das pessoas em metros. Nessa pesquisa, podemos propor a hipótese que a altura das pessoas siga uma distribuição normal, definida pela Equação (1.30), pois parece natural que muitas pessoas tenham a altura próxima da média e existam poucas pessoas muito altas ou muito baixas. Partindo da função de verossimilhança, Equação (1.36), teremos um produtório de 200 distribuições normais dependendo dos parâmetros μ e σ , ou seja, $L(\mu, \sigma; x)$. Como vimos, analiticamente, o máximo da função de verossimilhança pode ser acessado tomando a derivada e igualando a zero, como definido na Equação (1.37). Por outro lado, computacionalmente podemos gerar um intervalo de valores de μ e σ e substituir na função $L(\mu, \sigma; x)$ para determinar o máximo de forma numérica. A Figura 1.11A mostra

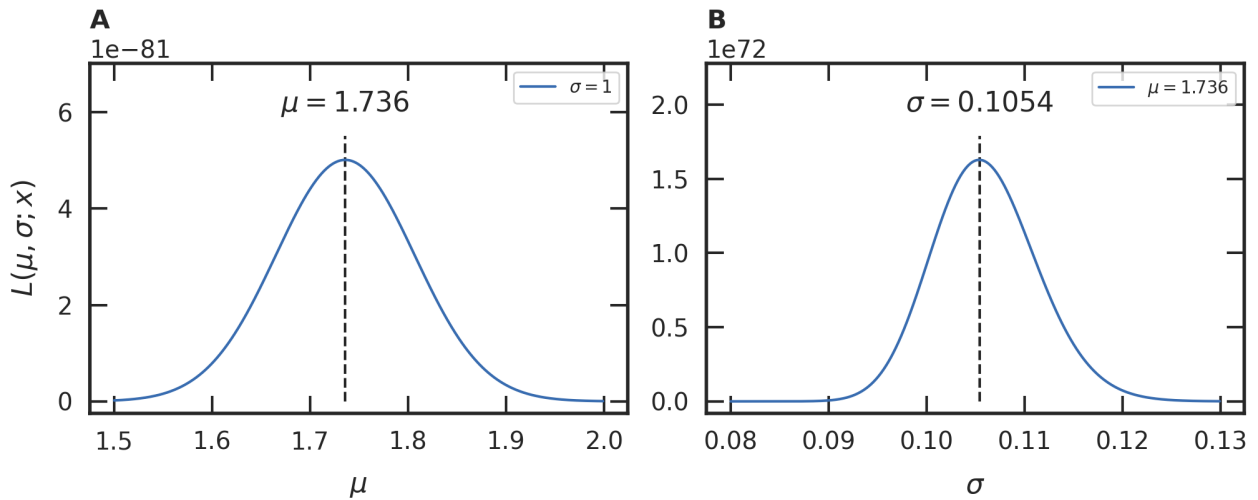


Figura 1.11: Análise da função de verossimilhança. Apresentamos a função de verossimilhança da distribuição normal para um intervalo de valores de (A) μ e (B) σ , utilizando os dados da Tabela B.1 do apêndice B. Fazendo uma varredura nos valores de μ em passos de 0,001, obtivemos o máximo da verossimilhança em $\mu = 1,736$, utilizando o mesmo procedimento para σ com passos de tamanho 0,0001 obtivemos $\sigma = 0,1054$.

o comportamento da função de verossimilhança no intervalo de 1,5 à 2,0 para μ , em que fixamos a variável σ em 1 (podemos fixar uma variável sem nos preocupar com a outra, pois elas são independentes). Nessa figura, fica claro que $L(\mu, \sigma; x)$ tem um máximo em torno de $\mu = 1,736$. Na Figura 1.11B, fixamos $\mu = 1,736$ e variamos σ dentro do intervalo de 0,08 até 0,13, e o máximo da verossimilhança se deu em $\sigma = 0,1054$. Na Figura 1.12, podemos conferir o histograma da amostra e a distribuição normal ajustada pela máxima verossimilhança.

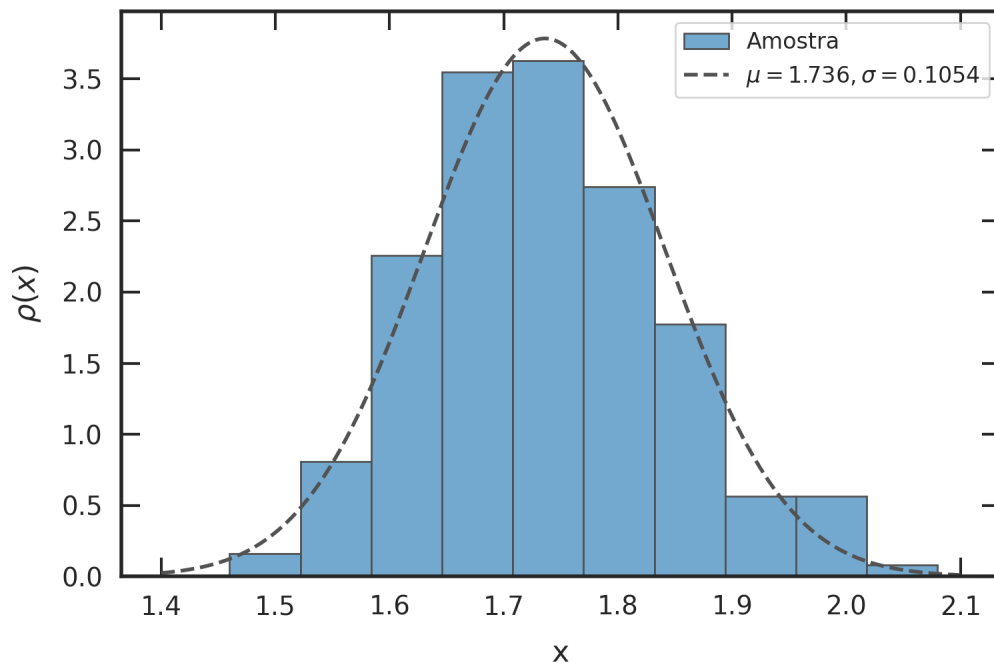


Figura 1.12: Ajuste da distribuição normal para uma amostra. Baseado nas estimativas feitas na Figura 1.11 ($\mu = 1,736$ e $\sigma = 0,1054$), apresentamos aqui o histograma dos dados e a distribuição normal ajustada.

1.10 *Bootstrap*

O intervalo de confiança (IC) é um tipo de estimativa por intervalo de um parâmetro desconhecido [63]. Esse intervalo depende do nível de confiança e pode variar de amostra para amostra. Porém, tomando um número razoável de amostras é possível construir um intervalo de confiança adequado para a variação da estimativa. Além disso, o nível de confiança pode ser fixado em um valor que passe credibilidade. Por definição, se é escolhido um nível de confiança de 99% para uma estimativa, o intervalo de confiança representa o intervalo em que 99% dos intervalos de confiança observados têm o valor real do parâmetro estimado. Deve estar claro que estimar o parâmetro como foi feito na seção anterior é apenas parte de uma investigação estatística rigorosa. O intervalo de confiança é fundamental para determinar a significância estatística e, portanto, aceitar ou rejeitar a hipótese proposta.

Na direção de determinar intervalos de confiança, uma técnica muito difundida é a reamostragem por *bootstrap* [64]. Em linhas gerais, o *bootstrap* pode ser empregado para atribuir medidas de precisão (viés, variância, intervalos de confiança, erro de previsão, etc.) para estimativas da amostra. Essa técnica permite estimar a distribuição amostral de quase todas as estatísticas usando métodos de reamostragem aleatória [65]. A Figura 1.13 mostra um esquema de como funciona o *bootstrap*. Na prática, seja $X = \{x_1, x_2, \dots, x_m\}$ um conjunto de observações, a ideia por trás do *bootstrapping* é reamostrar com reposição as observações da amostra X . Primeiramente, vamos considerar $X^* = \{x_1^*, x_2^*, \dots, x_m^*\}$ sendo uma reamostra de X , isto é, o elemento x_1^* representa o primeiro valor escolhido aleatoriamente de X . A seguir, é feito a reposição e então é escolhido o segundo valor aleatoriamente x_2^* . Seguimos assim sucessivamente até a reamostra X^* ter m elementos. Não é claro que X^* deve ter o mesmo número de elementos, m , que X , mas por convenção é comum fazer a reamostra com o mesmo tamanho da amostra. A partir da reamostra X^* , podemos calcular a estimativa de interesse que queremos saber de X e para obter o intervalo de confiança basta fazer esse processo de reamostragem e cálculo de estimativa um número grande de vezes. Dessa maneira, será gerada uma lista de quantidades da estimativa de interesse e será possível limitar um intervalo de confiança baseado no nível de confiança requerido pelo pesquisador. Não existe um número exato do quão grande deve ser o número de reamostras, mas, por pressuposto, um intervalo de confiança gerado por *bootstrap* é válido quando a distribuição de probabilidade do conjunto das quantidades estimadas segue uma distribuição normal.

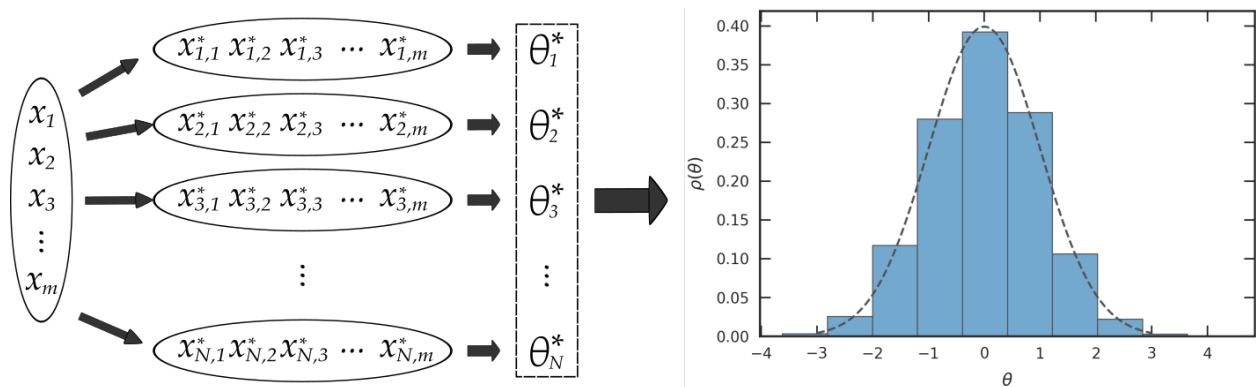


Figura 1.13: *Bootstrapping*. Um resumo esquemático da reamostragem por *bootstrap* para estimar o intervalo de confiança de um determinado parâmetro.

Partindo do exemplo da distribuição das alturas da seção 1.9, podemos estimar o intervalo de confiança dos parâmetros estimados μ e σ . Para isso geramos mil reamostras com o mesmo tamanho da amostra de alturas coletada e em cada uma delas estimamos o valor de μ^* e σ^* . Representamos as grandezas com o asterisco pois queremos enfatizar que a estimativa foi feita baseada nas reamostragens por *bootstrap*. Na Figura 1.14A, mostramos a distribuição do parâmetro estimado μ^* e destacamos (linha tracejada preta) o intervalo de confiança para um nível de confiança de 95%. Sabendo que $\mu^* = 1,722$ representa o valor 2,5% mais baixo

das estimativas e $\mu^* = 1,751$ o valor 2,5% mais alto das estimativas, podemos denotar nossa estimativa como $\mu = 1,736$ [95% IC: 1,721, 1,751]. Na Figura 1.14B, fizemos o procedimento análogo para o parâmetro estimado σ^* e obtivemos $\sigma = 0,1054$ [95% IC: 0,0946, 0,1152].

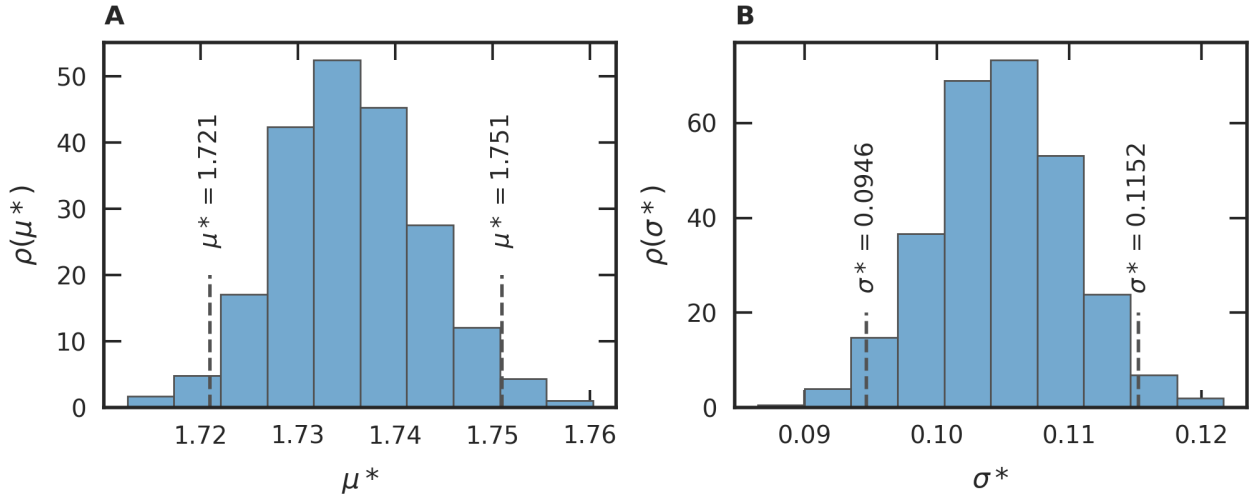


Figura 1.14: *Bootstrapping* das estimativas de μ e σ . (A) As barras azuis representam o histograma do parâmetro estimado μ^* , enquanto linhas pretas tracejadas representa o valor de μ^* dos 2,5% inferior ($\mu^* = 1,721$) e superior ($\mu^* = 1,751$). Analogamente, (B) mostra o histograma de σ^* (barras em azul) e as linhas pretas tracejadas evidenciam os valores 2,5% inferior ($\sigma^* = 0,0946$) e superior ($\sigma^* = 0,1152$).

1.11 Método dos mínimos quadrados

O método dos mínimos quadrados é a abordagem padrão na análise de regressão, ele determina o melhor ajuste para um conjunto de dados através da minimização dos resíduos [66]. Até a seção anterior, discutimos um caminho para investigar a distribuição de probabilidade de uma amostra. Nesta seção, vamos discutir uma técnica que permite quantificar e inferir a relação entre duas variáveis, um procedimento chamado de regressão. O método de mínimos quadrados é uma forma de regressão linear, isso quer dizer que o modelo deve ser linear nos parâmetros. Se $X = \{x_1, x_2, \dots, x_m\}$ e $Y = \{y_1, y_2, \dots, y_m\}$ são variáveis aleatórias, o modelo terá a forma:

$$Y = \alpha + \beta X + \epsilon, \quad (1.38)$$

em que α é um parâmetro do modelo chamado de constante (porque não depende de x), β é o coeficiente da variável x , e ϵ é o erro (que representa a variação de y que não é explicada pela variável x). A partir desse contexto, o objetivo do método dos mínimos quadrados é estimar o valor de α e β baseado no conjunto de dados X e Y . Para isso, precisamos definir o resíduo, denotado por r , que é o indicador da variação natural dos dados em relação ao modelo. Matematicamente o resíduo é definido como a diferença do valor observado e o valor

predito pelo modelo, ou seja,

$$r_j(\alpha, \beta) = y_j - \beta x_j - \alpha. \quad (1.39)$$

Tomando a soma quadrática dos resíduos

$$r_{\text{sum}}^2(\alpha, \beta) = \sum_{j=1}^m r_j^2, \quad (1.40)$$

o métodos dos mínimos quadrados estima os parâmetros minimizando r_{sum}^2 em relação a α e β . Ao derivar e igualar a zero surge um sistema de duas equações e duas variáveis que quando resolvido para α e β , leva a:

$$\beta = \frac{\sum_{j=1}^m (x_j - \langle x \rangle)(y_j - \langle y \rangle)}{\sum_{j=1}^m (x_j - \langle x \rangle)^2} \quad (1.41)$$

e

$$\alpha = \langle y \rangle - \beta \langle x \rangle. \quad (1.42)$$

Um requisito para o método dos mínimos quadrados é que o fator de erro (ϵ) seja distribuído aleatoriamente e essa distribuição seja normal.

1.12 Coeficiente de determinação

Assim como abordado na seção sobre *bootstrap*, a ciência estatística fornece algumas ferramentas para validar ou testar a qualidade de uma estimativa. Nesse contexto, o coeficiente de determinação é uma forma de quantificar a qualidade de uma regressão linear ajustada a um conjunto de dados observados [67]. Em estatística, o coeficiente de determinação é representado por R^2 ou até mesmo pode ser escrito por extenso R-quadrado. A definição matemática usualmente utilizada é dada por

$$R^2 = 1 - \frac{r_{\text{sum}}^2}{\sigma_Y^2}, \quad (1.43)$$

em que r_{sum}^2 é a soma quadrática dos resíduos para o par de parâmetros ajustados (α e β), enquanto σ_Y^2 é a variância do conjunto observado Y . O R^2 expressa a fração da variância do dado que é explicada por uma regressão linear. Um valor nulo de R^2 indica que o modelo é tão bom quanto uma constante representada pela média dos dados. Se R^2 é igual a 1, isso representa que o modelo explica toda a variância do dado [68]. Nesse caso, quanto maior o R^2 , mais explicativo é o modelo linear. Por exemplo, $R^2 = 0,75$ significa que o modelo linear explica 75% da variância da variável dependente a partir da variável independente do modelo.

Complexidade da popularidade e dinâmica das conquistas dentro dos jogos *online*

Tarefas de diferentes naturezas e níveis de dificuldade fazem parte da vida das pessoas. Nesse contexto, há um interesse científico na relação entre a dificuldade da tarefa e a persistência necessária para realizá-la. Apesar da generalidade desse problema, algumas tarefas podem ser simuladas na forma de jogos. Aproveitando isso, neste capítulo, empregamos dados de uma grande plataforma *online* chamada Steam para analisar jogos e o desempenho de seus jogadores [54]. Mais especificamente, investigamos a persistência na conclusão de tarefas com base na proporção de jogadores que alcançaram os *achievements*¹ nos jogos. No geral, apresentamos cinco descobertas principais. Primeiro, a distribuição de probabilidade para o número de *achievements* segue uma distribuição log-normal. Segundo, a distribuição dos jogadores também segue uma log-normal. Terceiro, a maioria dos jogos não exige um grau muito alto nem muito baixo de persistência. Quarto, os jogadores também preferem jogos que exigem um certo nível intermediário de persistência. Quinto, a proporção de jogadores em função do número de *achievements* diminui aproximadamente de forma exponencial. Como as funções log-normal e exponencial são consideradas distribuições sem memória, elas são formas matemáticas que descrevem efeitos aleatórios decorrentes da natureza do sistema. Portanto, nossas duas primeiras descobertas descrevem processos aleatórios de fragmentação de conquistas e jogadores, enquanto as três últimas fornecem uma medida quantitativa da preferência humana na busca de tarefas desafiadoras, alcançáveis e justificáveis.

¹Um *achievement* simboliza uma tarefa concluída dentro de um determinado jogo. Isso será discutido em detalhe na sequência deste capítulo.

2.1 Introdução

Somos constantemente confrontados com as mais diversas tarefas ao longo de nossas vidas. Algumas dessas tarefas podem ser simples, complexas ou até mesmo impossíveis, e para cada tarefa, a dedicação do indivíduo para realizá-la depende do seu grau de dificuldade [69, 70]. Nesse contexto, há um interesse em compreender a relação entre dificuldade, persistência e sucesso [71, 72]. A persistência na juventude, por exemplo, mostrou ser um preditor de resultados acadêmicos, renda e nível ocupacional [73–75]. Muitas vezes, investigações quantitativas sobre a taxa de sucesso na realização de tarefas não são fáceis de serem conduzidas porque esse tipo de dado é de difícil obtenção. Neste trabalho, em vez de analisar a taxa de sucesso de tarefas em geral, investigamos esse assunto no contexto dos *videogames*, do qual aproveitamos uma grande base de dados disponível.

Investigar aspectos relacionados a jogos é bastante pertinente na sociedade atual, uma vez que eles têm ganhado grande popularidade e importância ao longo dos últimos anos. Particularmente, o mercado de jogos está crescendo aceleradamente, com receitas chegando a quase 200 bilhões de dólares em 2023, sendo substancialmente maior do que o PIB da maioria dos países [76]. Em 2015, por exemplo, 60% dos habitantes dos Estados Unidos jogavam jogos digitais. A idade média dos jogadores era de cerca de 35 anos, e a proporção de jogadores masculinos e femininos era de 56% e 44%, respectivamente [77]. A aceitação dos jogos pela sociedade contemporânea, combinada ao caráter lúdico, fez surgir aplicações em atividades do cotidiano que simulam muitas tarefas humanas. Consequentemente, o conceito de “gamificação” está emergindo nos mais diversos contextos, como no trabalho e no treinamento de atividades físicas [78].

Essa onipresença dos jogos na sociedade tem despertado interesse científico em múltiplas áreas do conhecimento [79]. Em particular, os jogos estão sendo utilizados em estudos sobre a persistência humana. Consequentemente, teorias têm sido implementadas na tentativa de descrever os fatores que afetam a motivação individual nos jogos. Kahn et al. [80] desenvolveram e validaram uma escala destinada a medir a motivação para jogar *videogames* com seis componentes motivacionais: socializador, completador, competidor, escapista, focado na história e intelectual. Outros trabalhos adaptam abordagens já consolidadas em outros campos do conhecimento e as aplicam ao contexto dos jogos. Por exemplo, no campo da psicologia, o conceito de *flow* refere-se ao “processo de experiência ótima” [81]. A emergência do *flow* depende de condições críticas entre a fronteira do nível de habilidade do indivíduo no respectivo domínio e do nível de desafio da tarefa [82]. Com base nessa abordagem, foi demonstrado que diferentes jogos afetam a experiência de *flow* do jogador e, consequentemente, sua motivação [83]. Essa vertente da literatura científica deixa clara a importância de compreender melhor os aspectos da motivação humana. Em uma primeira instância, as investigações podem se concentrar nos jogos, porém os resultados obtidos dessa pesquisa

podem ajudar a entender a motivação humana para realizar tarefas mais gerais da vida.

Apesar de muitas teorias terem sido propostas, o maior desafio na condução desse tipo de pesquisa reside na obtenção de dados confiáveis. Sabe-se que dados autorrelatados podem ser tendenciosos devido a fatores como motivos de autopresentação, incapacidade de introspecção ou limitações das ferramentas de autorrelato [84,85]. No entanto, uma maneira de substituir esses métodos tradicionais das ciências humanas é através do uso de métodos automatizados de coleta de dados que permitem análises em escalas completamente diferentes [86,87]. Atualmente, existem plataformas *online* que vendem e gerenciam uma grande quantidade de jogos: Epic Games [88], GOG [89], EA Origin [90], entre outras. Uma das maiores plataformas de jogos é a Steam [91], que possui uma amostra representativa da população de jogadores, com milhões de usuários simultâneos distribuídos ao redor do mundo. A Steam tem um catálogo com mais de 16 mil jogos. Para a maioria deles, os desenvolvedores inserem marcadores de progresso ao longo do jogo, chamados de *achievements*. Neste trabalho, com base em dados da Steam, consideramos o número de *achievements* alcançados (conhecidos como “desbloqueados” na Steam) como uma medida de sucesso associada à persistência do jogador (veja mais detalhes na seção 2.2).

Portanto, estudamos a dinâmica de conclusão de tarefas usando os *achievements* dos jogos da Steam. A novidade do estudo é o uso de uma grande base de dados coletada automaticamente. Para cada jogo da Steam, obtivemos o nome, o número total de jogadores e a fração de jogadores que completaram cada *achievement*. Todos os dados correspondem a jogos lançados desde o início da plataforma em 2004 até janeiro de 2021. Em nosso trabalho, consideramos a persistência o principal fator para desbloquear os *achievements*, pois a sensação de motivação entre os jogadores aumenta à medida que se aproximam de uma conquista [92]. Por outro lado, do ponto de vista do jogo, associamos a proporção de jogadores que desbloquearam os *achievements* como um indicador vinculado ao fator de engajamento. Essa nomenclatura parece apropriada, uma vez que a maioria dos jogos *online* são sistemas comerciais projetados para motivar os jogadores a continuar jogando, incentivando-os a progredir de um nível a outro [93]. Assim, o principal objetivo do presente capítulo é descrever matematicamente a proporção de jogadores que desbloquearam os *achievements* nos jogos.

Inicialmente, conduzimos uma análise exploratória das variáveis globais do sistema e ajustamos distribuições de probabilidade ao número de *achievements* e ao número de jogadores por jogo. Essas duas distribuições são bem caracterizadas por uma log-normal, distribuição comum em outros contextos das ciências naturais e sociais. Durante este estudo exploratório, também descobrimos que a popularidade de um jogo (medida pelo número total de jogadores) está relacionada ao fator de engajamento. Em seguida, organizamos os jogos em nosso estudo em duas classes distintas: a primeira referente a jogos em que pelo menos um jogador completou todos os *achievements* e a segunda na qual nenhum jogador completou todos os *achievements*. Usando essa classificação, avaliamos a fração de jogadores

que completaram todos os *achievements* (fração de completadores) e a fração máxima de *achievements* completados pelos jogadores (fração completada). Focamos separadamente na forma das distribuições de probabilidade do número de jogos em função da fração de completadores ou em termos da fração completada. Em particular, identificamos que os jogos com mais jogadores são aqueles que exigem mais persistência, mas não parecem impossíveis. Avançando em nosso estudo, investigamos a fração de jogadores ao longo dos *achievements* de cada jogo. Separando os jogos em grupos com um fator de engajamento semelhante, um decaimento exponencial é identificado ao longo das conquistas do jogo. Portanto, mais uma vez, encontramos um tipo de padrão que é comum em estudos de vários sistemas naturais e sociais [94–96].

2.2 Conjunto de dados

Os dados utilizados neste trabalho foram coletados da plataforma Steam e estão disponíveis gratuitamente. A Steam é uma plataforma *online* desenvolvida pela Valve Corporation, que promove a distribuição digital e oferece gerenciamento de direitos digitais dos produtos de sua base [91]. Atualmente, o Steam é uma das principais plataformas que oferece esses serviços, alcançando mais de vinte milhões de usuários *online* simultaneamente [97]. Além disso, ela registra a atividade de seus usuários de forma automatizada, sem interferir na atividade dos jogadores. Esses fatos indicam que os dados do Steam contêm informações representativas e confiáveis sobre as atividades nos jogos.

Por motivos de confidencialidade e direitos, os dados disponíveis são particularmente limitados para proteger a identidade dos usuários. Esses dados podem ser acessados via API da Steam [98] ou por outros sites que também fornecem alguns desses dados. Por exemplo, o site AStats [99] foi utilizado neste trabalho para obter o nome dos jogos, número de jogadores e número de usuários que alcançaram cada *achievement* dos jogos considerados. Desse modo, os dados se referem apenas aos jogadores que conectaram suas contas da Steam ao AStats. Portanto, não temos acesso às informações de todos os jogadores da Steam. Os dados coletados compreendem os 16.806 jogos da Steam lançados até fevereiro de 2021. Uma parte desses dados não foi analisada devido à falta de informações identificadas nos registros de *achievements*. Após o processo de curadoria dos dados, restaram 15.239 jogos para serem investigados.

Nos jogos, um *achievement* é um meta-objetivo definido fora dos parâmetros do jogo. Os *achievements* marcam uma realização por parte do jogador e não estão necessariamente vinculadas à narrativa principal do jogo. Cumprir as condições do *achievement* e receber o reconhecimento dele pelo jogo é chamado de desbloqueio de *achievement*. Deve-se destacar um caso excepcional de alguns jogos que oferecem o modo “Ironman”, no qual os jogadores enfrentam desafios ou restrições adicionais, resultando frequentemente em uma

experiência mais difícil e rigorosa. Nos jogos que oferecem esse modo, é possível ativá-lo e desativá-lo. Se o modo “Ironman” estiver desativado, os jogadores podem usar trapaças e recursos para facilitar as tarefas dentro do jogo. No entanto, apenas com o modo ativado é possível desbloquear os *achievements*. Um dos principais objetivos dos desenvolvedores ao inserir *achievements* é incentivar os jogadores a continuarem jogando além de simplesmente completar a narrativa principal do jogo. Dessa forma, o desenvolvedor tenta aumentar a longevidade do jogo, fazendo o jogador descobrir todos os seus segredos e completar todos os seus desafios. Assim, os desenvolvedores determinam quantos *achievements* há em cada jogo, bem como quais ações são necessárias para alcançar cada *achievement*. Curiosamente, um jogo com muitos *achievements* não é necessariamente mais difícil ou demorado do que outro com poucos *achievements*.

O aspecto interessante de usar *achievements* como uma métrica para medir a persistência dos jogadores reside na visibilidade dessas conquistas no perfil *online* do jogador. A motivação do jogador para alcançar *achievements* está em maximizar sua própria pontuação geral em vários jogos e obter reconhecimento por seu desempenho devido à publicação de seus perfis de conquistas. Alguns jogadores perseguem o desbloqueio de *achievements* como um objetivo em si, sem se preocupar particularmente em aproveitar o jogo; essa comunidade de jogadores frequentemente refere-se a si mesma como “caçadores de *achievements*” [100]. Aproveitando esse cenário competitivo, interpretamos desbloqueios de *achievements* como tarefas e, para obter resultados quantitativos, usamos o número de jogadores como a principal métrica. Referimo-nos ao número de jogadores de um jogo como usuários da Steam que vincularam suas contas ao AStats, compraram o jogo, instalaram e jogaram pelo menos uma vez. Além disso, para cada jogo em nosso banco de dados, obtemos o número de jogadores que completaram cada *achievement*.

Nesse contexto, consideramos os *achievements* como indicadores de sucesso. No entanto, podem haver vários fatores ligados à decisão das pessoas de continuar jogando. De todos os fatores, a literatura foca em três principais: jogar com amigos (fator social), diversão intrínseca do jogo (fator imersivo) e desafio (fator individual) [101–106]. Assim, para simplificar, definimos que alcançar um *achievement* é resultado da persistência do jogador. Note que a definição de persistência usada neste trabalho resume todos os fatores responsáveis por manter o jogador desbloqueando os *achievements*. Do ponto de vista do jogo, o conjunto de persistências dos jogadores pode conduzir a uma medida de engajamento, aqui chamada de fator de engajamento. Mais uma vez, independentemente de o jogo ser divertido ou chato, fácil ou difícil, o fator de engajamento do jogo resume todos esses fatores e expressa o grau em que mantém seus jogadores buscando os *achievements*. Em geral, consideramos que uma grande proporção de jogadores alcançando os *achievements* em um determinado jogo significa um alto nível de engajamento entre os jogadores desse jogo, o que indica que o jogo possui um alto fator de engajamento. Matematicamente, definimos o fator de engajamento

de duas maneiras, pois ao longo do trabalho dividimos os jogos em dois grupos. O primeiro grupo representa os jogos nos quais pelo menos um jogador alcançou todos os *achievements* disponíveis; para esse grupo, consideramos o fator de engajamento como a fração de completadores, ou seja, a fração de jogadores que desbloquearam todos os *achievements* do jogo. No segundo grupo, consideramos os jogos nos quais nenhum jogador completou todos os *achievements*; para esse grupo, o fator de engajamento é a fração completada, a fração de *achievements* que o melhor jogador alcançou com sucesso.

2.3 Resultados

Começamos investigando o total de *achievements*, o total de jogadores, a fração de completadores e a fração completada. Um aspecto do total de *achievements* (número de partes de um jogo) de cada jogo mostra que esse total varia desde um até milhares de *achievements*. O valor máximo do total de *achievements* é 9.821, correspondendo ao jogo *LOGistICAL*. A distribuição de *achievements* para todos os jogos é mostrada na Figura 2.1A. Essa distribuição apresenta um pico aproximadamente em 10 *achievements*, o que representa 644 jogos. Essa figura também mostra o ajuste dos dados via distribuição log-normal (apresentada no capítulo 1) com a forma

$$p(N) = \frac{1}{\sqrt{2\pi} \sigma N} \exp\left(-\frac{(\ln N - \mu)^2}{2\sigma^2}\right), \quad (2.1)$$

em que N representa o total de *achievements*, $\ln$ é o logaritmo natural, $\mu = 3,20$, e $\sigma = 0,66$. Vale a pena dizer que 75% do dado tem até 40 *achievements*, e 95% dos jogos têm menos de 100 *achievements*. Esse fato está de acordo com a recomendação da *Steam* de que os jogos devem ter um máximo de 100 *achievements* [107]. Note ainda que a variável N é discreta, porém estamos aproximando-a por uma variável contínua ao empregarmos a distribuição log-normal na Equação (2.1).

Uma simples inspeção nos dados é suficiente para revelar que há uma abundância de jogos com um número relativamente pequeno de usuários e poucos jogos com muitos usuários. Em uma análise mais detalhada dos jogos com poucos jogadores, verificamos que o número de jogos decresce com o número de jogadores. Por exemplo, 647 jogos têm apenas 1 jogador, 397 jogos têm 2 jogadores, e 324 jogos 3 jogadores. Além disso, 78% dos jogos menos populares não têm mais do que 500 usuários. No outro extremo, o jogo mais popular, *Counter-Strike: Global Offensive*, tem 241.325 jogadores em nossa base de dados. A Figura 2.1B mostra a distribuição do número de jogos como uma função do número de jogadores, em que um claro comportamento de decaimento é observado. Essa distribuição está também em boa concordância com a distribuição log-normal, Equação (2.1), com parâmetros $\mu = 3,96$ e $\sigma = 2,18$. Como podemos ver, a cauda dessa log-normal é muito maior que aquela apresentada

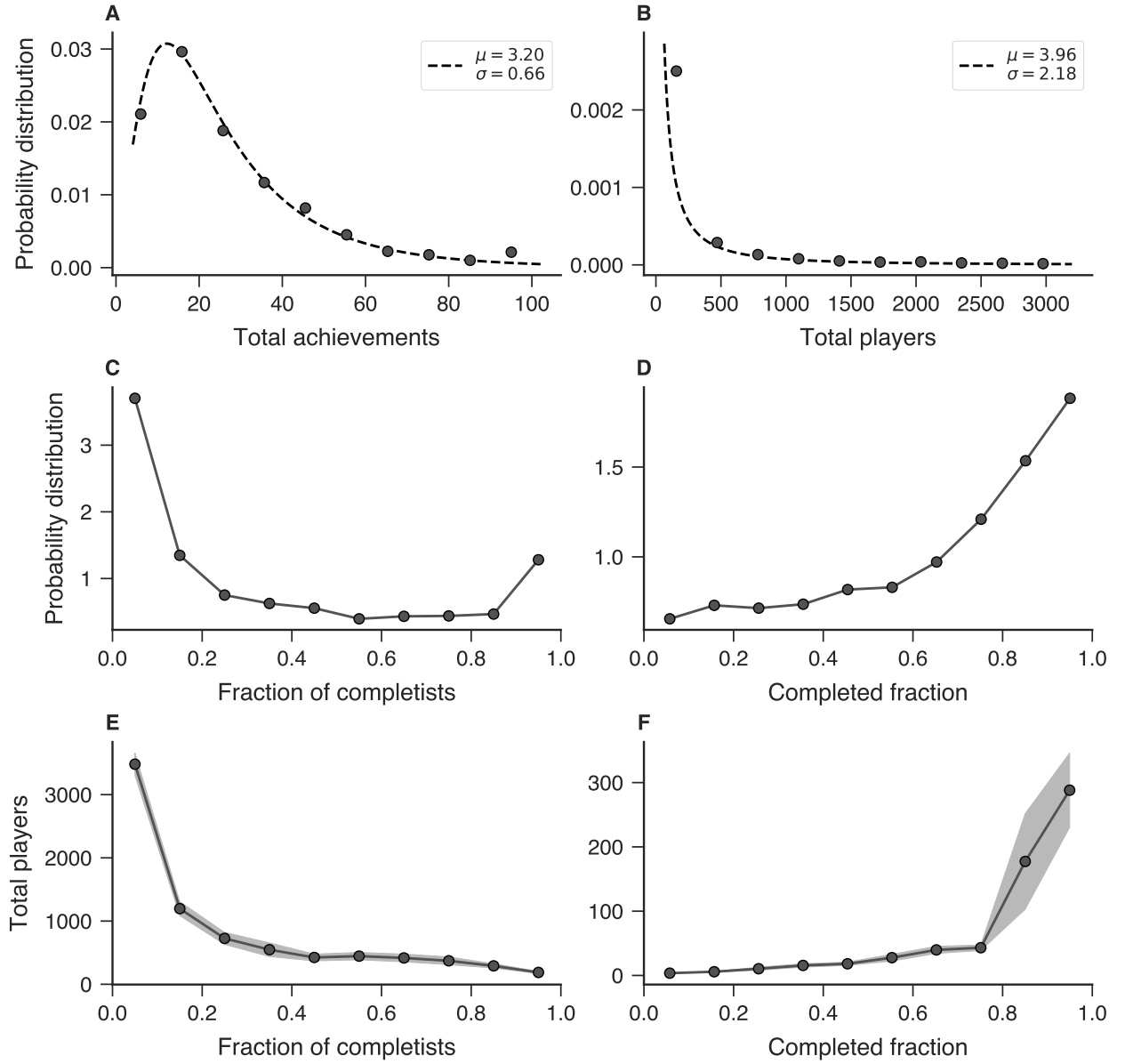


Figura 2.1: Distribuições de probabilidade das variáveis de jogo e jogador. Distribuição de probabilidade do número de jogos como função do (A) número de *achievements*, (B) número de jogadores, (C) fração de completadores e (D) fração completada. Os painéis (E) e (F) mostram a relação entre a fração de completadores e o número de jogadores e entre a fração completada e o número de jogadores, respectivamente. Nos painéis (A), (B), (C) e (D), os marcadores representam a probabilidade no eixo y e o centro do intervalo no eixo x. As linhas tracejadas em (A) e (B) referem-se ao ajuste da distribuição log-normal com parâmetros $\mu_A = 3,20$ [95% IC: 3,18, 3,22] e $\sigma_A = 0,66$ [95% IC: 0,64, 0,67], and $\mu_B = 3,96$ [95% IC: 3,91, 4,00] e $\sigma_B = 2,18$ [95% IC: 2,13, 2,23], respectivamente. Nos painéis (E) e (F) os marcadores representam a média por intervalo e a área sombreada indica o erro relativo da média.

na Figura 2.1A, visto que o parâmetro σ é maior no caso da Figura 2.1B.

Existem *achievements* que são relativamente difíceis de completar pelos jogadores, exi-

gindo um investimento de tempo significativo ou habilidades complexas de resolução de tarefas. Esse fato indica que há uma grande diversidade no grau de dificuldade dos jogos. Por exemplo, o número de jogos que pelo menos um jogador concluiu todos os *achievements* é 11.831, compreendendo aproximadamente 78% dos 15.239 jogos da nossa base de dados. Uma possível métrica para quantificar a dificuldade de concluir um jogo é a fração de usuários que o completou. Como mencionado na seção dos Métodos, essa fração é referida como a fração de completadores, podendo variar de 0, para jogos que nenhum jogador completou, até 1, para jogos que todos os jogadores completaram. Considerando apenas jogos que pelo menos um jogador conseguiu concluir todos *achievements*, a Figura 2.1C mostra a distribuição de probabilidade do número de jogos como função da fração de completadores. Essa distribuição é concentrada em torno dos menores valores da fração de completadores, isto é, há muito mais jogos completados apenas por uma pequena fração de jogadores. Quantitativamente, 6.076 jogos ($\sim 51\%$ dos jogos nos quais pelo menos um jogador completou todos *achievements*) correspondem à fração de completadores menor ou igual a 0,2.

Há também 3.408 jogos em que nenhum jogador conquistou todos os *achievements*, representando cerca de 22% de nossa base de dados. Nesse caso, focamos nossa análise para a fração de *achievements* que o melhor jogador alcançou com sucesso. Exatamente como definido na seção de Métodos, essa métrica é referida como a fração completada. Assim, se a fração completada é 0,6 para um dado jogo, apenas 60% dos seus *achievements* foram alcançados considerando a performance do melhor jogador. No nosso banco de dados, nenhum jogo tem a fração de completadores igual a 0, mas 181 jogos apresentam a fração menor ou igual a 0,1, e 1.151 têm essa fração entre 0,8 e 1. A Figura 2.1D mostra a distribuição de probabilidade do número de jogos como função da fração completada. Para os jogos mais difíceis, aqueles com pequena fração completada, observamos um aumento monotônico. No entanto, essa taxa de crescimento torna-se maior para valores mais altos da fração completada.

Relacionado às Figuras 2.1C e 2.1D, informações complementares sobre o nosso banco de dados são fornecidas em termos de distribuição de jogadores em função do fator de engajamento. Esses aspectos interrelacionados são exibidos nas Figuras 2.1E e 2.1F. Como podemos ver, o padrão das Figuras 2.1C e 2.1D são essencialmente replicadas nas Figuras 2.1E e 2.1F. Tanto a popularidade como o número de jogos decrescem com o aumento da fração de completadores, e crescem com o aumento da fração completada. Por exemplo, jogos com a fração de completadores maior que 0 e menor ou igual a 0,2 acumulam 17.273.449 jogadores, correspondendo a aproximadamente 86% de todos os jogadores. Além disso, jogos com fração completada maior ou igual a 0,8 contam com 266.816 usuários, representando um pouco mais de 1% de todos os jogadores.

Na direção de uma investigação mais detalhada, analisamos a fração de jogadores ao longo de todos os *achievements*. Nesse caso, para cada jogo, os *achievements* foram organizados em

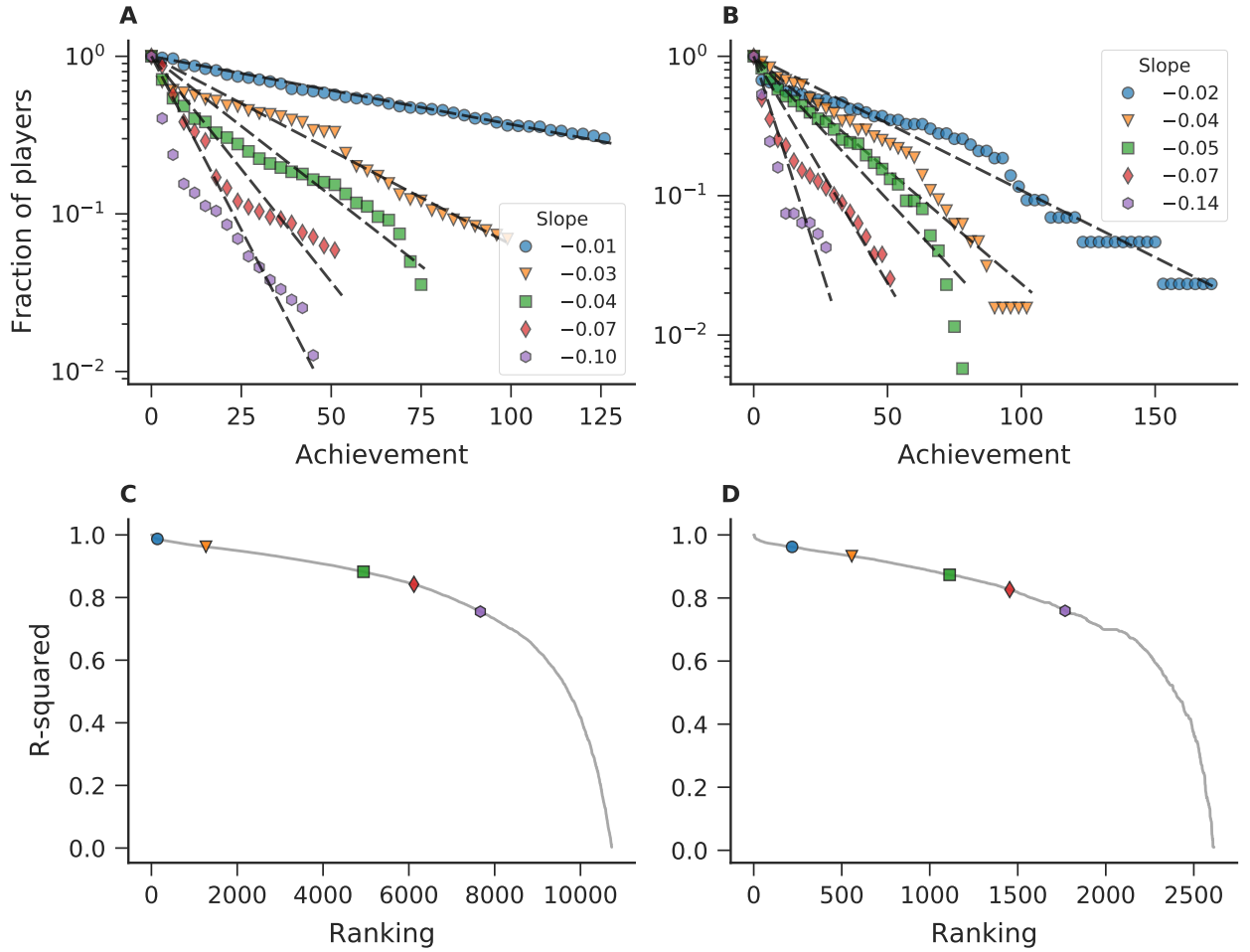


Figura 2.2: Decaimento exponencial da fração de jogadores como função dos *achievements* desbloqueados. Exemplos de decaimentos da fração de jogadores como função do número de *achievements* alcançados para (A) jogos com completadores e (B) jogos em que nenhum usuário completou todos os *achievements*, em escala mono-log. Os marcadores referem-se ao dado e as linhas retas são correspondentes aos ajustes exponenciais. O ranqueamento dos coeficientes de R-quadrado para os ajustes considerando (C) jogos com completadores e (D) jogos em que nenhum usuário completou todos os *achievements*. Os marcadores coloridos correspondem aos valores de R-quadrado das curvas mostradas nos painéis (A) e (B), e a linha cinza mostra o comportamento do ranqueamento do R-quadrado para todos os jogos.

ordem decrescente. Assim, para cada jogo, o primeiro *achievement* corresponde àquele que o maior número de jogadores alcançou com sucesso e o último *achievement* refere-se àquele que o menor número de jogadores alcançou. Na maioria das vezes, a fração de jogadores é próxima de 1 para o primeiro *achievement*, visto que quase todos os jogadores alcançaram um sucesso mínimo. Um estudo exploratório de jogos escolhidos aleatoriamente indicou um decaimento exponencial da fração de jogadores em função do número de *achievements*. Tomando vantagem do fato que o comportamento exponencial é bem representado como uma linha reta em um gráfico mono-log, algumas dessas escolhas aleatórias estão ilustradas

nas Figuras 2.2A e 2.2B. As linhas retas foram fixadas começando em frações iguais a 1 quando o número de *achievements* é 0; esta escolha simples foi adotada porque todos os jogadores de qualquer jogo correspondem a uma fração igual a 1. Dessa forma, há apenas um parâmetro para indicar o decaimento exponencial dos jogos, a inclinação da reta. Para um jogo com total de *achievements* N , supomos que a variável n representa um dado número de *achievements* ($0 \leq n \leq N$). Assim, a fração de jogadores do n -ésimo *achievement* ($f(n)$) é descrita por:

$$f(n) = \exp(\beta n), \quad (2.2)$$

em que β indica a inclinação de sua linha tracejada. Como podemos ver nas Figuras 2.2A e 2.2B, alguns comportamentos são próximos à exponencial e outros revelam claros desvios. Vale destacar que a Figura 2.2A é associada aos jogos com completadores e a Figura 2.2B exemplifica os jogos sem completadores.

Para quantificar a qualidade desses ajustes exponenciais, calculamos o coeficiente de determinação R-quadrado (R^2) [68]. Esse coeficiente expressa a porcentagem de variância dos dados que é explicado pelo modelo linear. Um valor de R^2 nulo indica que o modelo é tão bom quanto uma linha constante igual a média dos dados, já se o R^2 é igual a 1, isso representa que o modelo explica toda a variação nos dados. Para restringir a análise a um modelo linear, o logaritmo dos dados foi empregado ao longo do nosso estudo. A Figura 2.2C mostra o gráfico de *ranking* de R^2 para todos os jogos com frações de completadores não nulas, enquanto a Figura 2.2D exibe o mesmo tipo de gráfico para os jogos que nenhum jogador completou todos os *achievements*. Essas figuras apresentam um lento decaimento para os primeiros classificados, indicando que a maioria dos jogos têm um R^2 alto. Aqui em particular, 8.378 jogos tem R^2 maior que 0,7, correspondendo a aproximadamente 71% dos jogos com fração de completadores não nula. Paralelamente, para jogos em que nenhum jogador concluiu todos os *achievements*, 1.996 apresentam R^2 maior que 0,7, representando cerca de 59% desses jogos. Note que a Figura 2.2D contém apenas 2.616 dos 3.408 jogos em que nenhum jogador conseguiu concluir todos os *achievements*. Esse corte nos dados ocorre porque os valores negativos de R^2 foram omitidos. Esses valores negativos indicam situações nas quais uma reta horizontal é um melhor ajuste que um reta inclinada [108]. Esses resultados indicam que os jogos com completadores são usualmente melhor ajustados pelo decaimento exponencial do que os demais jogos.

Apesar da variabilidade em torno do modelo da fração de jogadores como função dos *achievements*, trabalhamos na direção de identificar padrões associados com esses decaimentos. No entanto, visto que há muitos jogos com diferentes números de *achievements*, empregamos a métrica normalizada $x = n/N$, definida como a razão entre o n -ésimo *achievement* e o total de *achievements* N de um jogo. Portanto, o *achievement* normalizado x assume valores entre 0 e 1. Avançando nessa linha de raciocínio, para cada jogo, vamos nos referir à fração de jogadores que alcançaram a proporção x de *achievements* como $f(x)$; assim, $f(x) = g(n)/G$,

em que $g(n)$ é o número de usuários que completou o n -ésimo *achievement* e G é o total de jogadores. Se $x = 0$, todos os jogadores são incluídos e temos $f(0) = 1$ e $g(0) = G$. Já no outro extremo, $f(1) = F$ e $g(N) = GF$, em que F é a fração de completadores.

Para identificar um comportamento coletivo médio, dividimos os dados em intervalos de F e calculamos o valor médio de $f(x)$ para cada intervalo. Isto é, agrupamos os jogos baseados na fração de completadores em intervalos igualmente espaçados e não sobrepostos de tamanho 0,1. Dentro de cada grupo, calculamos a série média e ajustamos um decaimento exponencial. Exemplos desse procedimento são mostrados na Figura 2.3A, em que os marcadores coloridos correspondem às séries médias e a linha tracejada ao ajuste exponencial de cada grupo. Como a figura está em escala mono-log, essa linha reta corresponde a um decaimento exponencial. Levando em consideração que a série média decai exponencialmente, podemos escrever $f(x) = \exp(\alpha x)$, em que α refere-se à inclinação da reta. Empregando $f(1) = F$, tem-se $F = \exp(\alpha)$ e $\alpha = \log(F)$. Na Figura 2.3B, apresentamos a inclinação α como função da fração de completadores F . Os símbolos em destaque representam a fração média de completadores em cada grupo e suas respectivas inclinações (α), a área sombreada em cinza demarca um desvio padrão para cima e um desvio padrão para baixo, e a linha tracejada é $\log(F)$. Nossos resultados permanecem robustos se o tamanho do intervalo escolhido da fração de completadores para agrupar os jogos varia entre 0,05 e 0,25. Usando $x = n/N$ e $f(x) = g(n)/G$, podemos alternativamente escrever o resultado empírico $f(x) = \exp(\log(F)x)$ como

$$g_c(n) = G \exp\left(\frac{\log F}{N} n\right), \quad (2.3)$$

em que o índice c foi introduzido para indicar jogos com completadores. Portanto, para cada jogo, esse resultado deve ser visto como o número médio de usuários que alcançou o n -ésimo *achievement* quando há uma fração F de completadores, N *achievements* e G jogadores.

Uma linha de raciocínio similar à aplicada nos dois parágrafos anteriores pode ser empregada para investigar o comportamento médio dos jogos sem completadores. Como no caso anterior, o número de jogadores que começam um dado jogo é G , ou seja, $g(0) = G$. Para a vasta maioria dos jogos sem completadores, verificamos que apenas uns poucos jogadores C de cada jogo desbloquearam seu *achievement* mais difícil, porém não impossível, $n_{\max}$. Por exemplo, há aproximadamente 82,2%, 9,4% e 3,0% dos jogos com C igual a 1, 2, e 3, respectivamente. Dessa forma, obtemos $g(n_{\max}) = C$. Similar e consistentemente com o comportamento exponencial apresentado na Figura 2.2B, consideramos $g(n) = G \exp(\beta n)$, em que β é o parâmetro a ser ajustado. Visto que $g(0) = G$ e $g(n_{\max}) = C$, temos $\beta = \log(C/G)/n_{\max}$. Esses resultados conduzem a

$$g_u(n) = G \exp\left(-\frac{\log(G/C)}{n_{\max}} n\right), \quad (2.4)$$

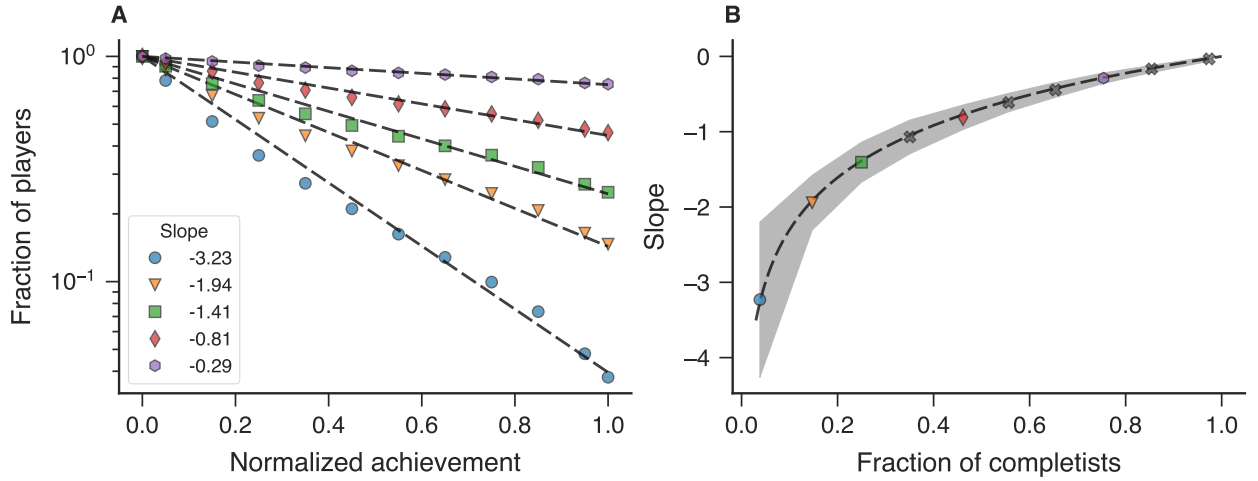


Figura 2.3: Decaimento exponencial médio para jogos com completadores. (A) Exemplos de decaimento da fração média de jogadores como função dos *achievements* normalizados na escala mono-log. Os jogos foram agrupados em intervalos variantes de fração de completadores, em incrementos de 0,1. Dentro de cada grupo, uma série média foi construída com 10 marcadores igualmente espaçados. Adicionalmente, o primeiro ponto da série média foi fixado como (0, 1) e o último ponto como (1, μ_F), em que μ_F é a média da fração de completadores do grupo. No fim desse procedimento, a série média de todos os grupos tinha 12 marcadores. Os marcadores azuis representam a média dos grupos de jogos com fração de completadores entre 0 e 0,1, os marcadores laranjas entre 0,1 e 0,2, os marcadores verdes entre 0,2 e 0,3, marcadores vermelhos entre 0,4 e 0,5 e marcadores roxos entre 0,7 e 0,8. As linhas tracejadas representam o modelo exponencial ajustado a cada grupo de completadores. (B) Relação entre a taxa de decaimento exponencial e a fração de completadores. Dessa forma, calculamos a inclinação média α e a média da fração de completadores F para cada grupo de jogos. As cores correspondem aos grupos mostrados em (A), isto é, o marcador azul é a inclinação do grupo de jogos com a fração de completadores entre 0,0 e 0,1, o marcador laranja é a inclinação do grupo de jogos com fração de completadores entre 0,1 e 0,2, e assim por diante. Os marcadores cinzas representam os grupos de jogos com fração de completadores não exemplificados em (A). A área sombreada cinza mostra a banda de desvio padrão e a linha tracejada é a função logarítmica $\alpha = \log(F)$.

em que o índice u representa os jogos sem completadores. Assim, empregamos $g_u(n)$ como o número médio de jogadores que alcançou o n -ésimo *achievement* para um jogo com G jogadores e C jogadores que concluíram o número máximo de *achievements* $n_{\max}$. Nessa direção, a fração de jogadores pode ser obtida definindo $f(n) = g_u(n)/G$ e, analogamente ao que foi feito no caso anterior, $f(x) = \exp(-\log(G/C)x/U)$, em que $x = n/N$ e $U = n_{\max}/N$. Lembrando que $f(x) = \exp(\alpha x)$, podemos escrever a relação $\alpha U = -\log(G/C)$, em que α é a inclinação do decaimento da fração de jogadores ao longo dos *achievements* normalizados. A Figura 2.4 mostra o comportamento médio da relação entre essas quantidades. Visto que para 82,2% dos jogos sem completadores, há apenas um jogador que alcançou $n_{\max}$ *achievements*, podemos aproximar C por 1 ($C \approx 1$). Consequentemente, aproximamos essa

relação por um modelo logarítmico $\alpha U = -\log(G)$, representado pela linha tracejada.

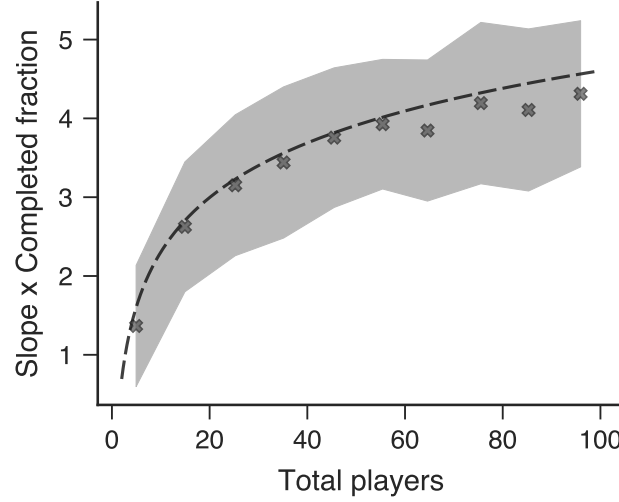


Figura 2.4: Relação entre a taxa de decaimento exponencial multiplicada pela fração completada e total de jogadores. Para cada jogo, a taxa de decaimento exponencial α foi calculada usando a fração de jogadores no eixo y e os *achievements* normalizados no eixo x (todos os valores nulos das séries foram removidos). Depois, o par ordenado $(G, \alpha U)$ foi tomado, e os jogos foram agrupados em intervalos contendo total de jogadores variando em incrementos de 10. Dentro de cada intervalo, foram calculadas as médias nos eixos x e y. Os marcadores cinzas representam os pares ordenados dessas médias para cada grupo. A área sombreada cinza mostra o desvio padrão e a linha tracejada representa a função logarítmica ($-\alpha U = \log(G)$).

2.4 Discussão

Como já apontado, os *achievements* são demarcadores de progresso de cada jogo que como consequência promovem sua fragmentação em partes. Essa fragmentação pode apresentar um certo grau de aleatoriedade, por exemplo, devido ao fato de que os jogos são diferentes uns dos outros e as desenvolvedoras são usualmente distintas. Nesses casos, podemos pensar em começar do todo e então fazer partições sucessivas aleatoriamente. Suponha uma parte de tamanho p_1 que é dividida e uma de suas componentes resultantes tem um tamanho $p_1 p_2$, em que p_2 é a fração de particionamento. Para n fragmentações consecutivas e aleatórias, esse resultado é estendido para o produto $p_1 p_2 \dots p_n$, isto é, obtemos um processo multiplicativo aleatório. Em situações como essa, é comum empregar, em uma primeira abordagem, a distribuição log-normal, Equação (2.1), para descrever a distribuição de tamanhos [109]. Particularmente, ressaltamos que há sistemas sociais que podem ser divididos em partes com algum grau de aleatoriedade e modelado com sucesso pela distribuição log-normal. Por exemplo, a distribuição log-normal ocorre ao investigar o número de votos recebidos por candidatos em eleições [110], a dinâmica de popularidade de memes [111],

os valores *dig* [112], e a distribuição de grau das redes de trocas de mensagens em fóruns na internet [113]. Exemplos como os citados e a abordagem teórica acima nos motivaram a modelar a distribuição do número de jogos como função do total de *achievements* por uma log-normal, como mostrado na Figura 2.1A. Alternativamente, tentamos ajustar outras distribuições, tais como a de Weibull, a gama e de Burr. No entanto, com a log-normal, obtivemos a melhor estatística no teste de Kolmogorov-Smirnov.

Como jogadores são distribuídos nos diferentes jogos, podemos considerar um modelo de partição desses usuários. Nessa direção, o modelo exposto no parágrafo anterior pode ser visto como uma primeira aproximação. No entanto, consideramos que todos os jogadores são divididos entre os jogos, em vez de cada jogo dividido em *achievements*. Esse raciocínio nos inspirou a empregar novamente uma log-normal, Equação (2.1), para modelar a distribuição de jogos como função do total de jogadores, como apresentado na Figura 2.1B. A exemplo do que foi pontuado em conexão com a Figura 2.1A, outras distribuições foram testadas, mas a log-normal teve a melhor estatística no teste de Kolmogorov-Smirnov.

Tomando vantagem das definições de fração de completadores e fração completada, imediatamente identificamos a classe dos jogos mais comuns e a classe dos mais populares. Para jogos com completadores, os mais comuns são os que não são finalizados tão facilmente (Figura 2.1C). Considerando jogos não completados pelos jogadores, os mais frequentes são os mais próximos de serem completados (Figura 2.1D). Em relação à popularidade, as Figuras 2.1E e 2.1F indicam padrões similares considerando jogadores em vez de jogos. Esses resultados são consistentes com pesquisas anteriores. Se considerarmos o fator de engajamento como uma medida do grau médio de persistência (associado à dificuldade) para concluir todos os *achievements*, então nosso trabalho reforça a ideia que a motivação cresce com o aumento da dificuldade da tarefa, enquanto o sucesso é possível e justificável [70, 72, 114]. Strojny e colaboradores encontraram que o envolvimento das pessoas cresce ao passo que o jogo se torna mais difícil, mas depois de uma certa dificuldade “ótima” o envolvimento decai rapidamente [115]. Além disso, parece claro que a intenção dos desenvolvedores é definir *achievements* visando esse grau “ótimo” de persistência. Os desenvolvedores fazem isso na tentativa de motivar os jogadores a continuarem jogando e consumindo tudo o que o jogo tem a oferecer.

Antes de discutir os resultados relacionados às Figuras 2.2, 2.3 e 2.4, vamos considerar uma fundamentação teórica para analisar a dificuldade efetiva em performar as diferentes partes de uma tarefa. Particularmente, podemos pensar no engajamento relativo de duas partes diferentes. Esse engajamento relativo pode ser definido como o número de pessoas que finalizam uma parte dividida pelo número de pessoas que finalizaram a parte anterior. Note que nessa comparação, é vantajoso considerar situações com muitas pessoas executando a dada tarefa a fim de superar as flutuações estatísticas indesejáveis. Se as partes de uma tarefa estão organizadas em ordem decrescente, como os *achievements* estão dispostos no atual

trabalho, é natural considerar o engajamento relativo de passos sucessivos. Empregamos r_i para denotar o engajamento relativo da i -ésima parte em comparação com a parte anterior. Assim, $r_i = n_i/n_{i-1}$ com $i = 1, 2, \dots$, em que n_i refere-se ao número de pessoas que finalizaram o i -ésimo passo e, particularmente, n_0 é o número de pessoas que começou a tarefa. Consequentemente, r_i assume valores de 0 a 1, com o valor 0 ocorrendo quando o número de pessoas que completou a i -ésima tarefa é nulo, enquanto o valor 1 ocorre quando todas as pessoas completaram a tarefa. Em geral, temos $n_1 = n_0 r_1$, $n_2 = n_1 r_2$ e assim por diante. Como consequência, o número de pessoas que terminam a i -ésima parte pode ser escrita em termos dos engajamentos relativos e do número de pessoas inicial, $n_i = n_0 r_1 r_2 \dots r_i$. Aplicando aos jogos, o engajamento relativo entre o primeiro e último *achievements* de um jogo coincide com a fração de completadores desse jogo. Assim, usando a notação acima, essa fração é igual à taxa n_N/n_0 e, consequentemente, reduz-se ao produto $r_1 r_2 \dots r_N$. Visto que a fração de completadores depende de muitos fatores, essa última relação indica que os engajamentos relativos são também conectadas com esses fatores. Além disso, os valores dos engajamentos relativos sucessivos podem oscilar significativamente e apresentar um perfil complexo. Notavelmente, o padrão mais simples desses engajamentos relativos ocorre quando os r_i 's são iguais, ou seja, $r_i = r$ para todo i com r constante. Para esse simples modelo, temos $n_i = n_0 r^i$ ou, alternativamente, o decaimento exponencial $n_i = n_0 \exp(bi)$ com b sendo uma constante positiva igual a $\ln(r)$, dado que r assume valores entre 0 e 1.

Retornando aos nossos resultados, eles indicaram que o modelo exponencial previamente exposto é uma boa aproximação para a maioria dos jogos (veja a Figura 2.2). Curiosamente, esse modelo exponencial pode ser qualificado como sem memória [116]. De fato, se o engajamento relativo é independente dos pares de *achievements* consecutivos considerados ($r_i = r$), não podemos identificar (lembrar) os r_i 's. Nessa direção, desvios do modelo exponencial podem indicar efeitos de memória entre um *achievement* e outro. Consistentemente com essa teoria, podemos definir um índice de ausência de memória como o coeficiente de determinação para as linhas retas ajustadas nas Figuras 2.2A e 2.2B. Assim, o valor máximo para esse índice é igual a 1 para um modelo totalmente sem memória. Dessa definição também segue que, como o índice é apenas afetado por desvios da exponencial (isto é, por valores $R^2 < 1$), ele é independente da sua taxa de decaimento exponencial (o valor de r).

Verificamos que, mesmo depois de agrupar os jogos, o modelo exponencial também fornece um ajuste valioso, como evidenciado nas Figuras 2.3 e 2.4. Particularmente, no caso de jogos sem completadores, há um desvio para um grande número de jogadores, como pode ser visto na Figura 2.4. Inclusive, apresentamos o comportamento médio dos jogos com até 100 jogadores, que representa 89% dos jogos sem completadores, porque para os jogos restantes a flutuação cresce muito. Observando a Figura 2.4, o modelo superestima o dado real (a linha tracejada está acima dos valores médios obtidos dos dados). Isso ocorre porque a linha tracejada representa a função logarítmica com $C = 1$ (o caso mais frequente), mas C pode

assumir valores maiores que 1, visto que ele pode ser proporcional ao total de jogadores de um jogo. Finalmente, o comportamento exponencial médio das equações (2.2) e (2.3) pode ser visto como jogadores desistindo a uma taxa fixa, em que essa taxa depende do fator de engajamento do jogo.

Apesar do grande conjunto de dados e dos resultados estatisticamente robustos, nosso trabalho também apresenta algumas limitações. Acima de tudo, os dados utilizados não distinguem condições específicas de jogos e jogadores. Conforme mencionado anteriormente, alguns jogos possuem o modo “Ironman”, e nossos dados não diferenciam esses jogos. Possivelmente, nesses jogos, os jogadores frequentemente escolhem manter o modo “Ironman” desativado. Nessa condição, os jogadores ainda podem gastar tempo e persistir em objetivos dentro do jogo; no entanto, os *achievements* não serão desbloqueadas. Dado isso, acreditamos que jogos como esses podem experimentar um decaimento exponencial mais rápido ao longo dos *achievements*, já que os jogadores têm a opção de desativar o desbloqueio de *achievements*. Alternativamente, esses *achievements* de uma parte dos jogadores que não são contabilizados poderiam contribuir para o surgimento de outros padrões matemáticos caso fossem contabilizados. Da mesma forma, padrões ligeiramente diferentes podem surgir simplesmente devido a uma diferença intrínseca no jogo, considerando que os jogos possuem diferentes gêneros, estilos e qualidades. Além das características do jogo, nossos dados não diferenciam os jogadores, que variam de jogadores casuais aos já mencionados “caçadores de *achievements*”. Ainda acreditamos que o uso de um conjunto de dados tão geral, como o empregado neste trabalho, oferece uma perspectiva sobre o padrão global da persistência humana. No entanto, se os propósitos e motivações dos jogadores fossem conhecidos, isso poderia levar a resultados mais práticos e direcionados. Aprofundando ainda mais essa discussão, os fatores que impulsionam a motivação de uma pessoa para continuar jogando podem ser dinâmicos ao longo do tempo [117, 118]. Dessa maneira, um banco de dados em grande escala que forneça dados sobre a satisfação dos jogadores ao longo do tempo poderia ser altamente benéfico do ponto de vista científico. Ainda mais, embora não consideremos novos jogos sendo publicados diariamente neste trabalho, os usuários podem simplesmente perder interesse e migrar para novas experiências não por falta de desafio ou tédio, mas devido ao fator de novidade de outros lançamentos.

No capítulo atual, focamos em determinar padrões matemáticos na dinâmica de conclusão de tarefas em jogos da Steam. Embora as definições utilizadas de “persistência” e “fator de engajamento” resumam muitas variáveis, acreditamos que os resultados obtidos neste trabalho contribuem para a compreensão da dinâmica da persistência humana na realização de tarefas. Pesquisas futuras podem estabelecer conexões entre o fator de engajamento e escalas para medir a persistência humana, ou mesmo outras escalas baseadas no tempo gasto e no número de tentativas para resolver uma tarefa lógica [119, 120]. Conectar teoria a dados de grande escala pode ainda proporcionar uma melhoria no desenvolvimento de técnicas de

ensino e aprendizagem. No campo da pedagogia, por exemplo, resultados quantitativos sobre a dinâmica de conclusão de tarefas poderiam ser implementados para maximizar o aprendizado das pessoas. Ou seja, com base no desempenho e na persistência de uma pessoa, a dificuldade das tarefas poderia ser regulada para maximizar os efeitos da aprendizagem. Focando em interesses econômicos, para a indústria de jogos, entender a correlação entre o fator de engajamento e o sucesso de um jogo pode ser útil. Além disso, resultados quantitativos ajudam a compreender melhor o comportamento médio de consumo dos jogadores, o que pode ser útil para melhorar estratégias de vendas e, assim, maximizar os lucros. Em geral, o banco de dados utilizado no trabalho atual é uma amostra representativa de jogadores de todo o mundo, mas trabalhos futuros podem testar a validade de nossas descobertas usando até mesmo outros bancos de dados.

Padrões da dinâmica de popularidade dos jogos *online*

Nos dias atuais, os jogos digitais têm se tornado cada vez mais presentes na vida das pessoas, tanto no lazer quanto em atividades “gamificadas” do cotidiano. Apesar dessa crescente presença dos jogos, análises de grande escala conduzidas a partir de dados reais permanecem escassas na literatura. Nesse sentido, neste capítulo, apresentamos uma investigação de padrões em séries de popularidade de mais de seis mil jogos da plataforma *Steam* [48]. Utilizando essas séries, depois de um estágio de pre-processamento dos dados, realizamos uma análise de *cluster* para agrupar as séries baseadas em suas formas. Nossos resultados indicaram a existência de cinco padrões de forma nomeados como: decrescente, montanhoso, crescente, explosivo e vale¹. Aproximadamente metade dos jogos demonstraram padrão decrescente de popularidade, 20,7% montanhoso, 11,8% crescente, 11,0% explosivo e 9,1% vale. Na parte final do trabalho, investigamos a prevalência e a persistência dos padrões de formas comparando, para as séries de popularidade mais longas, suas formas durante os meses iniciais às formas das séries inteiras. Encontramos que as séries de popularidade da maioria dos jogos tendem a manter seus padrões de forma ao longo do tempo, exceto para as séries constantes que aparecem nos primeiros meses de popularidade e depois transformam-se em séries montanhosas ou explosivas.

3.1 Introdução

Grupos sociais e comunidades são grandemente compreendidos por meio de conteúdos culturais². Em geral, conteúdos culturais referem-se a qualquer forma de informação, expres-

¹Em inglês, os nomes dos *clusters* são *decreasing*, *hilly*, *increasing*, *bursty*, and *valley*.

²Do inglês, *cultural contents*.

são e criatividade que refletem ou representam elementos de uma cultura particular [121]. Conteúdos culturais podem compreender literatura, arte, música, cinema, televisão, dança, folclore, tradições, rituais, idioma, jogos, e assim por diante. Ele também desempenha um papel crucial na formação e manutenção de uma identidade cultural, promovendo um senso de comunidade e transmitindo o conhecimento de uma geração para a próxima [122, 123]. Além disso, como as sociedades evoluem e interagem, conteúdos culturais podem também ser influenciados e influenciar outras culturas, conduzindo a um cenário cultural global interconectado e dinâmico [124]. Especialmente nos anos recentes, conduzidos por ferramentas digitais e tecnológicas, a disseminação de conteúdos culturais tem se tornado mais intensa e acelerada [125]. Em casos extremos, têm surgido fenômenos como *best-sellers* [126], viralização [127], memes [128, 129] e *fake news* [130, 131]. Dessa forma, enquanto conteúdos culturais visam caracterizar uma sociedade, a disseminação de tais conteúdos culturais pode afetar o comportamento coletivo daquela mesma sociedade [132].

Apesar da era digital estar em pleno desenvolvimento, o armazenamento digital e automatizado dos dados estabelecem um ambiente frutífero para o avanço de pesquisas de padrões de nosso comportamento coletivo [133]. Ancorado em bancos de dados de grande escala, muitos trabalhos usam “medidas de consumo” como uma aproximação para a quantidade de atenção coletiva que um tipo de conteúdo cultural recebe. Como exemplo, temos o número de compartilhamentos de *hashtags* no Twitter [134, 135], venda de ingressos de filmes [125], citações de publicações científicas [136], repetições de músicas no Spotify [137] e o número de visualizações em páginas da Wikipédia [138, 139]. Assim, a ideia de atenção coletiva está intimamente relacionada à noção de uma ciência social computacional que investiga vários padrões de consumo de conteúdo em nível populacional [136, 140].

Neste contexto, os *videogames* são talvez o principal produto cultural da nossa era digital, com implicações de longo alcance para a cultura, especialmente por meio da gamificação de vários aspectos da vida moderna, do trabalho ao aprendizado [141]. Além disso, o que no passado era visto como brincadeira de criança é agora um passatempo difundido entre jovens e adultos, independentemente do gênero [142], tornando os jogos digitais uma forma cada vez mais popular de entretenimento e atividade de lazer. Aproveitando o poder do entretenimento, a indústria de jogos tem se mostrado muito lucrativa nos últimos anos, gerando bilhões de dólares [76]. Apesar dessa relevância cultural e econômica, a maioria dos trabalhos relacionados aos *videogames* frequentemente utilizam amostras pequenas e são qualitativos ou focam apenas em certas características dos jogadores e jogos [141]. Por exemplo, considerando trabalhos desenvolvidos a partir de experimentos controlados, os pesquisadores mediram a motivação das pessoas ao longo das etapas de um jogo [115]. Usando auto-relatos, estudos determinaram fatores relacionados ao ganho ou perda de motivação individual para continuar jogando [80, 83].

De uma forma mais quantitativa, existe uma literatura mais recente e emergente rela-

cionada aos *videogames* baseada em análises estatísticas de grandes quantidades de dados. Esses trabalhos, por exemplo, investigaram características comportamentais de jogadores em diferentes jogos e países [54, 143, 144], estudaram características textuais de resenhas de jogos [145, 146], e analisaram séries temporais de popularidade de jogos [147, 148]. Sabendo que o objetivo dos desenvolvedores de jogos é fazer com que as pessoas os joguem, vamos considerar a série temporal do número médio de jogadores por hora de um jogo como uma aproximação para sua popularidade, uma aproximação para a atenção coletiva recebida por ele, a fim de detectar grupos de jogos que exibem padrões de popularidade semelhantes. Para isso, adotamos uma abordagem de *clustering*, inspirada em trabalhos recentes que aplicam esses algoritmos para investigar diversas características de séries temporais de sistemas complexos relacionados a criptomoedas [33, 149], carreiras científicas [150] e sistemas dinâmicos [151, 152]. Além disso, utilizamos a base de dados da Steam, que fornece o número médio de jogadores por hora de cada mês para 5.840 jogos, a fim de detectar grupos de jogos cujas séries temporais de contagem de jogadores apresentam um padrão de evolução temporal semelhante. Destacamos a novidade do nosso trabalho na literatura científica sobre jogos digitais tendo em vista tanto a abordagem quanto a magnitude do conjunto de dados.

3.2 Conjunto de dados

Nossos resultados são baseados em dados de jogos da Steam. Como já foi dito no capítulo anterior, a Steam é uma plataforma *online* que vende e gerencia os direitos autorais de produtos digitais de sua base [91] e possui milhões de usuários espalhados pelo mundo [97], sendo talvez a principal fonte dos dados utilizados em investigações mais quantitativas envolvendo jogos e jogadores [143, 145, 146]. Como o serviço da Steam é completamente *online*, ele pode armazenar uma enorme quantidade de dados sobre seus usuários de forma automatizada. Nesse sentido, alguns sites de terceiros disponibilizam parte desses dados de forma aberta. Para obter nossos resultados, usamos os dados fornecidos pelo site SteamCharts [153]. O SteamCharts é um site lançado em julho de 2012 que publica a contagem média de jogadores por hora e por mês dos jogos da Steam. Por exemplo, se um jogo tem um valor de popularidade igual a 100 em um determinado mês, isso significa que, em média, 100 pessoas estavam jogando o jogo a cada hora durante o dado mês. Combinando a Steam e o SteamCharts, obtemos o nome, a data de lançamento, o gênero e a série temporal de popularidade dos jogos entre agosto de 2012 e julho de 2023. Filtrando jogos que estavam faltando algumas informações, jogos que foram lançados antes de agosto de 2012 e jogos com menos de 1 ano de dados de série temporal (jogos lançados após julho de 2022), restaram 5.840 jogos a serem estudados. A informação mais importante de cada jogo é a série temporal do número médio de jogadores por hora por mês, que chamamos de série de popularidade, ou simplesmente sua popularidade.

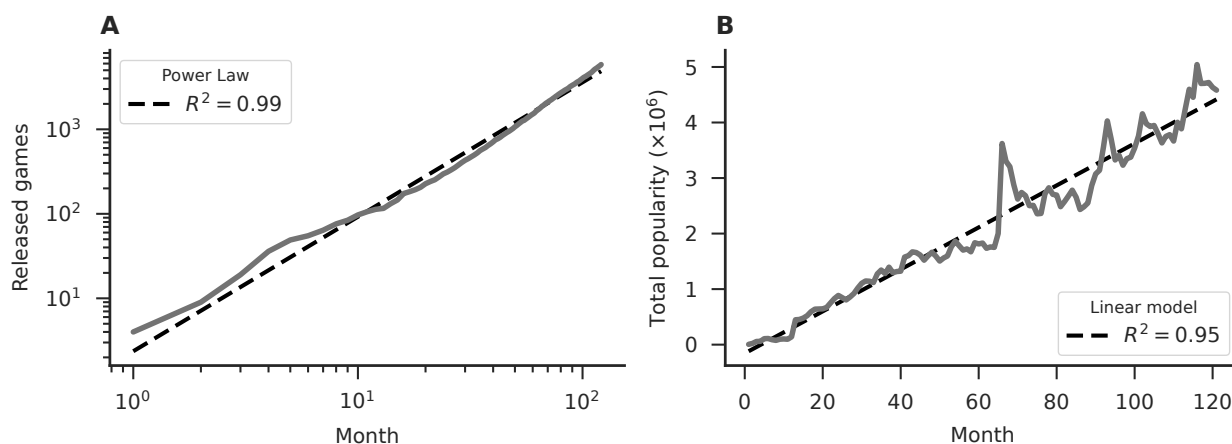


Figura 3.1: Evolução temporal do número de jogos e da popularidade total.

(A) O número acumulado de jogos lançados mensalmente na Steam é representado pela linha sólida enquanto a linha tracejada representa um ajuste de uma lei de potência com expoente igual 1,81. (B) A soma das popularidades dos jogos para cada mês é representada pela linha sólida e a linha tracejada representa um ajuste linear. Em ambos os painéis, os dados cobrem um período de 120 meses, de agosto de 2012 a julho de 2022.

Começamos a explorar nosso conjunto de dados com duas das principais variáveis relacionadas à Steam: o número de jogos e a popularidade total. A Figura 3.1A mostra o número cumulativo de jogos lançados de agosto de 2012 a julho de 2022. Como pode ser visto, o número total de jogos lançados exibe um crescimento superlinear que pode ser ajustado por uma lei de potência com muito boa aproximação. A Figura 3.1B retrata a evolução temporal da soma das popularidades mensais de todos os jogos na Steam. Essa métrica revela um crescimento aproximadamente linear na popularidade total dos jogos da Steam ao longo do tempo. Um aspecto interessante da dinâmica de popularidade da Steam é a flutuação, especialmente na segunda metade da curva, onde muitos picos distintos emergem de forma proeminente. Em particular, o pico mais distinto ocorre no mês 66 (janeiro de 2018), alcançando uma popularidade total de 3.621.887. Levando em consideração o período coberto por nossos dados, o máximo geral ocorre em março de 2022, alcançando um total de 5.043.435. De maneira geral, os resultados da Figura 3.1 destacam o crescimento da Steam nos últimos anos e dos jogos como um hobby ou atividade de lazer.

Investigando as mesmas variáveis em uma escala diferente, analisamos o número de jogos para cada gênero na Steam. Ao lançar um jogo, seu desenvolvedor o classifica dentro de alguns gêneros, similar ao que acontece com filmes. Existem 12 gêneros oficiais na Steam: Ação, Aventura, Casual, Acesso Antecipado³, Grátis para Jogar⁴, Indie, Multijogador Massivo (abreviado como Multijogador), Corrida, Jogo de Interpretação de Papéis (abreviado como RPG), Simulação, Esportes e Estratégia. Entre esses 12 gêneros, os desenvolvedores

³Do inglês, *Early Access*.

⁴Do inglês, *Free to Play*.

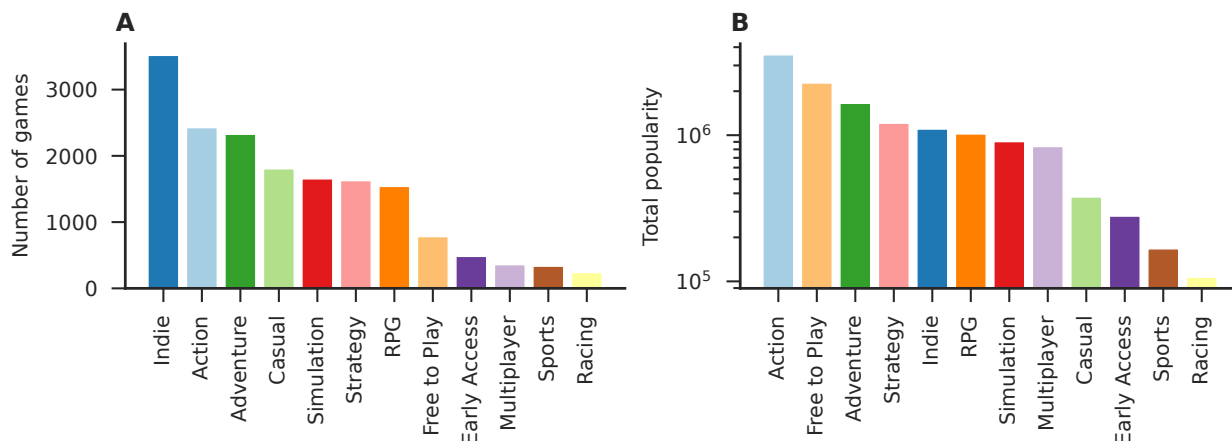


Figura 3.2: Número de jogos e jogadores relacionados a cada gênero de jogo na Steam. (A) Número acumulado de jogos lançados na Steam por gênero até julho de 2022. (B) Número de jogadores por gênero em julho de 2022. Para cada gênero, foi somada a popularidade dos jogos no mês de julho de 2022.

escolhem um ou mais para caracterizar seus jogos. Na Figura 3.2A, mostramos a quantidade de jogos lançados até julho de 2022 em cada gênero. O gênero com o maior número de jogos é Indie, que não caracteriza um estilo, mas indica jogos produzidos independentemente, tornando-se o gênero com o maior número de jogos lançados, já que são produzidos por pequenas empresas e, em geral, são jogos mais simples. Nessa linha de gêneros que não definem um estilo, existem também outros gêneros como Gratuito para Jogar (jogos gratuitos) e Acesso Antecipado (jogos disponíveis para jogar em sua fase beta, anterior ao lançamento oficial). Entre os gêneros que caracterizam aspectos de imersão, gêneros mais amplos como Ação e Aventura representam a maioria dos jogos. Por outro lado, o menor número de jogos lançados pertence ao gênero Corrida, provavelmente porque essa categoria é muito mais restrita e definida. Levando em conta a evolução temporal, todos os gêneros também cresceram com tendências superlineares, e uma grande parte dos gêneros pode ser bem aproximada por um modelo de lei de potência (veja a Figura 3.3).

Aprofundando a investigação por gênero, analisamos o número de jogadores por gênero para obter uma intuição sobre as preferências das pessoas. Na Figura 3.2B, mostramos a popularidade agregada dos jogos para cada gênero no mês de julho de 2022. Vale notar que um jogo com mais de um gênero contribui para a popularidade de todos os gêneros aos quais pertence. Apesar disso, percebemos que os jogadores têm uma forte preferência por jogos de Ação, o que possivelmente explica porque os desenvolvedores produzem muitos jogos desse gênero. Consistentemente com o que Indie representa, apesar de ter muitos jogos lançados, sua popularidade em relação a outros gêneros é mediana. Como já mencionado, produções independentes geralmente significam jogos mais simples que acabam não chamando a atenção dos jogadores. Por outro lado, há muito menos jogos Gratuitos para Jogar, mas eles alcançam um público muito mais amplo, sendo o segundo gênero mais po-

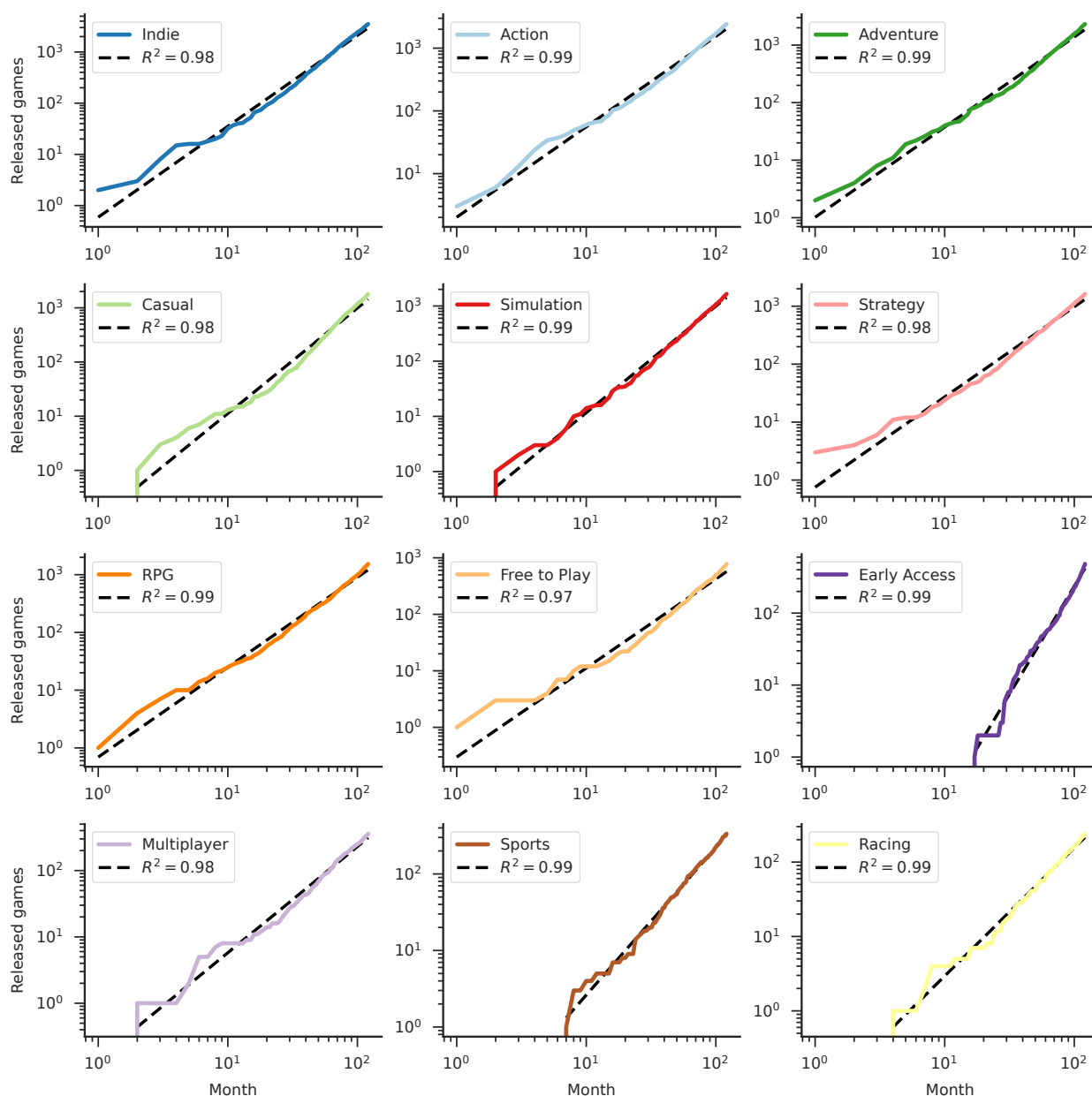


Figura 3.3: Evolução temporal do número de jogos de cada gênero. Número de jogos lançados para cada gênero na Steam ao longo dos 120 meses que compreendem o período de agosto de 2012 até julho de 2022. As linhas sólidas representam o dado enquanto as linhas tracejadas representam ajustes de leis de potência para cada gênero.

pular. Embora Ação tenha mais jogadores, os jogos Gratuitos para Jogar atingem uma taxa de engajamento muito alta, porque em números brutos, Ação tem 1.354 jogos e, em julho de 2022, tinha 3.970.486 de popularidade total, enquanto Gratuito para Jogar tinha 580 jogos e 2.504.271 de popularidade total durante o mesmo mês. Parece razoável pensar que jogos gratuitos têm maior alcance e, portanto, maior popularidade. No outro extremo, o gênero menos popular é Corrida e, analogamente, podemos reutilizar o mesmo argumento da especificidade do gênero. Examinando mais detalhadamente, a Figura 3.4 mostra a evolução

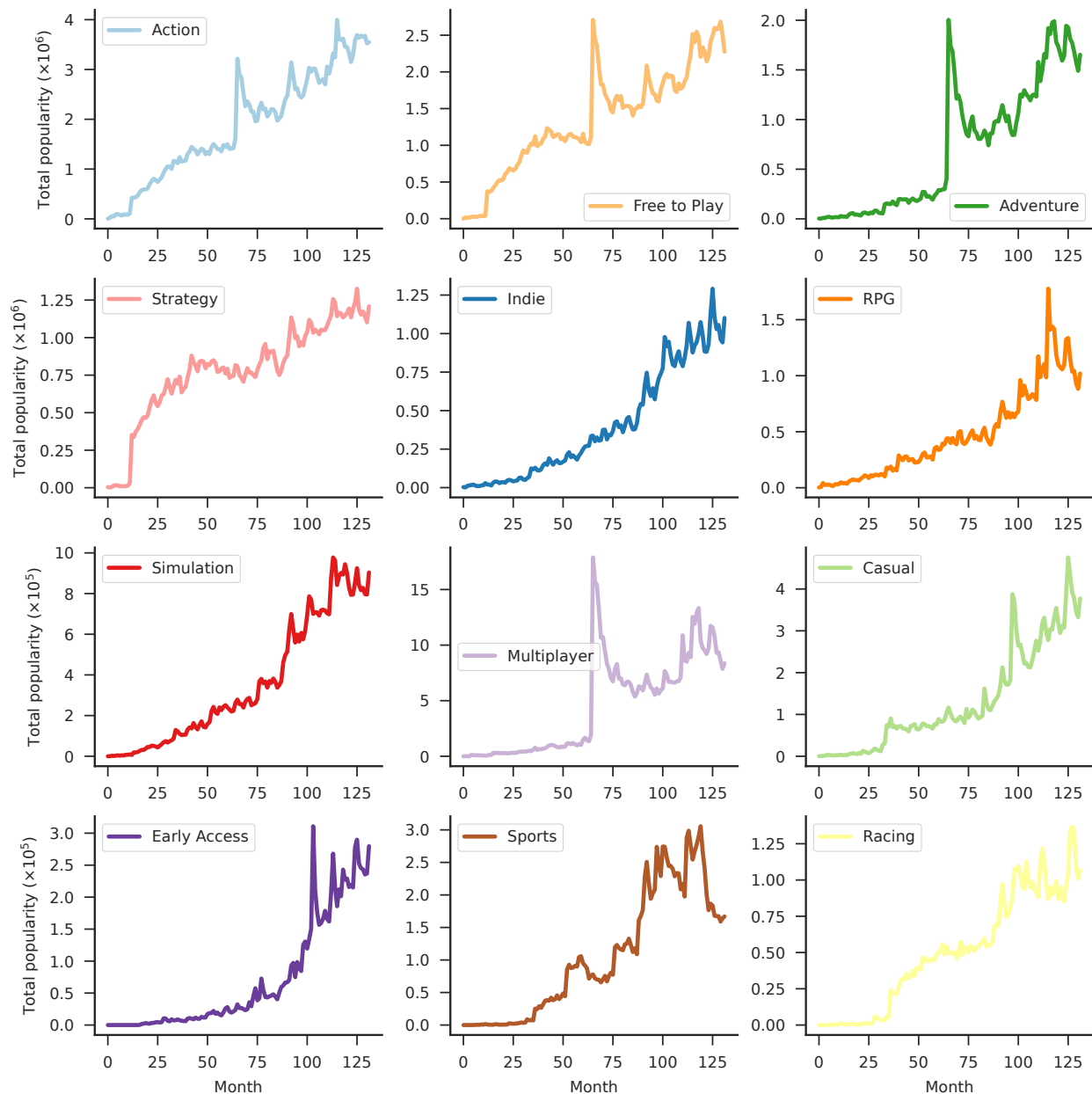


Figura 3.4: Evolução temporal do número de jogadores em cada gênero. Popularidade total em cada gênero da Steam ao longo dos 120 meses compreendendo o período de agosto de 2012 a julho de 2022.

temporal do número de jogadores em cada gênero. Apesar de todos os gêneros mostrarem aumento de popularidade, as características de crescimento são ligeiramente diferentes para cada um deles, com picos locais muito pronunciados se destacando. Restringindo-nos ao pico mais significativo na Figura 3.1B que ocorre no 66º mês (janeiro de 2018), observamos que é o resultado do lançamento oficial de um jogo chamado PUBG. Este jogo (pertencente aos gêneros Ação, Aventura, Gratuito para Jogar e Multijogador) atingiu a maior popularidade em um único mês em toda a base de dados, mas sua popularidade declinou rapidamente nos meses seguintes (veja a Figura 3.5).

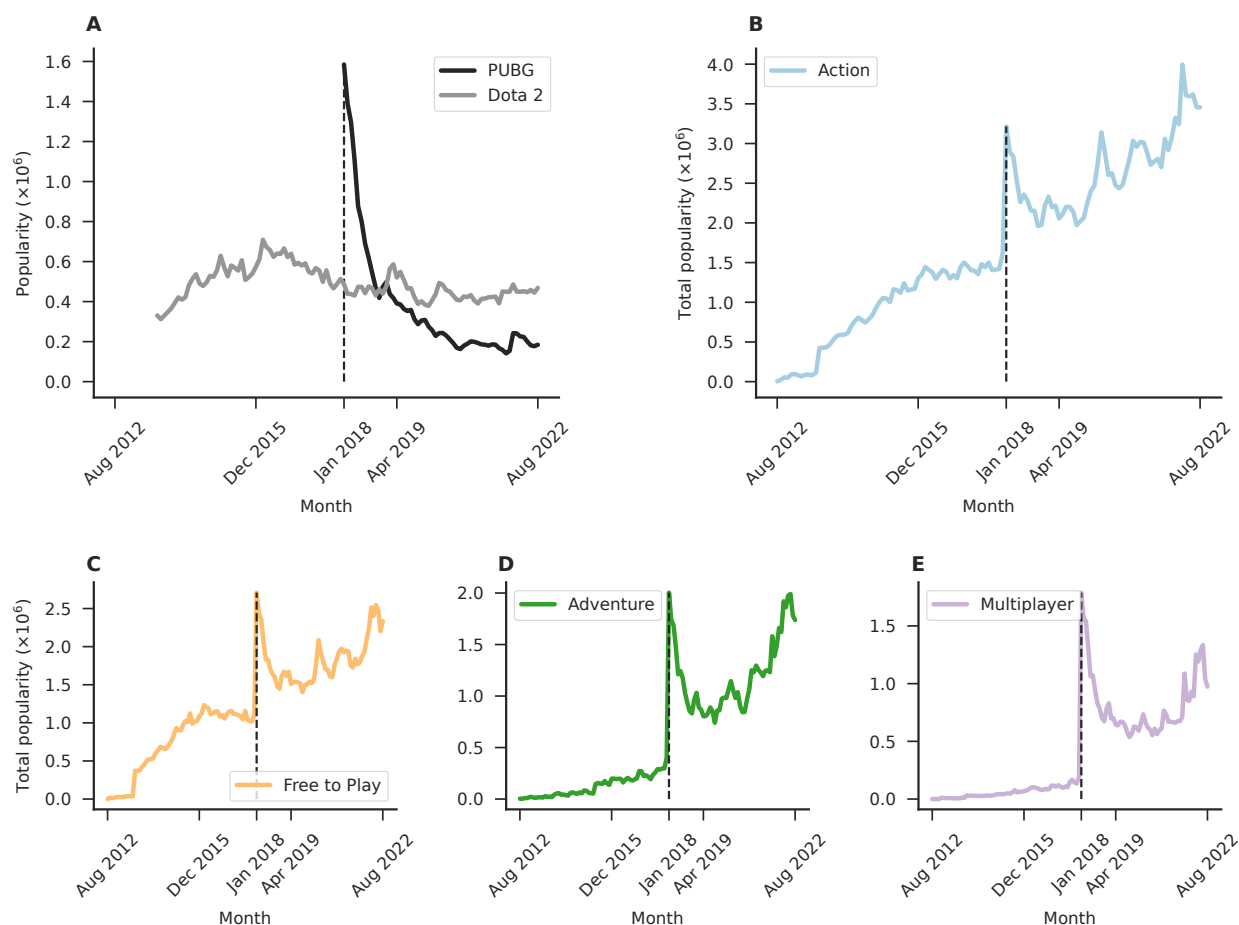


Figura 3.5: Impacto do lançamento do jogo PUBG. (A) As séries de popularidade dos jogos PUBG (black) e Dota 2 (cinza). O PUBG foi o jogo com mais jogadores em um único mês, alcançando seu pico de popularidade no mês do seu lançamento (janeiro de 2018). Em comparação, Dota 2 foi o segundo jogo com mais jogadores em janeiro de 2018. Os painéis (B)-(E) mostram o impacto do lançamento do PUBG na popularidade total dos gêneros que ele é caracterizado: Ação, Gratuito para jogar, Aventura e Multijogador.

Uma simples inspeção das séries de popularidade mostra que a maioria dos jogos tem poucos jogadores, enquanto alguns jogos concentram uma enorme massa de jogadores. Observando mais de perto, descobrimos que aproximadamente 73% das séries de popularidade têm uma média de menos de 100 jogadores ativos por hora por mês. Por outro lado, o jogo com a maior média de popularidade, Dota 2, teve cerca de 485.000 jogadores ativos por hora ao longo de sua história. A Figura 3.6A exibe a distribuição do número de jogos em função da média do número de jogadores, onde a distribuição apresenta uma cauda pesada. A linha tracejada representa uma distribuição log-normal com parâmetros $\mu = 3,35$ e $\sigma = 2,30$, o melhor ajuste aos dados de acordo com o teste estatístico de Kolmogorov-Smirnov [154] (veja a Tabela 3.1 para mais detalhes e outras distribuições testadas). A média é talvez o resumo estatístico mais comum calculado a partir de uma variável aleatória e geralmente é acompanhada por uma medida de dispersão, como o desvio padrão. Nesse sentido, também

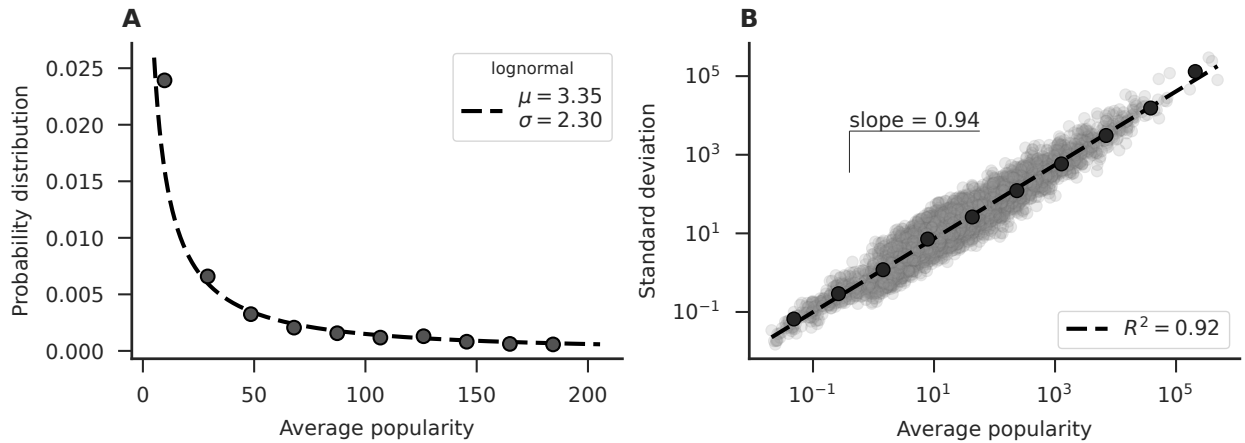


Figura 3.6: Resumo das séries temporais de popularidade dos jogos. (A) A distribuição de probabilidade do número de jogos como função da média de popularidade dos jogos. Os marcadores cinzas representam o dado e a linha tracejada representa a distribuição log-normal de parâmetros $\mu = 3,35$ e $\sigma = 2,30$. (B) Relação entre as médias e os desvios padrões das séries de popularidade. Os marcadores cinzas são os dados, os marcadores pretos são as médias tomadas em intervalos igualmente espaçados em escala logarítmica e a linha tracejada preta representa o modelo linear ajustado ao dado ($R^2 = 0,92$).

Distribuição	Distância de KS
Log-normal	0,0416
Weibull	0,1051
Beta	0,1594
Weibull dupla	0,3575
Gamma dupla	0,4047
Normal	0,4701

Tabela 3.1: Teste estatístico de Kolmogorov-Smirnov para diferentes modelos. A qualidade do ajuste para diferentes distribuições de popularidade modelando a média das séries de popularidade. Na primeira coluna apresentamos as distribuições testadas e na segunda coluna a distância de Kolmogorov-Smirnov. A distribuição log-normal tem o melhor ajuste, com a menor distância de Kolmogorov-Smirnov.

calculamos o desvio padrão para cada série de popularidade e mostramos sua relação com a média. Na Figura 3.6B, vemos, como tendência geral, que quanto mais a média de popularidade aumenta, mais o desvio padrão aumenta. Interessantemente, de acordo com nosso ajuste (representado pela linha preta tracejada), o crescimento é aproximadamente linear, já que a inclinação é $0,94 \approx 1$ na escala log-log, possivelmente sugerindo um processo anômalo de “difusão de atenção” [155].

3.3 Métodos

Em um próximo passo, caracterizamos a forma dos padrões de séries temporais de popularidade de *videogames*, ou seja, a maneira como uma série de popularidade evolui ao longo do tempo, sem levar em conta sua magnitude, apenas sua forma. Dessa maneira, agrupamos séries que apresentam formas semelhantes, mesmo que apresentem contagens médias de jogadores por hora em diferentes escalas. Tendo esse objetivo em mente, transformamos cada série de popularidade em um *z-score*, ou seja, subtraímos os valores de cada série por sua média e os dividimos pelo respectivo desvio padrão. Assim, usamos séries cuja variação é dada em unidades de desvio padrão em relação à média da própria série. No entanto, as séries dos *z-scores* são muito ruidosas, apresentando uma grande variação no número de jogadores de um mês para o outro. Por causa disso, aplicamos um *kernel* de suavização à série, chamado de filtro gaussiano [156], para reduzir flutuações aleatórias e tornar o padrão da série mais evidente. O filtro gaussiano depende do parâmetro σ , e este parâmetro está relacionado ao grau de suavização. Quanto maior o valor de σ , mais suave a série se torna. Na Figura 3.7 mostramos exemplos de séries de popularidade ao lado de suas séries de *z-score* e filtros gaussianos com $\sigma = 1, 2$ e 3 . Para $\sigma = 1$, algumas séries ainda permanecem bastante ruidosas; para $\sigma = 2$, as séries se tornam suaves sem perder sua forma, enquanto para $\sigma = 3$, as séries são tão suaves que perdem a amplitude de variação e alguns detalhes importantes da dinâmica. Levando essas observações em consideração, escolhemos $\sigma = 2$.

A partir dessas séries de *z-score* suavizadas, calculamos a matriz de distâncias entre elas usando o algoritmo *Dynamic Time Warping* (DTW) [157, 158]. A distância DTW é uma medida de dissimilaridade baseada na forma e isso nos permite comparar séries de tamanhos diferentes criando um caminho de deformação não linear entre elas, proporcionando mais flexibilidade para combinar sequências que apresentam padrões semelhantes. Após calcular a matriz de distâncias, aplicamos o método *Uniform Manifold Approximation and Projection* (UMAP) [159, 160] (com parâmetros padrão) para realizar uma redução de dimensionalidade. Em resumo, a partir da matriz de dissimilaridade, o UMAP realiza uma redução de dimensionalidade com base em fundamentos matemáticos de geometria Riemanniana e topologia algébrica. O UMAP primeiro cria um complexo simplicial difuso, que pode ser representado como um grafo ponderado, e então projeta os dados em um espaço de menor dimensão através de um algoritmo de *layout* de grafo dirigido por forças. Assim, o algoritmo cria uma representação em rede da matriz de dissimilaridade, onde os nós representam jogos e arestas ponderadas conectam jogos com dinâmica de popularidade semelhante. Essa técnica de redução de dimensionalidade é projetada na forma de uma rede complexa, organizando e ligando os nós de forma a equilibrar o arranjo respeitando as estruturas locais e globais. Aqui, no entanto, dispensamos o grafo associado ao espaço de baixa dimensão e aplicamos o algoritmo Infomap para identificar a estrutura de comunidade da primeira rede criada

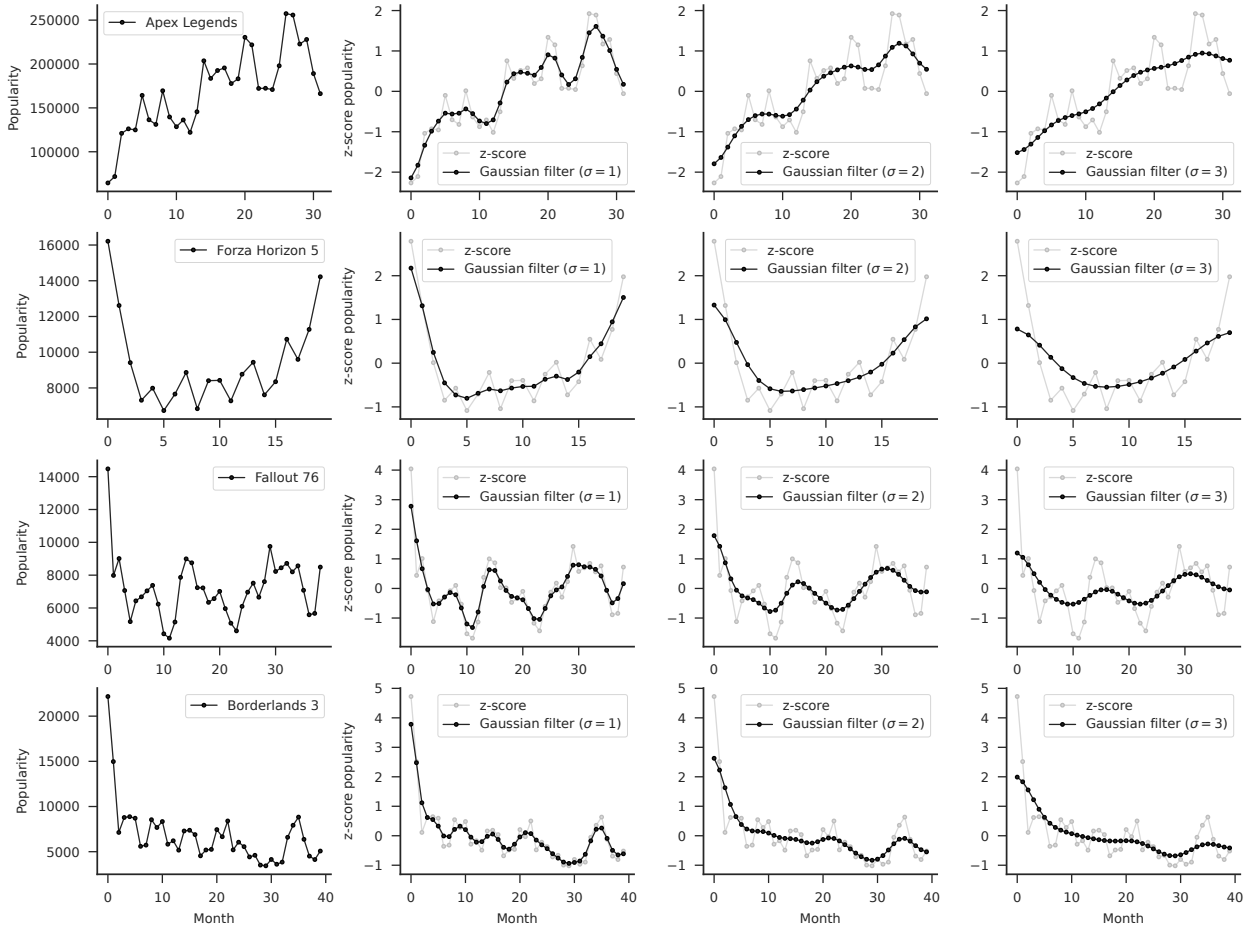


Figura 3.7: Exemplos de séries de popularidade de jogos da Steam e o processo de padronização dessas séries. A primeira coluna mostra a série de popularidade de quatro jogos: Apex Legends, Forza Horizon 5, Fallout 76 e Borderlands 3. A segunda, terceira e quarta coluna mostram as séries suavizadas usando o filtro gaussiano com $\sigma = 1$, $\sigma = 2$ e $\sigma = 3$, respectivamente. Nas últimas três colunas, os marcadores em cinza representam as séries do *z-score*, enquanto os marcadores pretos correspondem às séries do *z-score* suavizadas.

pelo UMAP. O Infomap é uma técnica de detecção de comunidades em redes baseada em conceitos de teoria da informação e que utiliza passeios aleatórios como uma aproximação para o fluxo de informação na rede. Dentro do algoritmo é calculada a chamada *map equation* [161–163] para determinar os limites teóricos de quão concisamente é possível descrever um passeio aleatório infinito na rede com uma determinada atribuição de nós a comunidades. Ao minimizar a *map equation*, o Infomap revela a estrutura de comunidades da rede. Como o algoritmo Infomap se baseia no fluxo probabilístico de passeios aleatórios pela rede, diferentes execuções podem levar à detecção de comunidades distintas. Para mitigar esse efeito e garantir resultados mais robustos, realizamos dez mil execuções do algoritmo Infomap. Nessas execuções, detectamos números de comunidades (aqui também chamadas de *clusters*) variando de 3 a 7, sendo 5 o número mais recorrente de comunidades detectadas (para mais

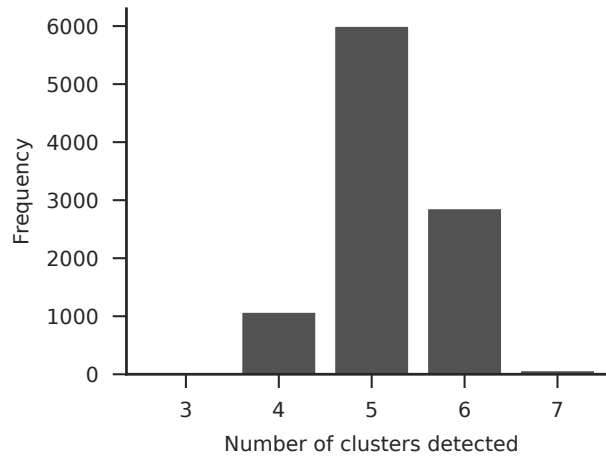


Figura 3.8: Frequência do número de *clusters* detectados pelo algoritmo Infomap.

Como o algoritmo é sensível a *input seed* (semente de entrada), foi executado dez mil vezes usando diferentes sementes. O histograma mostra a frequência do número de *clusters* detectados em todas as execuções do algoritmo. A detecção de 5 *clusters* foi a moda e ocorreu 5.999 vezes.

detalhes, veja a Figura 3.8). Entre essas dez mil execuções, escolhemos aquela que expressa a melhor partição da rede em *clusters* com base no coeficiente de silhueta [164] calculado com os valores da matriz de distâncias DTW. Este coeficiente quantifica a consistência do procedimento de agrupamento, dando uma medida da similaridade dos elementos do mesmo *clusters* e da dissimilaridade dos elementos de diferentes *clusters*. De modo geral, quanto maior o valor médio do coeficiente de silhueta, melhor a configuração dos agrupamentos encontrados.

Finalmente, após todos esses passos, agrupamos séries de popularidade cujas curvas têm formatos semelhantes. Essa mesma rotina de algoritmos foi recentemente proposta por Sunahara *et al.* [150] como uma modificação do algoritmo de Lee *et al.* [165].

3.4 Resultados

O melhor agrupamento de séries de popularidade e a forma média das séries pertencentes aos *clusters* que encontramos são mostradas na Figura 3.9. No total, detectamos 5 *clusters*, cada um representando uma evolução diferente de popularidade. O Cluster 1 (azul), que representa 47,4% dos jogos, exibe um padrão predominante de decaimento, deixando claro que uma parte significativa dos jogos atinge seu pico de popularidade nos primeiros dias após o lançamento e depois gradualmente perde popularidade ao longo do tempo. O Cluster 2 (vermelho), representando 20,7% dos jogos, apresenta o padrão mais irregular entre todos os *clusters*. A característica mais distintiva deste *cluster* é seu padrão ondulante, mostrando múltiplos picos de popularidade ao longo do tempo. No apêndice C, expomos algumas figuras que detalham nossa análise sobre a dinâmica de popularidade dos jogos. Por exemplo, na

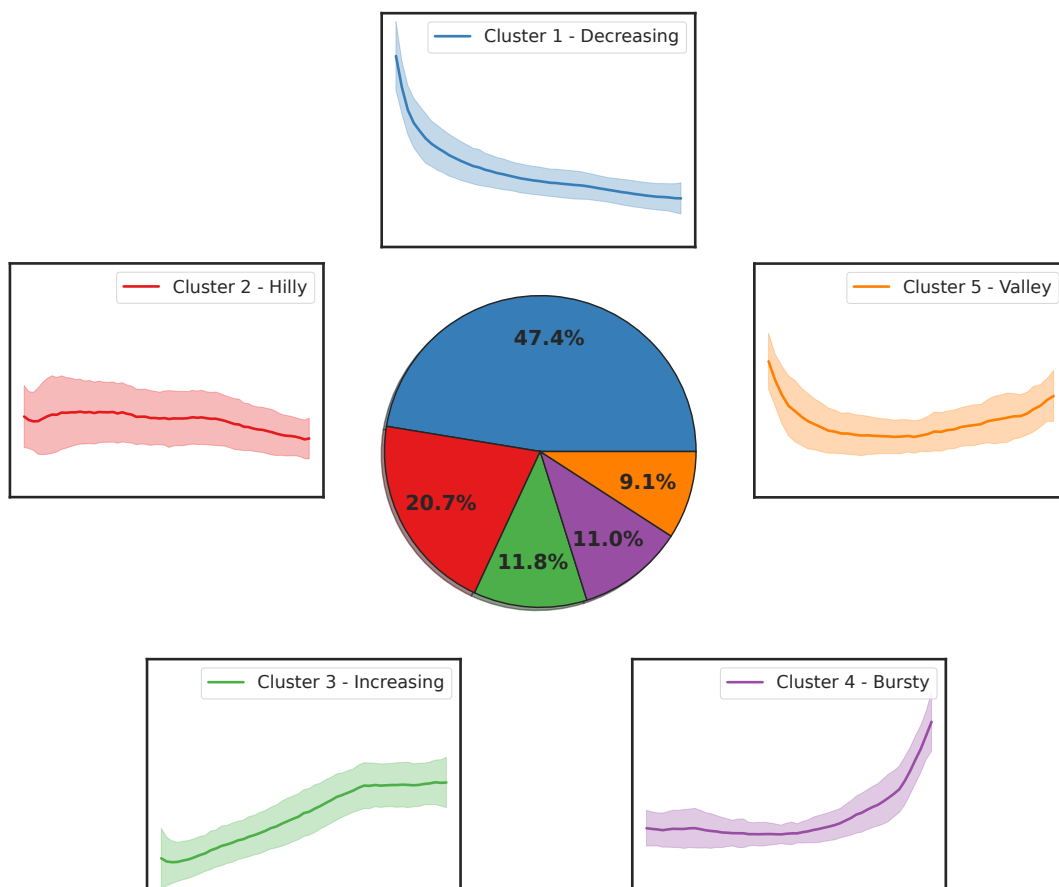


Figura 3.9: Formas das séries de popularidade dos jogos da Steam. O painel central mostra a proporção de jogos dentro de cada grupo que foi detectado pelo algoritmo de agrupamento: 47,4% dos jogos estão no Cluster 1 (decrecente, em azul); 20,7% no Cluster 2 (montanhoso, em vermelho); 11,8% no Cluster 3 (crescente, em verde); 11,0% no Cluster 4 (explosivo, em roxo); 9,1% no Cluster 5 (vale, em laranja). Os painéis ao redor do centro mostram a forma média das séries temporais dentro de cada *cluster* e uma faixa de desvio padrão.

Figura C.1, mostramos duas séries de popularidade típicas para cada *cluster*. Apesar das duas séries ilustradas como exemplos do Cluster 2 parecerem ter uma tendência de declínio, algumas séries mostraram uma tendência aproximadamente constante e, em outros casos, pode haver até um leve crescimento. Assim, observamos que a característica mais distintiva das séries no Cluster 2 são os múltiplos picos de popularidade que ocorrem de tempos em tempos. O Cluster 3 (verde), 11,8% dos jogos, exibe um padrão de crescimento consistente e duradouro. Além de apresentar um padrão de crescimento, os jogos no Cluster 3 têm a maior média de popularidade (ver Figura C.2). No entanto, vale notar que, em um grande número de jogos nesse *cluster*, o crescimento inicial é mais acelerado comparado ao último terço da série temporal. De fato, para muitos jogos, há uma saturação nos últimos meses (ver exemplos na Figura C.1). O Cluster 4 (roxo), 11,0% dos jogos, consiste em jogos estáveis e relativamente desconhecidos em sua maioria; no entanto, a partir de um certo

mês, eles exibem um aumento repentino de popularidade. O Cluster 5 (laranja), 9,1% dos jogos, exibe o padrão mais inesperado, inicialmente decaindo e depois seguido por uma recuperação na popularidade. Devido a esse padrão de popularidade que se assemelha a um vale, hipotetizamos que jogos nesse cluster teriam séries mais longas, pois isso permitiria que esses jogos fossem esquecidos e depois lembrados como “bons jogos”. No entanto, a Figura C.3 mostra o contrário; o Cluster 5 exibe a maior concentração de densidade de probabilidade em valores mais baixos de comprimento de séries, com a menor mediana entre todos os *clusters*.

A representação dos gêneros de jogos varia entre os *clusters*. Para examinar esse aspecto, primeiro calculamos a diferença entre a proporção de jogos de um determinado gênero dentro de cada *cluster* e a proporção geral de jogos do mesmo gênero (veja a Figura 3.10). Além disso, dividimos essa diferença pela proporção geral de jogos pertencentes ao gênero. A Figura 3.10 ilustra essa diferença na representação dos gêneros dentro de cada *cluster*. Jogos Indie e Casuais estão sobrerrepresentados no Cluster 4, sugerindo que uma parte desses jogos atinge “sucesso” mais tarde. Por outro lado, Simulação e Estratégia estão sub-representados no Cluster 4, sugerindo que seus padrões de popularidade são mais consistentes ao longo do tempo, sendo, portanto, mais raro que fenômenos de “bela adormecida” ocorram nesses gêneros. Enquanto isso, RPG parece ser um gênero de jogo propenso a ser completamente esquecido; está sobrerrepresentado no Cluster 1 (decrecente) e sub-representado nos outros *clusters*, especialmente no Cluster 3 (crescente). Curiosamente, jogos Free to Play estão sobrerrepresentados no Cluster 1 e sub-representados nos outros também; no entanto, são jogos muito populares (Figura 3.2). Isso sugere que, em muitos casos, esses jogos têm um início meteórico e depois gradualmente desaparecem. A descoberta mais surpreendente foi a notável representação dos jogos de Acesso Antecipado no Cluster 3. Jogos Multiplayer exibem dinâmicas estáveis de crescimento ou declínio de popularidade, já que esses jogos estão sobrerrepresentados nos Clusters 1 e 3 e sub-representados nos outros. Os dois gêneros mais nichados, Esportes e Corrida, estão sobrerrepresentados no Cluster 3, demonstrando o potencial de sua força atrativa em capturar novos jogadores. Os dois gêneros que não exibiram sobrerrepresentação nem sub-representação notável em nenhum *cluster* foram Ação e Aventura.

Em um próximo passo, analisamos se os padrões de popularidade estão relacionados ao comprimento das séries. Para isso, dividimos as séries de popularidade em quatro grupos de mesmo tamanho com base nos quartis da distribuição dos comprimentos (conforme mostrado na Figura 3.11). Os quatro grupos englobam jogos cujas séries têm entre 12 a 27 meses de duração, 28 a 49 meses, 50 a 74 meses e jogos com séries mais longas que 74 meses. A Figura 3.12 mostra como cada grupo de comprimento é composto por jogos pertencentes a diferentes *clusters* de popularidade. Observamos que os percentuais de jogos com padrões de forma de popularidade decrescente (Cluster 1) e vale (Cluster 5) diminuem à medida

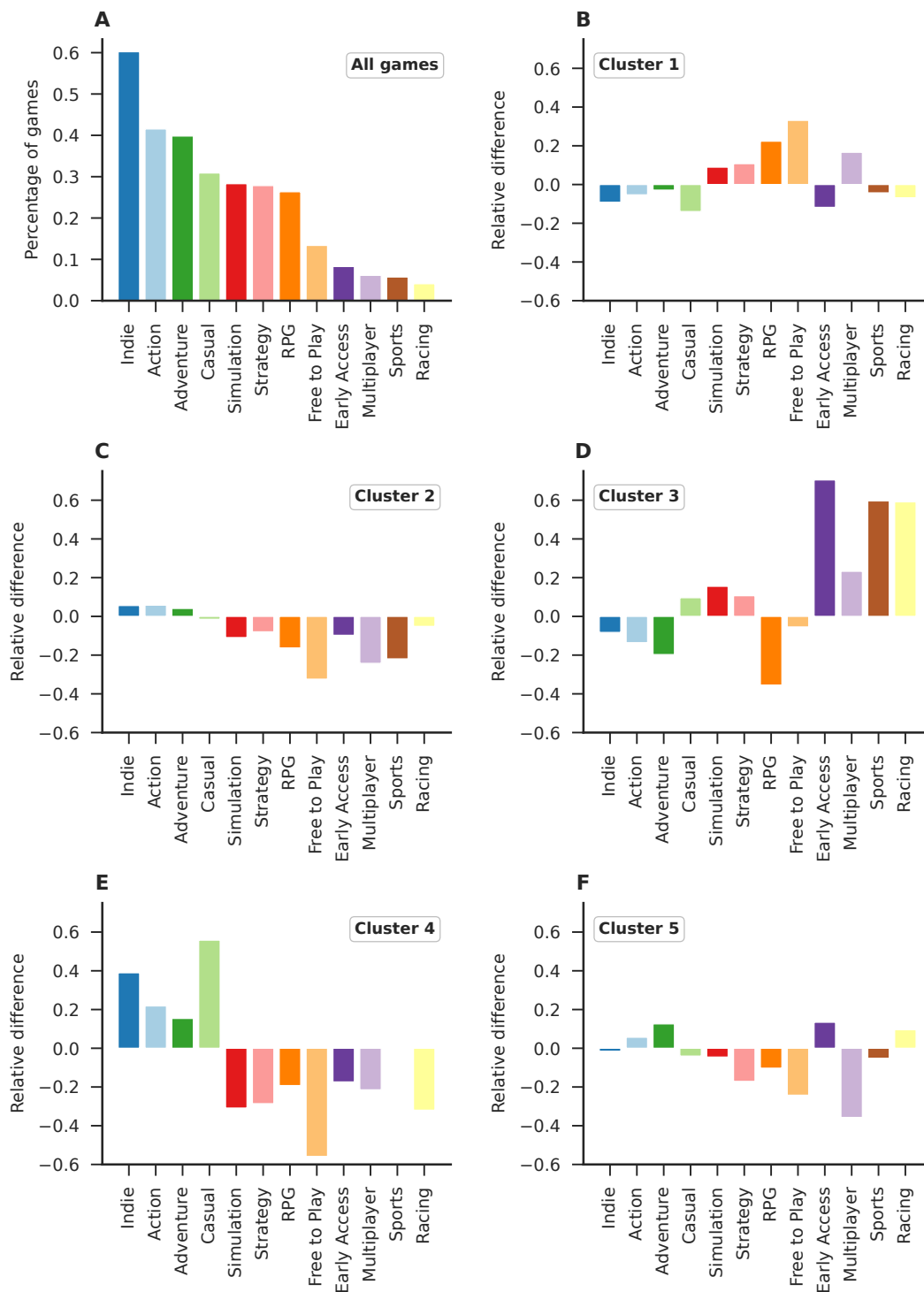


Figura 3.10: Representação dos gêneros e suas proporções relativas nos *clusters* detectados. (A) O percentual de jogos em nosso banco de dados pertencentes a um ou mais dos 12 gêneros oficiais da Steam. As diferenças relativas na representação dos gêneros de jogos para jogos pertencentes aos *clusters* de 1 a 5 são mostradas nos painéis (B)-(F).

que suas séries se tornam mais longas. No entanto, o padrão decrescente permanece o mais recorrente, independentemente do comprimento da série. O padrão montanhoso (Cluster 2)

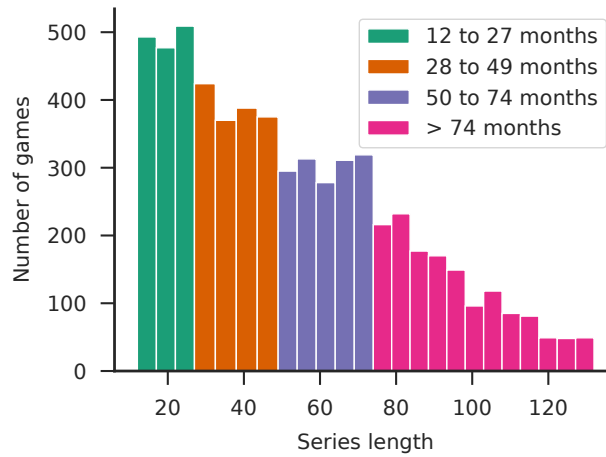


Figura 3.11: Histograma da frequência de jogos em função da duração das séries de popularidade. Com base neste histograma, os jogos foram divididos em quatro grupos: séries entre 12 a 27 meses de duração (verde), 28 a 49 meses (laranja), 50 a 74 meses (roxo) e mais de 74 meses (rosa).

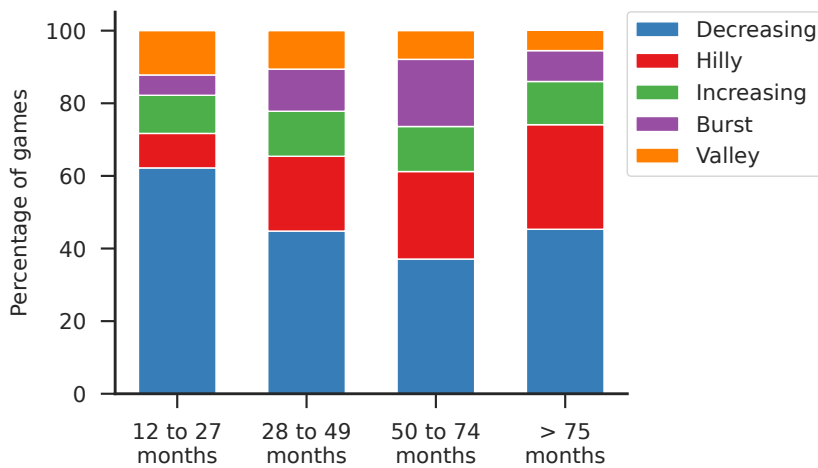


Figura 3.12: Representação dos padrões de forma detectados em jogos agrupados por tamanho da série. Para cada grupo de tamanho de série, o percentual de jogos que exibem cada um dos cinco padrões de forma detectados colore uma parte da barra vertical.

vê um aumento na representação à medida que as séries se tornam mais longas. Surpreendentemente, o padrão explosivo (Cluster 4) mostra um percentual significativamente maior de séries entre 50 a 74 meses de duração. Enquanto isso, o padrão crescente (Cluster 3) mantém seu percentual em diferentes comprimentos, sugerindo que “jogos de sucesso” são lançados a uma taxa fixa.

Na parte final da nossa análise, investigamos a prevalência e a persistência dos padrões de forma de popularidade ao longo do tempo. Em outras palavras, para todos os jogos, comparamos o padrão de popularidade nas fases iniciais com o da série inteira. Considerando o grupo de séries mais antigas (com mais de 75 meses), empregamos o método de detecção

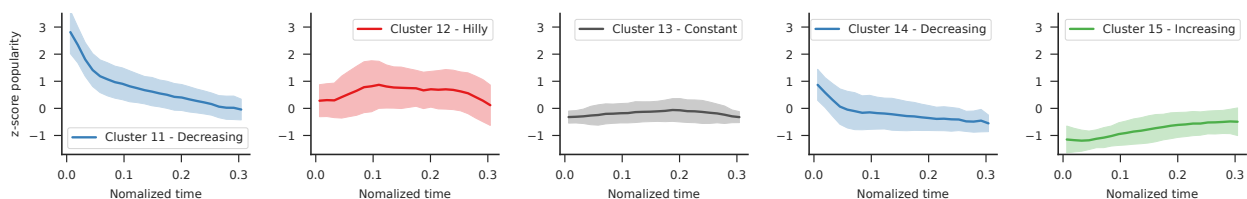


Figura 3.13: Padrões das séries de popularidade para os primeiros 24 meses de jogos com séries de popularidade superiores a 74 meses. O painel exibe o comportamento médio dos *clusters* detectados pela melhor realização do algoritmo Infomap de acordo com o coeficiente de silhueta. Dois *clusters* exibiram o mesmo padrão: os Clusters 11 e Cluster 14 mostraram um padrão decrescente (azul). Outros dois padrões foram detectados novamente, montanhoso (Cluster 12, vermelho) e crescente (Cluster 15, verde). O Cluster 13 (cinza) apresentou um novo padrão constante com variação relativamente pequena em torno da média.

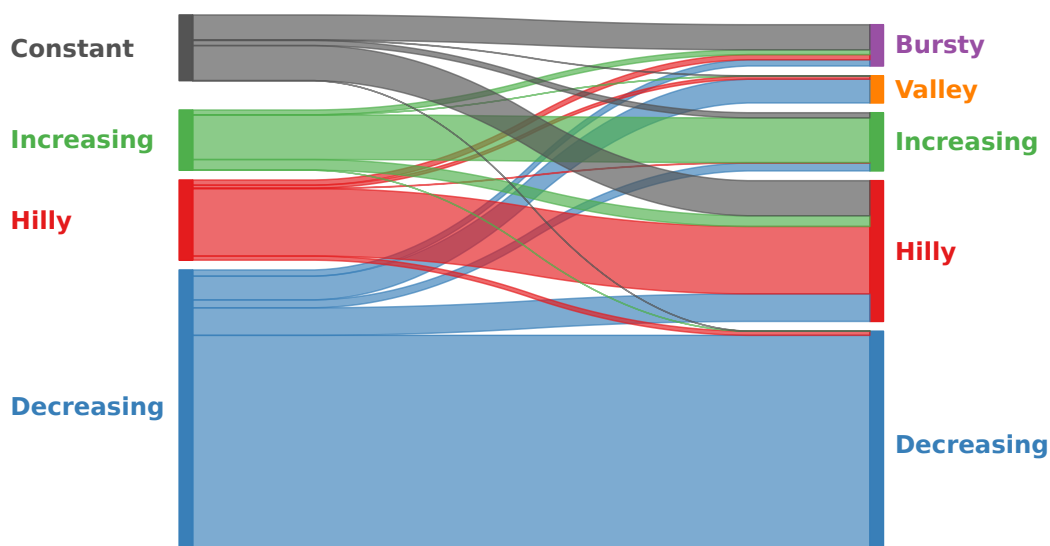


Figura 3.14: Comparação entre a prevalência e a persistência dos padrões de forma nas fases iniciais e finais das séries de popularidade. As barras à esquerda mostram a fração de jogos pertencentes a cada *cluster* de padrão nos primeiros 24 meses, enquanto as barras à direita mostram a prevalência desses padrões de forma considerando a série de popularidade completa. Conexões entre os padrões à esquerda e à direita significam a migração de um padrão de forma para outro à medida que as séries de popularidade evoluem ao longo do tempo. Além disso, a espessura das conexões representa a proporção de jogos.

de *clusters* mencionado anteriormente (seção 3.3) para identificar os padrões de forma de popularidade dos primeiros 24 meses. A Figura C.4 mostra a frequência de *clusters* detectados após dez mil realizações do algoritmo Infomap. Enquanto isso, a Figura 3.13 exibe os *clusters* detectados usando a melhor realização de acordo com o coeficiente de silhueta. Nestes *clusters* detectados, três padrões de forma já foram exibidos: decrescente (Cluster 11

e Cluster 14), montanhoso (Cluster 12) e crescente (Cluster 15). A novidade foi o surgimento de um padrão constante com pouca variância (Cluster 13). Os padrões vale e explosivo não foram detectados. Conectando os resultados da análise de *clusters* para as fases iniciais com os achados apresentados anteriormente para a série inteira, medimos a persistência dos padrões. Restringindo nossa análise apenas a jogos com mais de 75 meses, a Figura 3.14 mostra os padrões de forma de popularidade exibidos pelos jogos considerando os primeiros 24 meses (barras à esquerda) e a série inteira (barras à direita). Descobrimos que 77% dos jogos mantêm o padrão decrescente, 84% permanecem montanhoso e 74% permanecem crescentes. As mudanças mais significativas ocorrem em jogos que exibem padrões de forma constante e decrescente no início de suas séries de popularidade. Jogos com padrão constante basicamente se tornam montanhosos (53%) e explosivos (37%), enquanto 10% dos decrescentes tornam-se montanhosos e 8% tornam-se vales. Exemplos de séries de popularidade de jogos que sofrem mudanças de padrão ao longo do tempo podem ser vistos na Figura 3.15.

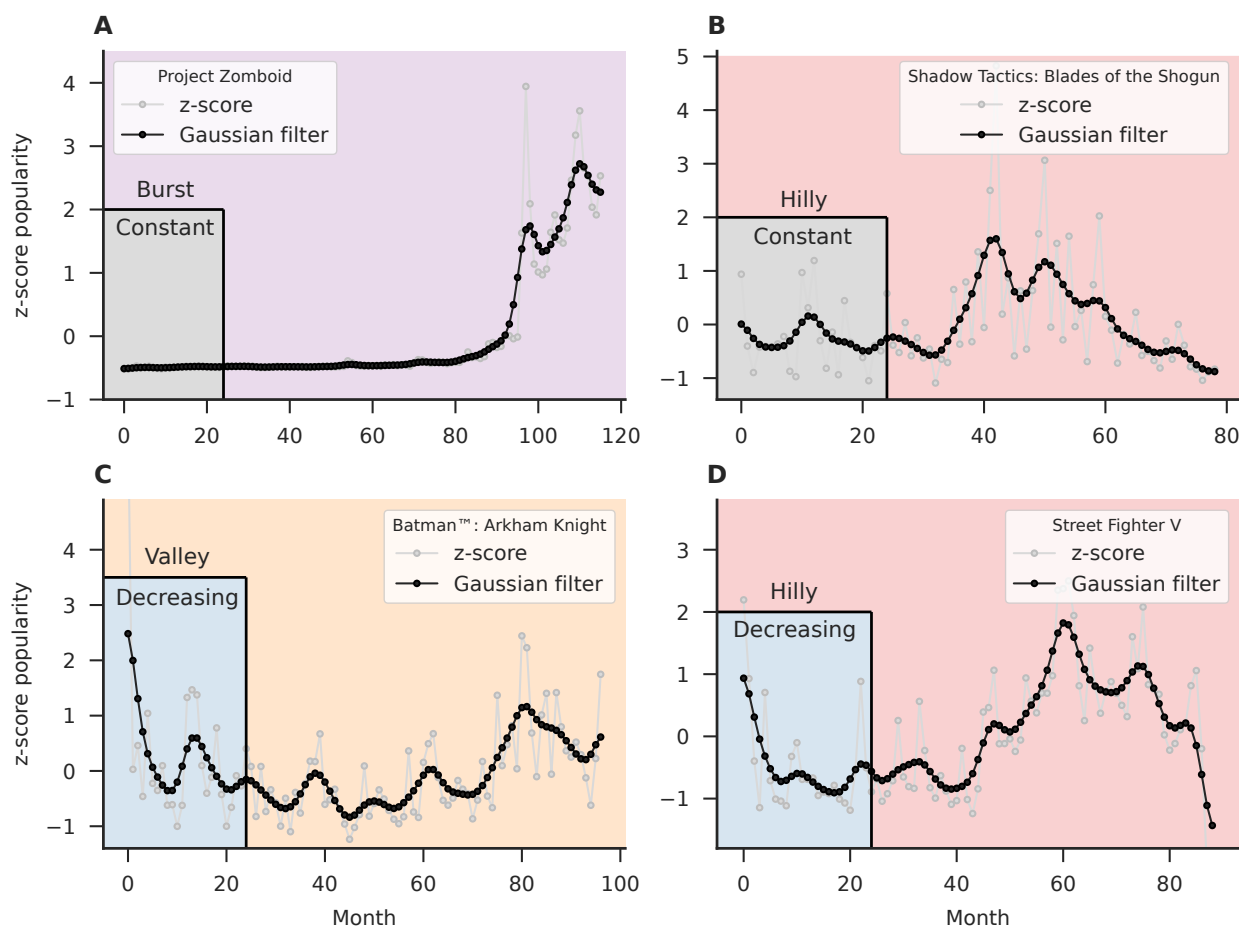


Figura 3.15: Exemplos de transição de padrão de forma. Exemplos de quatro jogos (nomeados dentro dos painéis) cuja forma da série de popularidade transita entre (A) constante e explosivo, (B) constante e montanhoso, (C) decrescente e vale e (D) decrescente e montanhoso.

3.5 Discussão

Apresentamos um estudo quantitativo sobre jogos digitais e seus jogadores. Especificamente, utilizamos uma abordagem de análise de *cluster* para detectar padrões na dinâmica da popularidade dos jogos. Para isso, usamos a média horária de jogadores por mês para 5.840 jogos da Steam. Em nossa análise exploratória, focamos nos números gerais da Steam. A Figura 3.1 mostra o crescimento maciço da plataforma, seguindo aproximadamente um crescimento superlinear (com expoente de lei de potência igual a 1,81) para o número de jogos e um crescimento linear para o número de jogadores em função do tempo. Até o presente momento, não há sinal aparente de saturação, sugerindo que esse ecossistema ainda tem potencial para crescer ainda mais tanto em termos de jogos quanto de jogadores. A partir de uma inspeção visual da Figura 3.1B, podemos observar a ocorrência de picos de popularidade bem definidos na segunda metade da série. Embora não tenhamos calculado estatísticas específicas para esses picos, observamos que eles ocorrem com mais frequência nos meses de janeiro e julho, que são tipicamente meses de férias. Vale notar que as mesmas conclusões que apresentamos ao longo deste parágrafo são aplicáveis ao analisar todos os gêneros (ver Figuras 3.3 e 3.4).

Em um esforço para caracterizar os aspectos mais gerais das séries de popularidade dos jogos, calculamos suas médias e desvios padrão. A Figura 3.6A exibe a distribuição de probabilidade do número de jogos em função da popularidade média ajustada por uma log-normal. Esse resultado é consistente com a distribuição de probabilidade da popularidade de outros tipos de conteúdo cultural [166, 167], sugerindo a existência de um padrão para a popularidade média. Embora possa ser intuitivo supor que a distribuição da popularidade média seja Gaussiana, especialmente considerando o teorema do limite central [168], acreditamos que os fatores determinantes para ganhar ou perder popularidade não são aditivos, mas multiplicativos (como argumentamos no capítulo 2). Em outras palavras, podemos inferir que a atenção coletiva das pessoas é fragmentada entre os jogos disponíveis e, portanto, esse processo de fragmentação gera uma distribuição log-normal [54, 148]. Interessantemente, variáveis aleatórias assimétricas com caudas pesadas, como é o caso da distribuição log-normal, frequentemente exibem eventos raros e extremos, conhecidos como *outliers* [169]. Essa característica é consistente com nossas descobertas, onde há uma diferença significativa entre alguns jogos extremamente populares e o restante. No entanto, a Figura 3.6B mostra que a dispersão em torno da média das séries de popularidade é quase sempre da mesma magnitude que a própria média, tornando a média um descritor pouco confiável para a dinâmica da popularidade dos jogos.

Neste cenário incerto em relação à média, focamos na forma das séries de popularidade dos jogos. Sabendo que as formas podem variar amplamente (veja a Figura 3.7), adotamos uma abordagem de *clustering*. Conforme descrito na seção 3.3, as séries de popularidade

de cada jogo foram normalizadas usando *z-score* e, em seguida, suavizadas com um filtro gaussiano. Apesar de fixarmos $\sigma = 2$ no filtro gaussiano, resultados semelhantes poderiam ser obtidos usando $\sigma = 1$ ou $\sigma = 3$. No total, 5 *clusters* foram detectados (veja a Figura 3.9), os quais nomeamos de acordo com o padrão de suas formas: decrescente (Cluster 1), montanhoso (Cluster 2), crescente (Cluster 3), vale (Cluster 4) e explosivo (Cluster 5). O ponto mais interessante em nossa análise de detecção de *clusters* é a conexão com outros resultados já documentados na literatura. Primeiramente, objetos culturais passam por um processo de esquecimento, onde sua popularidade exibe um decaimento temporal [96]. Para os jogos, este resultado também é apoiado por distribuições altamente assimétricas de tempo de jogo na Steam [143]. Nesse contexto, sugerimos que o Cluster 1 consiste em jogos cuja dinâmica de popularidade é essencialmente governada pelo decaimento temporal. Lembrando que quase metade dos jogos em nosso banco de dados estão no Cluster 1, um processo de decaimento temporal pode descrever a dinâmica de popularidade de quase metade dos jogos. Em segundo lugar, ao avaliar o Cluster 3, sugerimos que sua dinâmica de popularidade é em grande parte resultado do mecanismo de *preferential attachment* (anexação preferencial), onde popularidade gera popularidade [170]. Este *cluster* representa um pequeno grupo de jogos (menos de 12%) com médias de popularidade mais altas e comportamento continuamente crescente ao longo do tempo. Na mesma linha de raciocínio, acreditamos que o padrão do Cluster 3 é um forte candidato para descrever a popularidade dos jogos *outliers*. À primeira vista, a detecção de jogos *outliers* pode parecer insignificante, mas o problema dos *outliers* é um dos mais antigos na estatística [171]. Em terceiro lugar, o crescimento tardio e acelerado da popularidade nos jogos do Cluster 4 se assemelha à dinâmica esperada das “belas adormecidas”, estudadas pela primeira vez com séries temporais de citações de publicações científicas. Nesse contexto, trabalhos que permanecem em um período de hibernação e depois experienciam uma popularidade meteórica são conhecidos como “belas adormecidas” [172]. Finalmente, se considerarmos intuitivamente que a popularidade dos jogos crescerá até um valor de saturação e depois diminuirá para um valor mínimo (possivelmente zero) [173], os Cluster 2 e Cluster 5 exibem os padrões mais inesperados. O Cluster 2 representa jogos com séries de popularidade mais imprevisíveis e irregulares, com oscilações que obscurecem tendências médias em alguns casos. Conjecturamos que essas oscilações na popularidade podem estar, por exemplo, associadas a períodos de férias escolares e promoções de vendas de jogos com descontos. O Cluster 5 apresenta um padrão de dinâmica de popularidade totalmente inesperado, desafiando nossa intuição, começando com um certo nível de popularidade que declina e é gradualmente recuperado após algum tempo.

Também revelamos como os gêneros estão associados aos padrões de forma detectados. Isso pode ser útil para entender quais jogos têm mais probabilidade de alcançar popularidade crescente ou decrescente com base em seus gêneros (veja a Figura 3.10). Nesse contexto, alguns gêneros merecem atenção especial. Jogos Indie estão super-representados no Clus-

ter 4, sugerindo que muitos jogos Indie bem-sucedidos ganham popularidade tardiamente. Acreditamos que, por serem jogos de baixo orçamento, eles podem ter dificuldade em atrair atenção por meio de grandes campanhas de *marketing*. No entanto, após algum tempo, a popularidade pode crescer por meio do boca a boca. Jogos de acesso antecipado estão super-representados no Cluster 3, e nossa hipótese para esse fato está relacionada ao seu processo de desenvolvimento. Esses jogos são disponibilizados para serem jogados mesmo quando incompletos e são atualizados pelos desenvolvedores. Em muitos casos, os desenvolvedores atualizam os jogos por meses ou até anos com base no *feedback* dos jogadores [145]. Acreditamos que o resultado desse processo de desenvolvimento são jogos que agradam melhor aos seus jogadores. Jogos Multiplayer estão super-representados nos Cluster 1 e Cluster 3, mostrando que suas trajetórias tendem a ser monotônicas, aumentando ou diminuindo. Por definição, esses jogos exigem que as pessoas joguem juntas e acreditamos que isso reforça o boca a boca, gerando um efeito bola de neve tanto para o crescimento quanto para a queda da popularidade.

Finalmente, demonstramos a persistência dos padrões de forma comparando a parte inicial de uma série de popularidade com a forma da série completa. Avaliando os primeiros 24 meses, detectamos quatro padrões de forma: decrescente, montanhoso, crescente e constante. Esses quatro padrões dão origem aos cinco padrões de forma já discutidos (veja a Figura 3.14). Como regra geral, os padrões tendem a persistir ao longo do tempo, mostrando a consistência dos padrões de forma detectados para caracterizar a dinâmica da popularidade dos jogos. O único padrão que se extingue ao longo do tempo é o padrão constante, que evolui para montanhoso (quando a variação aumenta) ou explosivo (quando há um aumento repentino na popularidade).

Do ponto de vista econômico, a popularidade dos jogos não é uma medida direta de sucesso financeiro. Isso se deve ao fato de que a monetização depende da estratégia adotada por cada desenvolvedor, com dois modelos principais: jogos que precisam ser comprados e jogos gratuitos com conteúdo exclusivo pago [174]. Há também o caso de jogos que se monetizam de forma híbrida, onde é necessário comprar o jogo, e há conteúdos exclusivos que podem ser comprados dentro do jogo. Portanto, além de contribuir para o debate sobre a dinâmica da popularidade de objetos culturais, nossas descobertas podem ser úteis para a indústria de jogos, pois entender a dinâmica da popularidade pode orientar os desenvolvedores a como melhor investir em seus jogos.

Em conclusão, nossa pesquisa fornece resultados inéditos sobre a popularidade dos *videogames*, acrescentando à literatura de investigações baseadas em dados sobre esse produto cultural. Considerados os resultados obtidos, acreditamos que nosso trabalho contribui para o interesse atual na existência de padrões universais na dinâmica de popularidade de conteúdo cultural. Além disso, nosso trabalho pode inspirar investigações adicionais sobre as características dos padrões que definem os *clusters* de séries de popularidade em outros

contextos.

Efeitos de ambientes sombreados no comportamento de peixes

Apesar de esforços significativos serem empregados para entender a complexidade e a emergência dos movimentos coletivos em grupos de animais, o papel de diferentes configurações externas neste tipo de movimento permanece em grande parte inexplorado. Neste capítulo, combinando análise de séries temporais e ferramentas de redes complexas, apresentamos uma investigação do efeito da sombra no comportamento de uma espécie de peixe, Carpa Prateada (*Hypophthalmichthys molitrix*) [53]. Encontramos que a sombra encoraja a residência dos peixes durante as horas do dia, mas o grau de preferência para a sombra varia entre ensaios e lagoas. Durante o dia e sob a sombra, os peixes são muito mais lentos e exibem menos persistência em suas velocidades. Essa diferença de comportamento é ainda mais evidente quando comparamos com as velocidades dos peixes durante a noite, período em que os peixes mostraram maiores velocidades e persistência. Além disso, nossa pesquisa mostra que a sombra afeta significativamente o comportamento de cardumes desses peixes reduzindo sua polarização, número de interações intra-individuais e a formação de relacionamento de tríade estável. No entanto, os peixes mantêm um maior grau de ordem local sob a sombra quando em comparação com o período noturno.

4.1 Introdução

Movimento coletivo é um comportamento comum de espécies de animais que se reúnem em bandos, manadas ou cardumes [175, 176]. Esses movimentos coletivos rendem uma gama de padrões espaciais e temporais, que são bem conhecidos por exibir espetáculos visuais na natureza [177]. Além de uma beleza contemplativa, entender a complexidade e a emergência do movimento coletivo em grupos de animais têm sido por muitos anos parte da vida de

pesquisadores de diversas áreas da ciência. Esforços significantes têm sido empregados para propor e investigar modelos capazes de conectar regras de nível individual aos fenômenos macroscópicos observados [178–182]. No entanto, os avanços em abordagens têm sido, em grande parte, desacompanhados de observações empíricas criando uma lacuna entre teoria e experimento que só recentemente começou a diminuir. De fato, estudos recentes baseados em novos métodos para rastrear animais em grupos abriram caminho para questões empíricas que seriam consideradas impraticáveis alguns anos atrás [183–188]. Exemplos incluem a identificação de vizinhos influentes de peixes no movimento do grupo [189], efeitos de predação na interação de cardumes de peixes [190] e na fuga coletiva de pombos [191], emergência de movimento giratório [192], e o papel das interações sociais nos ciclos de vida de abelhas [193].

Apesar de interessantes trabalhos já terem sido conduzidos, muitos deles são focados em uma única configuração externa e pouco se sabe sobre os efeitos de diferentes configurações externas no comportamento coletivo dos animais [194]. Nesse contexto, uma possibilidade interessante refere-se à seleção de estruturas suspensas que criam sombra. Muitas espécies têm mostrado uma afinidade por estruturas com sombra que podem também desempenhar um papel na dispersão e estabelecimento de espécies invasoras [195]. De fato, existe evidência empírica que a sombra aumenta o tempo de residência de algumas espécies de peixe. Por exemplo, o Bluegill (*Lepomis macrochirus*), um peixe de água doce nativo dos Estados Unidos, é curiosamente conhecido por se abrigar à sombra de árvores ao longo das margens e, de fato, exibe uma forte preferência por áreas sombreadas, mesmo na presença de predadores [196]. Trabalhos recentes têm mostrado que a carpa prateada (*Hypophthalmichthys molitrix*) passa grande parte do tempo perto de uma eclusa e de uma represa no rio Mississippi durante o dia [197], sugerindo que a sombra pode encorajar os peixes. As estruturas de sombra são também conhecidas por atrair espécies de peixes marinhos para habitats de mangue [198], aumentar o número de salmões (*Oncorhynchus kisutch*) em canais [199] e aumentar o uso de determinados habitats por diferentes espécies, como o esgana-gato (*Gasterosteus aculeatus*) [200] e o peixe branco do lago (*Coregonus clupeaformis*) [201]. A preferência por sombra é frequentemente associada com baixo risco de predação não apenas por peixes [196, 199], mas também por pássaros [202] e roedores [203, 204]. Ao mesmo tempo, algumas espécies como peixe-zebra (*Danio rerio*) [200] não exibem uma preferência clara por sombra, enquanto outras espécies como Rei Salmão (*Oncorhynchus tshawytscha*) parecem perceber a sombra como um ambiente de risco [205].

Considerando o interesse no entendimento da preferência por sombra em muitas espécies e a escassez de trabalhos focados em estudar o comportamento coletivo em diferentes contextos, aqui investigamos como a sombra afeta várias facetas dos comportamentos individuais e coletivos da carpa prateada. Contamos com experimentos em larga escala em quatro lagoas onde a posição de pequenos cardumes (~10 indivíduos) de carpas prateadas

são rastreados usando telemetria acústica¹ em sete ensaios de 48 horas. Nesses experimentos, um ambiente sombreado é criado ancorando uma grande plataforma flutuante dentro de cada lagoa, nos permitindo quantificar a seleção da sombra e, mais importante, como o comportamento desses peixes diferem quando sob ou fora da estrutura de sombra durante o dia e a noite. A carpa prateada exibe uma preferência pela estrutura de sombra durante o dia, mas a intensidade dessa preferência varia ao longo dos ensaios e lagoas. Nossa pesquisa mostra que a carpa prateada é significativamente mais lenta sob a sombra. No entanto, ela atinge seus maiores picos de velocidade no período da noite. A persistência do movimento também decresce quando a carpa prateada está sob a sombra e aumenta durante a noite. Além disso, a sombra afeta significativamente o comportamento do cardume desses peixes reduzindo sua polarização, número de interações intra-individuais, e a formação de relacionamentos de tríades estáveis. Ainda assim, descobrimos que a carpa prateada preserva um grau intermediário de ordem local quando sob sombra, acompanhando as posições de seus vizinhos, mesmo em estados semelhantes a enxames.

No que segue, descrevemos o experimento usado para obter as trajetórias dos peixes, seguido pela análise da preferência pelas áreas com sombra dessa espécie. Depois, quantificamos o efeito da sombra na velocidade dos peixes e o grau de persistência na série de velocidade dos peixes. Assim, focando no comportamento coletivo, estimamos dois parâmetros de ordem relacionados ao grau de alinhamento e rotação do cardume de carpas prateadas. Finalmente, criamos uma representação em redes complexas da interação dos peixes baseada nas suas relações espaciais para destacar como a sombra afeta a relação entre as carpas prateadas.

4.2 Métodos

Os experimentos e o conjunto de dados usados neste estudo compreendem sete ensaios que ocorreram em quatro lagoas no Columbia Environmental Research Center, Columbia, Missouri, EUA. As lagoas são de aproximadamente 40 metros de comprimento, 25 de largura e de 1 a 2 metros de profundidade. Os experimentos começaram com dez carpas prateadas jovens ($345,2 \pm 17,5$ mm de tamanho total e $377 \pm 64,7$ g de peso) em cada lagoa. Todos os indivíduos foram equipados com transmissores acústicos (795-LD; Innovasea Systems Inc., Boston, MA) implantados cirurgicamente na cavidade celomática (20 indivíduos) ou fixados externamente ao peixe (20 indivíduos). O trabalho com animais está em acordo com todos os padrões aplicáveis para a ética da experimentação e integridade da pesquisa e aprovado pelo Comitê Institucional de Cuidados e Uso de Animais do Serviço Geológico dos EUA no Centro de Pesquisa Ambiental de *Columbia* (número IACUC: AEH-18-CERC-01). As

¹A telemetria acústica baseia-se na detecção de sinais sonoros emitidos por pequenos transmissores que são implantados nos animais

posições dos peixes foram registradas em média a cada 1,3 segundos usando uma matriz de telemetria acústica (Innovasea Systems Inc.) em cada lagoa. Uma plataforma flutuante de 5 metros quadrados foi colocada em uma das cinco posições possíveis em cada lagoa (linhas tracejadas nas Figuras 4.1A, 4.1B e D.1). Os peixes foram colocados na lagoa e ficaram 24 horas para se acostumar com a lagoa e a posição da sombra. Após isso o ensaio começava e eram registradas as posições dos peixes por 48 horas ininterruptas (começando 00:00 e terminando às 23:59 do dia seguinte). A plataforma flutuante era movida para uma nova posição em cada ensaio, e dessa forma, eram respeitadas 24 horas de intervalo para os peixes se ambientarem. Essa plataforma cria um ambiente de sombra embaixo de si que imita uma estrutura de sombra real encontrada em habitats de peixes.

Filtramos todas as posições recebidas pelos transmissores em cada ensaio examinando a variabilidade da posição (desvio padrão de incrementos de posição consecutivos) de cada peixe em uma janela deslizante de uma hora. As trajetórias com variabilidade de posição inferiores a 20% da variabilidade geral de cada peixe foram inspecionadas visualmente. As trajetórias nas quais percebemos um padrão de queda ou defeito do transmissor foram totalmente removidas das análises. O conjunto de dados final compreendeu mais de 24 milhões de posições dos indivíduos que permaneceram com os transmissores até o ensaio final, como detalhado na Tabela 4.1. No fim do sétimo ensaio as lagoas foram drenadas e todos os 40 peixes foram recuperados vivos.

Para estimar a série temporal das velocidades, primeiramente segmentamos as trajetórias dos peixes em duas classes: dentro e fora da estrutura de sombra. Então, selecionamos todos os segmentos de trajetória com mais de 50 observações e construímos a série temporal da velocidade calculando incrementos de posições consecutivas e dividindo pelo intervalo de tempo correspondente. Finalmente, essas séries de velocidade foram separadas em três grupos: sob a sombra durante o dia; fora da sombra durante o dia; e durante a noite.

Estimamos o expoente de Hurst h de segmentos da série de velocidade $v(t)$ via *detrended fluctuation analysis* (DFA) [206, 207]. O DFA consiste dos seguintes passos. Primeiro, definimos a série integrada $Y(t) = \sum_{i=1}^t [v(i) - \langle v(t) \rangle]$, em que $\langle v(t) \rangle$ é a velocidade média do segmento da série. Depois, amostramos $Y(t)$ em $m_n = m/n$ partições não sobrepostas de tamanho n , em que m é o comprimento da série temporal. Para cada partição, ajustamos um modelo linear e subtraímos de $Y(t)$, definindo o perfil sem tendência na escala n como $Y_n(t) = Y(t) - p_j(t)$, em que $p_j(t)$ representa o ajuste linear na j -ésima partição. Finalmente calculamos a função de flutuação quadrática média $F(n) = \sum_{j=1}^{m_n} \langle Y_n(t)^2 \rangle_j / m_n$, em que $\langle Y_n(t)^2 \rangle_j$ é o valor quadrático médio de $Y_n(t)$ em todo o dado da j -ésima partição. Para séries temporais autossimilares, $F(n)$ mostra uma dependência tipo lei de potência com a variável n , e então $F(n) \sim n^h$, em que h é o expoente de Hurst. Dessa forma, estimamos h calculando a inclinação da lei de potência linearizada ($\ln F(n) \sim h \ln n$) via métodos dos mínimos quadrados (seção 1.11). Valores de $h > 0,5$ indicam comportamento de longa

Ensaio	Lagoa	Número de peixes	Número de detecções
1	26	10	1.296.149
2	26	10	1.280.127
3	26	8	1.056.631
4	26	8	1.068.860
5	26	7	939.736
6	26	4	545.645
7	26	4	541.051
1	27	10	1.257.650
2	27	8	1.047.315
3	27	7	953.318
4	27	7	883.221
5	27	8	1.044.909
6	27	7	973.654
7	27	7	970.958
1	30	9	1.161.989
2	30	8	933.214
3	30	6	799.124
4	30	5	688.476
5	30	5	713.752
6	30	5	720.318
7	30	5	708.362
1	31	7	940.619
2	31	8	1.123.747
3	31	7	940.349
4	31	5	659.301
5	31	3	420.820
6	31	3	419.256
7	31	3	415.454

Tabela 4.1: **Resumo do conjunto de dados.** Número de detecções de posição em cada ensaio e lagoa do conjunto de dados final. Os ensaios com menos de 10 peixes são aqueles que apresentaram padrão de queda ou defeito do transmissor em um ou mais peixes.

persistência, enquanto que $h \approx 0,5$ representa ausência de correlação ou correlação de curto alcance. Também determinamos que o expoente de Hurst da versão embaralhada da série das velocidades é aproximadamente 0,5.

Para calcular os parâmetros de ordem, interpolamos a trajetória de cada peixe usando o método de vizinhos mais próximos e reamostramos essas séries temporais com uma resolução de 4 segundos para sincronizar as trajetórias de todos os peixes em cada ensaio e lagoa.

Usando essas trajetórias sincronizadas, o parâmetro de ordem de polarização é definido por [208, 209]

$$O_p = \frac{1}{N} \left| \sum_{i=1}^N \vec{u}_i \right|, \quad (4.1)$$

em que $\vec{u}_i$ é o vetor unitário da direção da velocidade do i -ésimo peixe em um dado tempo para um ensaio e lagoa particulares com N peixes rastreados. Os valores de O_p são restritos ao intervalo $[0, 1]$ e $O_p \approx 1$ representa a configuração dos peixes completamente alinhados, enquanto $O_p \approx 0$ indica a situação sem alinhamento. Por sua vez, o parâmetro de ordem de rotação é calculado por [208, 209]

$$O_r = \frac{1}{N} \left| \sum_{i=1}^N \vec{u}_i \times \vec{r}_i \right|, \quad (4.2)$$

em que $\vec{r}_i$ é o vetor unitário apontando do centro do cardume para a posição do i -ésimo peixe. Essa quantidade representa o valor médio normalizado do momento angular do peixe, ou seja, pode variar no intervalo $[0, 1]$. O limite inferior de O_r representa a situação com nenhuma rotação, enquanto que o valor superior indica uma configuração de forte rotação.

Das trajetórias sincronizadas definidas no parágrafo anterior, construímos uma rede de interação. Para um dado ensaio e lagoa com cinco ou mais peixes, selecionamos todas as posições em um tempo particular t e criamos o diagrama de Voronoi. Esse diagrama é um particionamento do plano de posições dos peixes em polígonos convexos em que cada polígono contém a posição de um peixe, e todos os pontos dentro do polígono são mais próximos do seu peixe particular do que dos demais peixes (veja Figura 4.4A). Usando o diagrama de Voronoi no tempo t , criamos um grafo $g_t = (\nu, \varepsilon)$, em que os vértices ν correspondem ao peixe e o conjunto de arestas ε conecta todos os pares de indivíduos que são vizinhos na tesselação de Voronoi.

Construindo esses grafos g_t para cada tempo t (amostrado em intervalos de 4s) e combinando-os em intervalos de tempo não-sobrepostos de tamanho Δt , criamos uma rede pesada $G_\tau = (V, E, W)$. Nessa rede pesada, τ representa o ponto central no intervalo de tempo, os vértices V representam peixes, o conjunto de arestas E conecta todos os pares de indivíduos que foram vizinhos pelo menos uma vez no intervalo de tempo Δt , e W é o conjunto de pesos de arestas representando a fração de tempo que um par de indivíduos permaneceu vizinhos durante o intervalo de tempo dividido pelo número de peixes menos um. Fixamos $\Delta t = 10$ min, no entanto utilizando diferentes valores de Δt obtivemos resultados semelhantes (veja Figura D.3). Além disso, consideramos a rede para representar a categoria sob a sombra quando mais de 10% das posições dos peixes estão localizadas sob as estruturas de sombra dentro do intervalo de tempo (na Figura D.4 mostramos que os resultados são invariantes independente da escolha do limiar de 10%).

Por último, da rede pesada $G_\tau = (V, E, W)$, calculamos duas grandezas para o melhor entendimento da relação entre os indivíduos do cardume. Primeiramente, o peso κ associado a um vértice é a soma de todos os pesos de suas arestas incidentes [210]. Logo, o grau pesado médio k é a média dos valores dos pesos de todos os vértices da rede. Depois calculamos a medida de transitividade da rede. Em seguida, calculamos a transitividade pesada da rede, que mede o quanto os vértices de um grafo se agrupam. Para grafos simples e não direcionados, a transitividade (ou o coeficiente global de agrupamento) é definida como a razão entre o número de triádes fechadas e o número total de triádes (abertas e fechadas) [210]. Essa medida quantifica a probabilidade geral da rede ter vértices adjacentes interconectados, assim revelando a existência de comunidades fortemente conectadas. Existem muitas generalizações de transitividade para redes pesadas [211], mas para esse trabalho usamos a definição de Onnela *et al.* [212] que substitui o número de triádes fechadas na definição não pesada pela soma normalizada das intensidades das triádes (em que a intensidade refere-se à média geométrica do peso das arestas), como implementado no pacote NetworkX [213] escrito na linguagem Python.

4.3 Resultados

Para começar a investigar os efeitos da sombra no comportamento da carpa prateada, analisamos o dado da posição das carpas em cada ensaio e lagoa. As Figuras 4.1A e 4.1B mostram o mapa de densidade para a posição de dois peixes de diferentes lagoas e ensaios durante o dia e a noite. No apêndice D expomos outras figuras que detalham nossa análise sobre o comportamento das carpas prateadas. Por exemplo, na Figura D.1 mostramos o mapa de densidade para a posição de todas as lagoas e ensaios. Observamos que um desses peixes mostra forte preferência pela estrutura de sombra durante as horas do dia, enquanto que o outro peixe não parece ter a mesma preferência. Variabilidade semelhante ocorre quando consideramos a posição de todos os peixes e avaliamos a seleção de sombra calculando a fração de tempo que eles passam sob estruturas de sombra ao longo das horas do dia. As Figuras 4.1C e 4.1D exibem essa análise para todos os peixes de uma mesma lagoa e ensaio usados para criar o mapa de densidade. Notamos que o comportamento individual dos peixes nas Figuras 4.1A e 4.1B concorda com as frações de tempo gastas sob a estrutura de sombra por seus cardumes. Os peixes da lagoa e do ensaio da Figura 4.1C gastaram muito do seu tempo sob a sombra durante o dia e evitaram a estrutura de sombra à noite. Contrariamente, os peixes da lagoa e do ensaio da Figura 4.1D nunca selecionaram a estrutura de sombra entre a meia noite e o meio dia e além disso raramente ficaram sob a sombra durante a tarde e nas primeiras horas da noite. Essas duas lagoas e ensaios são de fato os exemplos extremos da variabilidade na preferência dos peixes pela sombra (veja Figura D.2 para todas as lagoas e ensaios). Para obtermos uma ideia mais geral, utilizamos a regressão logística para

quantificar a preferência pela estrutura de sombra. A regressão logística foi utilizada para testar se a variável binária indicando dia e noite é capaz de prever se o peixe está dentro ou fora da sombra. A partir disso descobrimos que a associação entre essas covariáveis varia substancialmente entre as lagoas e os ensaios, mas a luz do dia aumenta significativamente a probabilidade de encontrar peixes sob a estrutura de sombra em todos os casos (Figura 4.1E).

Enquanto existem evidências suportando a hipótese que a estrutura de sombra encoraja a residência da carpa prateada, não sabemos se o comportamento desses peixes difere substancialmente quando eles selecionam a sombra comparado a quando eles estão fora da sombra. Para começar a responder essa questão, calculamos a velocidade do grupo de peixes em três categorias: peixe sob a sombra durante o dia, peixe fora da sombra durante o dia e a noite. A Figura 4.2A retrata a distribuição cumulativa complementar da velocidade (isto é, a probabilidade $p(v)$ de achar um peixe com velocidade maior que v) de cada categoria. Esses resultados claramente mostram que os peixes se movem muito mais lentamente sob a sombra durante o dia. A velocidade média sob a sombra é $\langle v \rangle \approx 0,13$ m/s e os peixes raramente alcançam mais do que um metro por segundo nessa situação. Por outro lado, fora da sombra os peixes estão em média quase duas vezes mais rápido que dentro da sombra. Os peixes são um pouco mais lentos em média à noite ($\langle v \rangle \approx 0,21$ m/s) do que durante o dia, mas é durante a noite que os peixes alcançaram os maiores picos de velocidade. Essas diferenças na velocidade sugerem que os peixes estão ociosos sob a sombra, usando esse ambiente como local de abrigo e descanso. Em contrapartida, as maiores velocidades fora da sombra são provavelmente associadas com o comportamento de forrageamento². Além disso, as altas velocidades durante a noite também sugerem que esses peixes usam a escuridão para explorar a lagoa em segurança, já que essa condição visual pode decrescer o risco de predação nos ambientes reais.

Na direção das distribuições de probabilidade de velocidade, a dinâmica das séries temporais de velocidade dos peixes pode também indicar diferentes comportamentos sob ou fora da sombra durante o dia e a noite. Uma possibilidade interessante é quantificar o grau de persistência na velocidade dos peixes. Se os peixes estão ociosos sob a sombra e movem-se mais objetivamente fora dela, esperamos que a série temporal da velocidade seja menos persistente quando o peixe está sob a sombra durante o dia. Para verificar essa hipótese, estimamos o expoente de Hurst h da série temporal das velocidades agrupadas nas mesmas categorias já citadas. Como dissemos antes, o expoente de Hurst $h > 0,5$ indica comportamento de persistência de longo alcance, ou seja, incrementos na série temporal são mais prováveis de serem seguidos por incrementos de mesmo sinal. Esse efeito é tanto mais forte quanto mais próximo h está de 1, o maior valor possível do grau de persistência. Os valores de $h < 0,5$ representam um comportamento de antipersistência em que os incrementos na série temporal alternam de sinal com mais frequência. Uma série temporal é não correla-

²Forrageamento é a busca e a exploração de recursos alimentares.

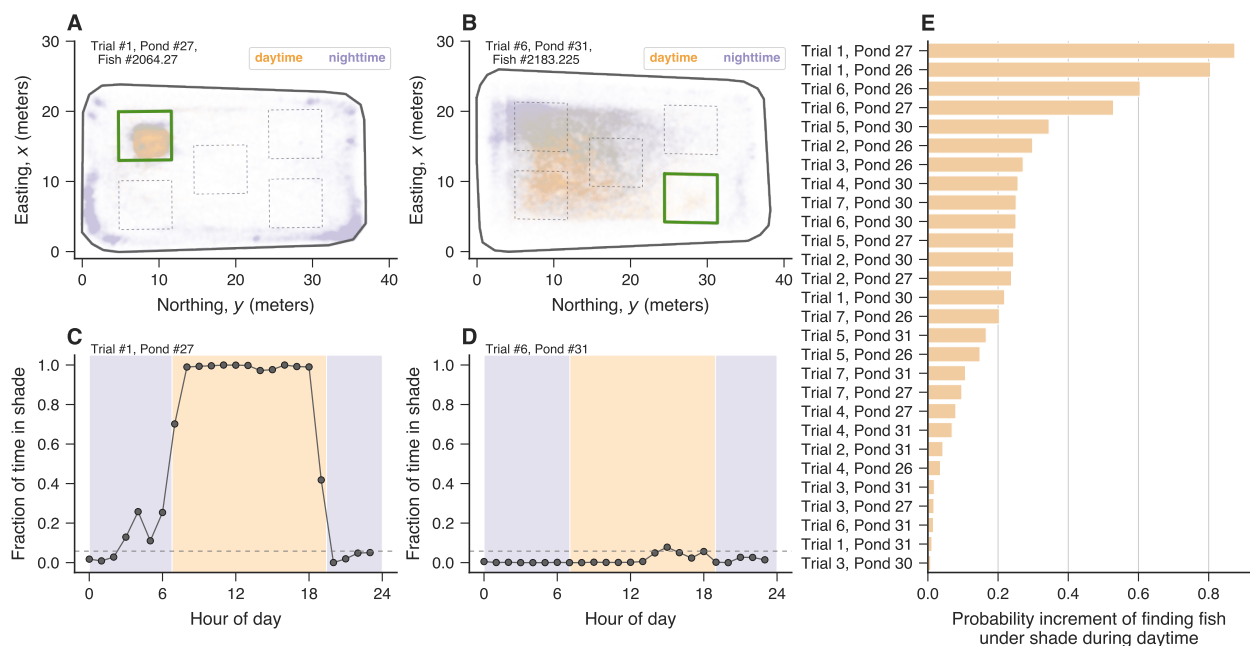


Figura 4.1: Preferência da carpa prateada pela estrutura de sombra. (A)-(B)

Mapas de densidade das posições de dois peixes de diferentes ensaios e lagoas. Linhas cinzas sólidas representam os limites das lagoas, os quadrados em linhas cinzas tracejadas indicam todas as posições possíveis para a plataforma flutuante (estrutura de sombra), e o quadrado verde indica o local dessa plataforma nos ensaios e lagoas cujos números são apresentados em cima e à esquerda nos painéis (A) e (B). Regiões das lagoas coloridas em laranja indicam as posições registradas durante as horas do dia, enquanto regiões coloridas em roxo representam as posições coletadas à noite. Observamos que o peixe em (A) tem uma preferência extrema pela área de sombra durante as horas do dia, preferência essa que não é clara para o peixe em (B). Nos painéis (C)-(D), os círculos mostram a fração do tempo que o peixe passou sob a estrutura de sombra em janelas de tempo de uma hora para os ensaios e lagoas em (A) e (B). A linha tracejada representa a fração esperada da posição sob a sombra caso os peixes se movessem aleatoriamente na lagoa, e a cor de fundo indica as horas de sol (laranja) e as horas de noite (roxo). Os peixes da lagoa #27 permaneceram na sombra durante quase todas as horas de sol nos dois dias do ensaio #1 e evitaram a sombra durante a noite. Os peixes da lagoa #31 selecionaram a estrutura de sombra muito mais raramente durante o dia (principalmente no período da manhã) no ensaio #6. No painel (E), mostramos o incremento de probabilidade de encontrar uma posição sob uma estrutura de sombra, conforme estimado por meio de regressão logística para cada ensaio e lagoa. A relação entre um peixe selecionar a sombra ou não e a variável independente binária dia ou noite varia substancialmente, mas não deixa de ser significativa para todos os ensaios e lagoas (p -valores $< 0,001$).

cionada se $h = 0,5$. A Figura 4.2B mostra que a velocidade dos peixes são marcadas pela persistência dentro e fora da sombra durante o dia e durante a noite (aproximadamente 90% das séries temporais têm $h > 0,5$). Existe uma diferença significativa na distribuição de h entre as três situações. Observamos que a distribuição do expoente de Hurst é deslocada para a esquerda no regime sob a sombra quando comparado com as outras duas situações. O

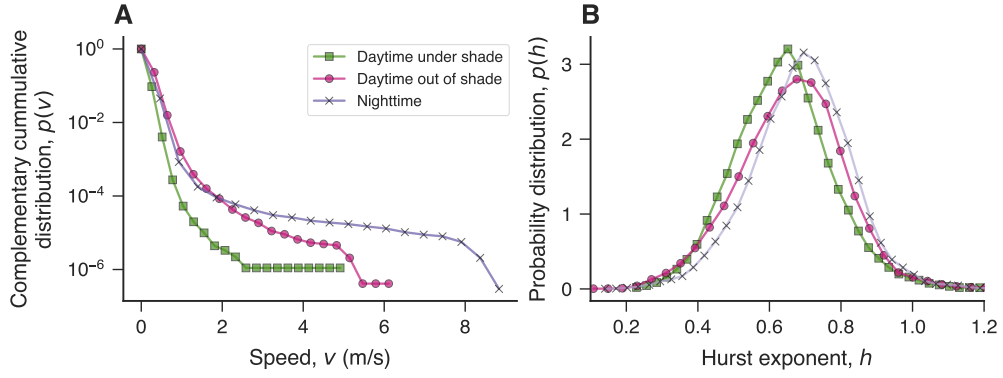


Figura 4.2: Diferença de velocidade dos peixes dentro ou fora da sombra durante o dia e à noite. (A) Distribuições cumulativas complementares da velocidade dos peixes v agrupada em três categorias: durante o dia sob a sombra (quadrados verdes), durante o dia fora da sombra (círculos rosa) e durante a noite (cruzes roxas). Os peixes movem-se significativamente mais devagar sob a sombra ($\langle v \rangle = 0,1286 \pm 0,0001$ m/s) e a velocidade média é maior quando o peixe está fora da sombra ($\langle v \rangle = 0,2324 \pm 0,0001$ m/s) do que à noite ($\langle v \rangle = 0,2086 \pm 0,0001$ m/s); no entanto, a distribuição de velocidade é mais assimétrica à noite, indicando que os peixes alcançam suas maiores velocidades durante a noite. (B) A distribuição de probabilidade do expoente de Hurst h estimado de séries temporais da velocidade agrupadas nas mesmas três categorias. As séries de velocidade mostram um comportamento de persistência de longo alcance nas três situações, mas a média do expoente de Hurst é um pouco menor sob a sombra durante as horas do dia ($\langle h \rangle = 0,639 \pm 0,002$) quando comparado com fora da sombra ($\langle h \rangle = 0,661 \pm 0,002$) ou à noite ($\langle h \rangle = 0,689 \pm 0,002$).

valor médio de h sob a sombra ($\langle h \rangle = 0,639 \pm 0,002$) é menor que fora ($\langle h \rangle = 0,661 \pm 0,002$) e à noite ($\langle h \rangle = 0,689 \pm 0,002$). Esses resultados apoiam a hipótese de que os peixes se movem com menos persistência sob a sombra e também indicam que seus movimentos são mais persistentes à noite.

Outra questão intrigante é se o ambiente de sombra afeta o comportamento coletivo das carpas prateadas. Para investigar essa possibilidade, calculamos dois parâmetros de ordem que quantificam o grau de alinhamento e o grau de rotação de cardumes de peixes [208, 209]. Como definido nos Métodos (seção 4.2), essas grandezas são chamadas de parâmetro de ordem de polarização O_p e parâmetro de ordem de rotação O_r . Os valores de O_p variam de 0 à 1 e medem o quão alinhado o movimento dos peixes é: $O_p \approx 0$ indica um estado de enxame em que os peixes se movem totalmente desorientados em diferentes direções, enquanto $O_p \approx 1$ representa o estado polarizado em que os peixes se movem alinhados uns com os outros. O valor de O_r também está limitado de 0 a 1 e esse parâmetro de ordem descreve o grau de rotação do cardume de peixe em torno do seu centro de massa: $O_r \approx 0$ representa um estado sem rotação enquanto $O_r \approx 1$ indica um estado de forte rotação do grupo. Calculamos os valores de O_p e O_r para cada instante de tempo para todas as lagoas e ensaios, com cinco ou mais peixes nas mesmas três categorias. As Figuras 4.3A, 4.3B

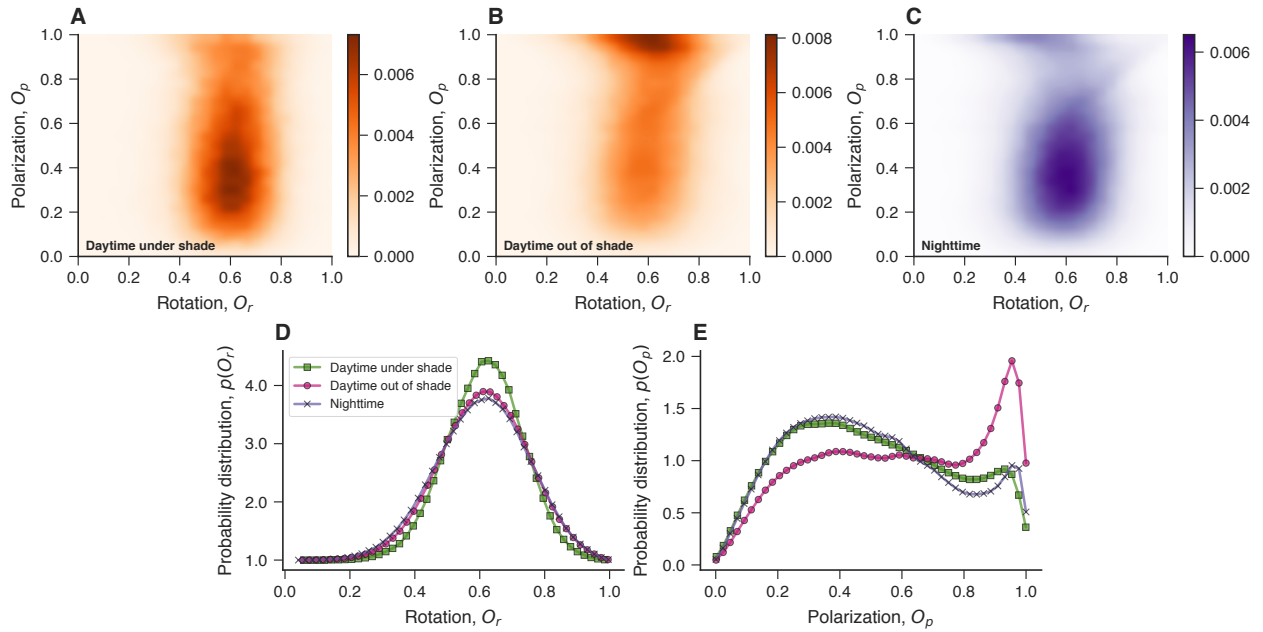


Figura 4.3: O efeito da sombra no comportamento de cardume da carpa prateada. A distribuição de probabilidade conjunta dos parâmetros de ordem de rotação (O_r) e polarização (O_p) quando agrupados nas três categorias: (A) sob a sombra durante o dia, (B) fora da sombra, e (C) durante a noite. A distribuição marginal do parâmetro de ordem de rotação (D) e polarização (E) para as mesmas categorias. Além da distribuição ligeiramente mais localizada de O_r durante o dia sob sombra, não há diferença significativa no grau de rotação coletiva entre as três categorias. Ao contrário disso, o grau de alinhamento medido por O_p é significativamente maior quando os peixes estão fora da sombra durante o dia comparado com as outras duas categorias.

e 4.3C exibem a distribuição conjunta dos dois parâmetros de ordem para cada categoria. Adicionalmente, as Figuras 4.3D e 4.3E mostram as distribuições marginais correspondentes para os parâmetros de ordem de rotação e polarização, respectivamente.

Os gráficos de O_p versus O_r mostram que o grau de alinhamento desses peixes difere significativamente entre as três categorias. Observamos que a distribuição conjunta é muito mais localizada em torno de altos valores de O_p quando os peixes estão fora da sombra (Figura 4.3B) comparado com as outras duas categorias. Esses comportamentos distintos tornam-se mais claros quando percebemos que a distribuição marginal de O_p tem um pico perto de 1 para posições fora da sombra (Figura 4.3E). O valor médio desses parâmetros de ordem é significativamente maior quando os peixes estão fora da sombra ($\langle O_p \rangle \approx 0,59$) do que sob a sombra ($\langle O_p \rangle \approx 0,50$) e à noite ($\langle O_p \rangle \approx 0,50$); portanto, a probabilidade de achar indivíduos fortemente alinhados ($O_p > 0,9$) é aproximadamente duas vezes maior fora da sombra do que nas outras duas categorias. Esses resultados sugerem que os peixes adotam estados de alta polarização como uma estratégia para fugir de possíveis predadores quando estão fora da estrutura de sombra. Surpreendentemente, o grau de polarização sob a sombra

(Figura 4.3A) é similar ao observado durante a noite (Figura 4.3C). Assim, embora os peixes movam-se mais rápida e persistentemente durante a noite do que sob a sombra, o grau de alinhamento entre os membros do grupo é baixo e não difere muito entre essas duas situações. Esses estados de baixa polarização durante a noite e sob a sombra provavelmente representam comportamentos responsivos à baixa possibilidade de predação durante a noite em habitats reais e também indicam que os peixes consideram a estrutura de sombra igualmente segura durante o dia.

Por outro lado, a estrutura de sombra não impacta muito no grau de rotação entre as três categorias (Figura 4.3). A distribuição marginal de O_r (Figura 4.3D) mostra que o grau de rotação é apenas um pouco menos espalhado sob a sombra, mas o valor médio desse parâmetro de ordem é aproximadamente o mesmo ($\langle O_r \rangle \approx 0,61$) nas três situações. Os peixes raramente adotam um estado de forte rotação em grupo, visto que apenas $\approx 1\%$ dos tempos foram caracterizados por $O_r > 0,9$, sem considerar as categorias. A escassez desses estados de forte rotação é provavelmente reflexo do número de indivíduos utilizados no experimento. Em uma pesquisa com a carpa dourada (*Notemigonus crysoleucas*), Tunström *et al.* [209] reportam que grupos de 30 peixes em um tanque raso de 2,1 m por 1,2 m raramente adotam estados de rotação em grupo, mas conforme o grupo cresce a frequência desses estados aumenta. Assim, pesquisas futuras com maior número de carpas prateadas podem nos ajudar a entender se esses peixes exibem fortes estados de rotação e se ambientes sombreados desempenham um papel na emergência desses estados.

Para investigar efeitos de sombra no comportamento coletivo das carpas prateadas, criamos uma representação em redes complexas das interações baseadas nas suas relações espaciais. Como detalhado nos Métodos (seção 4.2), primeiro obtemos a tesselação de Voronoi relacionada às posições dos peixes em um dado tempo e construímos as conexões entre os indivíduos que compartilham a mesma fronteira no diagrama de Voronoi (Figura 4.4A). Esse tipo de relação espacial entre os indivíduos se aproxima das propriedades das redes sensoriais e, portanto, é uma boa aproximação para as interações entre os peixes [214]. Em seguida, somamos as informações sobre todas as conexões dentro de uma janela de 10 minutos para criar uma rede pesada em que os peixes são os nós, e o peso das conexões entre eles representam a fração do tempo em que o par de peixes permaneceu vizinho durante o intervalo de tempo dividido pelo número de peixes menos um. Finalmente, agrupamos essas redes nas mesmas três categorias: sob a sombra durante o dia, fora da sombra e durante a noite. A Figura 4.4B retrata uma visualização da rede típica para cada uma dessas categorias. Nessa visualização, as conexões entre indivíduos que permaneceram vizinhos no diagrama de Voronoi para uma considerável fração de tempo são representadas por arestas espessas e escuras, enquanto arestas finas e claras indicam pares de peixes que foram vizinhos por muito menos tempo.

Uma simples inspeção visual das redes da Figura 4.4B indica que a interação entre os

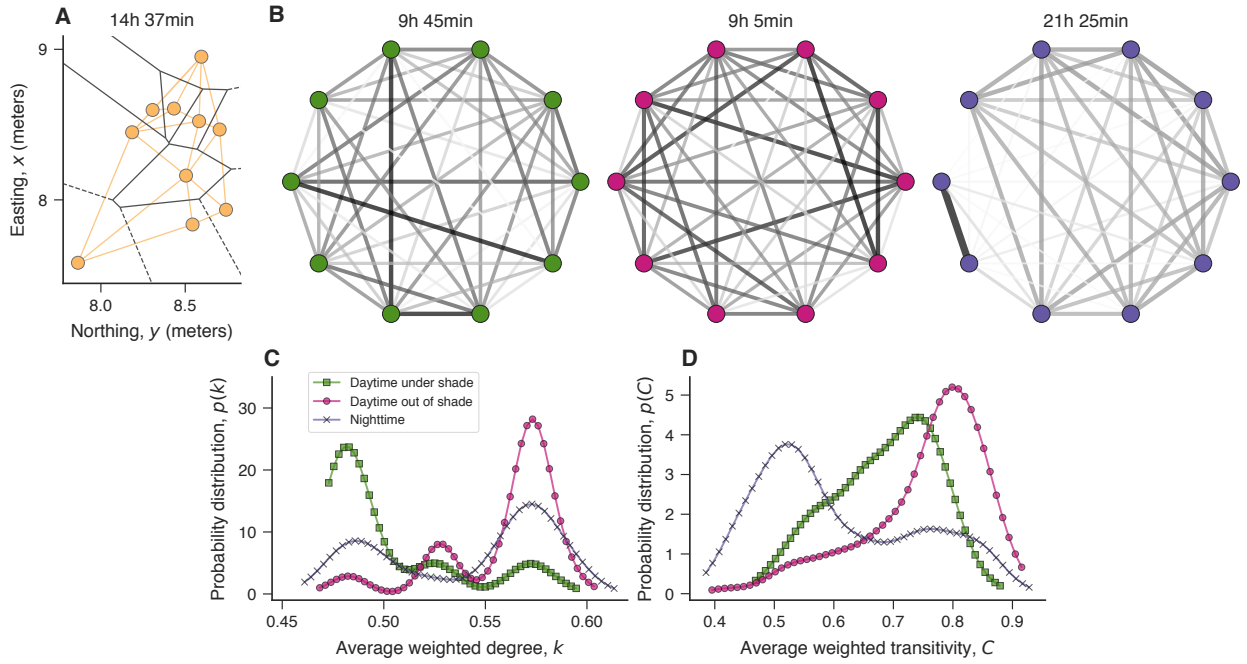


Figura 4.4: A rede de interações da carpa prateada durante o dia dentro e fora da sombra e à noite. (A) Um exemplo da tesselação de Voronoi de um grupo de dez peixes. As linhas pretas delineiam as células de Voronoi, enquanto os círculos laranjas indicam a posição dos peixes para esse tempo em particular (14h 37min). A rede de interação é criada conectando indivíduos que compartilham uma borda no diagrama de Voronoi (as linhas laranjas mostram todas as conexões nesse exemplo). (B) Os exemplos de redes pesadas obtidas pela soma das redes de interações dentro do intervalo de dez minutos. As ligações pesadas representam a fração de tempo que um par de peixes permaneceu vizinho no diagrama de Voronoi durante o intervalo de tempo dividido pelo número de peixes menos um. As linhas mais escuras e grossas entre os nós mostram que as conexões são mais fortes entre esses peixes. Os painéis (C) e (D) mostram a distribuição de probabilidade do grau pesado médio k e a transitividade pesada média C quando os dados são agrupados nas três categorias: durante o dia na sombra (quadrados verdes), durante o dia fora da sombra (círculos rosas) e à noite (cruzes roxas).

peixes durante o dia sob e fora da sombra e à noite são diferentes. A rede representando as posições fora da sombra tem conexões muito mais intensas entre todos os indivíduos, enquanto que as outras redes mostram um número menor destas conexões (particularmente durante a noite). Para explorar sistematicamente essas diferenças, primeiro calculamos o grau pesado médio k dessas redes. O peso κ de um nó é a soma dos pesos das arestas conectadas a ele e, portanto, representa a soma das frações de tempo de todos os peixes que foram vizinhos de um determinado peixe dividido pelo número total de peixes menos um. Essa quantidade é um indicador do número médio de vizinhos (pesado pela fração de tempo e normalizado pelo número de peixes) que um peixe tem durante um intervalo de dez minutos; por exemplo, se peixes permanecem em uma configuração tipo estrela, o peixe

central teria um valor de κ aproximadamente igual a um, enquanto κ é aproximadamente igual a um sobre o número de peixes menos um para todos os peixes periféricos. O grau pesado k é o valor médio dos pesos κ dos nós levando em consideração toda a rede e, assim, k é uma aproximação para o número de interações entre os peixes do cardume.

A Figura 4.4C mostra a distribuição de probabilidade do grau pesado médio k quando agrupamos a rede em três categorias. Observamos uma distinção clara entre a rede relacionada aos peixes dentro da sombra e fora. As posições sob a sombra rendem redes com grau pesado médio significativamente mais baixo comparado com as posições fora da sombra. Esse resultado mostra que a interação dos peixes diminui sob a sombra, e dado que sabemos a polarização é baixa na sombra (Figura 4.3E), podemos afirmar que esse baixo número de interações é parcialmente explicado pela instabilidade de formações que caracterizam o estado tipo enxame. Por outro lado, o alto número de interações mostradas pelos peixes fora da sombra reflete formações mais estáveis e polarizadas nas quais os peixes monitoram as direções de seus vizinhos. Dessa forma, a distribuição do grau pesado médio associado às posições noturnas é mais uniforme que nos outros dois casos anteriores, indicando que o número de interações à noite é mais aleatório.

Também calculamos o coeficiente de transitividade pesado médio C [212] dessas três redes. Para redes não ponderadas, o coeficiente de transitividade (também conhecido como coeficiente de aglomeração) mede a densidade de tríades fechadas, indicando a tendência dos vizinhos mais próximos a um nó estarem conectados entre si [210]. Similarmente, a transitividade pesada quantifica a densidade de tríades levando em consideração a média geométrica dos pesos dessas formações triangulares. Assim, tríades fechadas nas quais uma dessas ligações tem baixa intensidade contribui pouco para a transitividade pesada, enquanto tríades fechadas tendo ligações com maior peso contribuem mais. Podemos interpretar esse coeficiente de transitividade como um indicador de ordem local na posição dos peixes de modo que vizinhanças locais estáveis (ou seja, movimentações que tendem a manter os mesmos primeiros vizinhos) produzam valores mais altos de C do que vizinhanças locais instáveis.

A Figura 4.4D mostra a distribuição de probabilidade do coeficiente de transitividade pesado médio C quando agrupamos as redes nas três categorias. Notamos que a distribuição de C é notavelmente diferente entre as rede noturna e as redes diurna tanto sob quanto fora da sombra. As posições noturnas geram redes com valores de transitividade menores e a distribuição tem um pico em torno de $C \approx 0,5$, enquanto as posições diurnas criam redes com valores de transitividade significativamente maiores. Quando comparamos as posições dentro e fora da sombra, observamos que as posições fora da sombra produzem os maiores valores de transitividade (pico da distribuição em torno de $C \approx 0,85$), do que posições sob a sombra (pico da distribuição em $C \approx 0,75$). As posições diurnas fora da sombra são assim marcadas por fortes relações de tríades fechadas, indicando não apenas que o movimento

global do cardume é polarizado, mas também que a vizinhança local dos peixes permanece muito estável durante movimentações nesta condição. Por outro lado, e apesar do grau de polarização ser similar durante o dia sob a sombra e à noite (Figura 4.3E), nossos resultados mostram que a formação de tríades fechadas e fortes é bem menos comum durante a noite. Assim, embora os peixes frequentemente permaneçam em estado de enxame de dia sob a sombra e à noite, as posições à sombra preservam um grau de ordem local intermediário em que apesar das diferentes direções de movimento os peixes ainda monitoram as posições de seus vizinhos.

4.4 Discussão

Apresentamos uma extensa investigação dos efeitos de sombra no comportamento das carpas prateadas. Similarmente a outras espécies [196, 198–201], verificamos que a sombra encoraja a residência das carpas prateadas durante o dia, mas o grau de preferência pela sombra varia significativamente entre nossos experimentos. Além de descobrir que os peixes são mais propensos a estar sob estruturas de sombra durante o dia do que à noite, nossa pesquisa revela que o comportamento da carpa prateada é significativamente diferente dependendo se os peixes estão dentro ou fora da sombra durante o dia ou à noite. Durante o dia, os peixes nadam muito mais lentamente sob a sombra do que fora da sombra, mas é à noite que eles alcançam suas maiores velocidades. Além disso, as séries temporais das velocidades apresentam diferentes graus de persistência de longo alcance entre essas três condições. A persistência de longo alcance decresce quando os peixes estão dentro da sombra e torna-se mais intensa à noite.

Complementarmente, investigamos como o comportamento coletivo dessa espécie de peixe se dá dentro e fora da sombra durante o dia e durante a noite utilizando duas abordagens. Na primeira, contamos com dois parâmetros de ordem que quantificam o grau de polarização e rotação do cardume. Dessa forma, em nossos experimentos, encontramos que a carpa prateada raramente adota estados de forte rotação, ou seja, a sombra e a noite não afetam seu comportamento. Por outro lado, no período diurno, os peixes assumem o estado de enxame sem nenhuma direção preferencial de movimento quando estão sob a estrutura de sombra e adotam um estado de forte polarização, movendo-se alinhados uns com os outros, quando fora da sombra. Na segunda abordagem, criamos redes de interação para as carpas prateadas baseadas em suas relações espaciais via tesselação de Voronoi, uma técnica que se aproxima das redes sensoriais dos peixes [214]. Nessa direção, encontramos que essas redes de interação são notavelmente diferentes dentro e fora da sombra durante o dia e à noite. Os peixes interagem menos entre si à sombra e tendem a manter um alto número de vizinhos estáveis quando estão fora da sombra durante o dia; à noite, o número ponderado de interações é distribuído mais uniformemente. Observamos ainda uma hierarquia quanto

ao surgimento de tríades fortes e fechadas entre os peixes. A densidade das tríades fortes e fechadas é muito maior quando os peixes estão fora da sombra durante o dia, diminui quando estão debaixo da sombra, e diminui ainda mais à noite.

Em conjunto, nossos resultados indicam que carpas prateadas diminuem suas velocidades, a persistência de sua movimentação e sua polarização de grupo sob estruturas de sombra, o que é consistente com estados de descanso ou de movimento latente. Apesar disso, esses peixes parecem manter algum nível de ordem local como indicado pelos valores intermediários de transitividade em suas redes de interação nesta condição. Durante o dia e fora da sombra, carpas prateadas movem-se rapidamente, persistentemente, e em grupos mostrando polarização intensa e formações estáveis marcadas por valores altos de transitividade e um grande número de vizinhos estáveis. Esses resultados sugerem que as carpas prateadas mantêm um estado de atenção coletiva quando estão se movendo na lagoa ao longo do dia, presumivelmente para forragear. Finalmente, esses peixes alcançam suas maiores velocidades e persistência de longo alcance à noite, enquanto a polarização do grupo é semelhante à situação sob a sombra. Ainda mais, as interações entre os peixes são muito mais fracas à noite comparadas às outras duas condições, pois as redes de interações são caracterizadas por uma distribuição de graus pesados mais uniformes e baixos valores de transitividade. Assim, os peixes parecem aproveitar o escuro para explorar com segurança o ambiente em estados de grupo marcados por pouca organização, provavelmente porque essa condição visual é associada a uma baixa pressão de predação em habitats naturais.

Considerações finais e perspectivas

Em nossa aplicação a um sistema social, apresentamos resultados sobre jogos digitais e seus jogadores. Tomamos vantagem do grande banco de dados da *Steam*, com milhares de jogos e dados registrados automaticamente para conduzir nossos estudos. Analisando a proporção de jogadores que alcança cada *achievement*, obtemos cinco grandes resultados sobre os jogos *online*: a distribuição de probabilidade do número de *achievements* é log-normal; a distribuição de probabilidade da popularidade dos jogos também segue uma log-normal; as desenvolvedoras preferem fazer jogos difíceis, porém não impossíveis; os jogadores também preferem jogos com uma certa dificuldade intermediária; o número de jogadores como função dos *achievements* alcançados decai, em média, exponencialmente. Em outra direção, analisando o número médio de jogadores por mês, detectamos e demonstramos a robustez de cinco padrões que regem a dinâmica de popularidade, eles foram nomeados: decrescente (representa 47,4% dos jogos), montanhoso (20,7%), crescente (11,8%), explosivo (11,0%) e vale (9,1%). Além disso, definimos quais padrões de popularidade são mais prováveis em cada gênero de jogo. Em resumo, nossos achados poderiam definir diretrizes para maximizar os efeitos de ensino e aprendizagem, regulando a dificuldade de acordo com o desempenho e a persistência das pessoas. Já em uma vertente econômica, a detecção de padrões universais na dinâmica de popularidade de conteúdos culturais, em especial de jogos *online*, pode ser útil para melhorar estratégias de monetização e maximização de lucro por parte dos desenvolvedores. Finalmente, destacamos que o banco de dados utilizado no trabalho atual é uma amostra representativa de jogadores de todo o mundo, mas trabalhos futuros podem testar a validade de nossos achados usando outros bancos de dados.

Em nossa aplicação a um sistema biológico, identificamos padrões de comportamento de carpas prateadas (*Hypophthalmichthys molitrix*) sob a sombra e fora dela. Para esse estudo, utilizamos os dados cedidos pela Columbia Environmental Research Center, em Columbia no Missouri, Estados Unidos. Os experimentos foram conduzidos em lagoas artificiais, em

que havia uma plataforma móvel capaz de projetar sombra. Foram feitos sete ensaios em cada uma das quatro lagoas. Em geral, nossas análises descreveram a preferência dos peixes, além de seus padrões de movimentação individuais e coletivos. Um sumário desse estudo envolve sete principais resultados: verificamos que a sombra encoraja a residência das carpas prateadas durante o dia; os peixes nadam muito mais lentamente sob a sombra do que fora durante o dia, no entanto é à noite que eles alcançam suas maiores velocidades; a persistência de longo alcance decresce quando os peixes estão dentro da sombra e torna-se mais intensa à noite; encontramos que a carpa prateada raramente adota estados de forte rotação e que a sombra e à noite não afeta o comportamento delas; fora da estrutura de sombra as carpas adotam estados de maior polarização, movendo-se alinhadas umas com as outras; os peixes tendem a manter um maior número de vizinhos estáveis quando está fora da sombra durante o dia; a densidade das tríades fortes e fechadas é muito maior quando os peixes estão fora da sombra durante o dia, reduz quando está dentro da sombra e diminui ainda mais à noite. Ao fim dessa investigação, acreditamos que nossos achados têm alguma importância para as pesquisas de ecologia e conservação, pois apresentamos padrões de comportamento de uma espécie invasora do rio do Mississípi que precisa ser controlada. Na direção de perspectivas futuras, sob o enfoque de comportamento de animais aquáticos, outros trabalhos podem ser propostos para investigar diferenças entre o padrão da carpa prateada e outras espécies.

APÊNDICE A

Demonstração do Teorema central do limite

Para demonstrar o Teorema central do limite, vamos partir de $x_1, x_2, x_3, \dots, x_n$ sendo variáveis aleatórias independentes e identicamente distribuídas com média a e variância b . A seguir, escrevemos uma variável z desse conjunto de variáveis x_j que é dada por

$$z = \frac{1}{\sqrt{nb}} \left(\sum_{j=1}^n x_j - na \right). \quad (\text{A.1})$$

A questão fundamental do Teorema central do limite é determinar a distribuição de probabilidade da variável z , no limite de $n \rightarrow \infty$.

Para determinar $\rho(z)$, podemos trabalhar com a transformada de Fourier das distribuições de probabilidade. Primeiro, vamos definir $\rho_1(x_1), \rho_2(x_2), \dots, \rho_n(x_n)$ como sendo as distribuições de probabilidade das variáveis $x_1, x_2, x_3, \dots, x_n$, respectivamente. Sabendo que todas as variáveis são identicamente distribuídas, vamos considerar por simplicidade $\rho_j(x) = \tilde{\rho}(x)$, para todo j . Consideraremos $g(k)$ a transformada de Fourier de $\tilde{\rho}(x)$. Para determinar $g(k)$ partiremos de um truque matemático simples:

$$g(k) = \exp \{ \ln g(k) \} = \exp \left\{ \ln \left(\int_{-\infty}^{\infty} e^{ikx} \tilde{\rho}(x) dx \right) \right\}. \quad (\text{A.2})$$

Tomando a expansão em série de Taylor para o expoente, obtemos

$$g(k) = \exp \left\{ \ln \left(\int_{-\infty}^{\infty} \sum_{j=0}^{\infty} \frac{(ikx)^j}{j!} \tilde{\rho}(x) dx \right) \right\} = \exp \left\{ \ln \left(1 + \sum_{j=1}^{\infty} \frac{(ik)^j}{j!} \langle x^j \rangle \right) \right\}. \quad (\text{A.3})$$

Note que conseguimos escrever a $g(k)$ em função dos momentos da distribuição $\tilde{\rho}(x)$. Sabendo

que a série de Taylor do logaritmo conduz a $\ln(1+x) \approx x - \frac{1}{2}x^2$, podemos escrever:

$$g(k) = \exp \left\{ \sum_{j=1}^{\infty} \frac{(ik)^j}{j!} \langle x^j \rangle + \left(\sum_{j=1}^{\infty} \frac{(ik)^j}{j!} \langle x^j \rangle \right)^2 + o(k^3) \right\} = \exp \left\{ ika - \frac{1}{2}k^2b + o(k^3) \right\}. \quad (\text{A.4})$$

Perceba que foi possível escrever a transformada de Fourier apenas em função da média $a = \langle x \rangle$, da variância $b = \langle x^2 \rangle - \langle x \rangle^2$ e potências de k de ordem superior a 2.

O passo seguinte é determinar a relação entre as transformadas de Fourier de $\tilde{\rho}(x)$ e de $\rho(z)$. Apesar de ainda não sabermos a função da distribuição de probabilidade de z , vamos denota-la por $\rho(z)$. E admitimos que sua transformada de Fourier é $G(k)$. Partindo disso, utilizaremos que $G(k)$ pode ser escrito como uma operação de média de $\exp\{ikz\}$:

$$\begin{aligned} G(k) &= \langle \exp\{ikz\} \rangle = \left\langle \exp \left\{ \frac{ik}{\sqrt{nb}} \left(\sum_{j=1}^n x_j - na \right) \right\} \right\rangle \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left\{ ik \frac{ik}{\sqrt{nb}} \left(\sum_{j=1}^n x_j - na \right) \right\} \rho(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n. \end{aligned} \quad (\text{A.5})$$

Sabendo que $x_1, x_2, \dots, x_n$ são variáveis aleatórias independentes e idênticas, então $\rho(x_1, x_2, \dots, x_n) = \tilde{\rho}_1(x_1)\tilde{\rho}_2(x_2)\dots\tilde{\rho}_n(x_n)$. Além disso, considerando $K = k/\sqrt{nb}$, verificamos que

$$\begin{aligned} G(K) &= \exp\{-iKan\} \prod_{j=1}^n \int_{-\infty}^{\infty} \exp\{iKx_j\} \rho_j(x_j) dx_j \\ &= \exp\{-iKan\} \prod_{j=1}^n g_i(K) = [g(K) \exp\{-iKa\}]^n. \end{aligned} \quad (\text{A.6})$$

Dessa maneira, determinamos $G(K)$ em função de $g(K)$. Portanto, basta apenas substituir o resultado encontrado em (A.4) na equação acima e tomar a transformada inversa. Assim, teremos

$$G(K) = \exp \left\{ -\frac{1}{2}K^2bn + o(K^3) \right\}. \quad (\text{A.7})$$

No limite de $n \rightarrow \infty$, temos $o(K^3) \rightarrow 0$ e, portanto,

$$G(K) = \exp \left\{ -\frac{1}{2}bnK^2 \right\}. \quad (\text{A.8})$$

Lembrando que $bnK^2 = k^2$, a $G(k)$ pode ser escrita como

$$G(k) = \exp \left\{ -\frac{1}{2}k^2 \right\} \quad (\text{A.9})$$

e sua transformada inversa é dada por

$$\rho(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\{-ikz\} \exp\left\{-\frac{1}{2}k^2\right\} dk. \quad (\text{A.10})$$

A integral da Equação A.10 é conhecida como Integral Gaussiana. Ela pode até parecer uma integral complicada, mas existe um truque muito simples para resolvê-la, primeiramente vamos reescrever o expoente de modo a aparecer um quadrado perfeito:

$$\rho(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{k}{\sqrt{2}} + \frac{iz\sqrt{2}}{2}\right)^2\right\} \exp\left\{-\frac{z^2}{2}\right\} dk. \quad (\text{A.11})$$

Utilizando a mudança de variável $l = \frac{k}{\sqrt{2}} + \frac{iz\sqrt{2}}{2}$, a integral pode ser reescrita como

$$\rho(z) = \frac{e^{-\frac{z^2}{2}}}{\pi\sqrt{2}} \int_{-\infty}^{\infty} e^{-l^2} dl. \quad (\text{A.12})$$

Agora, o único problema é resolver a integral

$$I = \int_{-\infty}^{\infty} e^{-l^2} dl. \quad (\text{A.13})$$

Para resolvê-la, elevaremos os dois lados ao quadrado para obter

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-l^2} dl\right)^2 = \left(\int_{-\infty}^{\infty} e^{-l_1^2} dl_1\right) \left(\int_{-\infty}^{\infty} e^{-l_2^2} dl_2\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(l_1^2+l_2^2)} dl_1 dl_2. \quad (\text{A.14})$$

Sabendo que as duas variáveis (l_1 e l_2) são independentes uma da outra, podemos pensar que a integral I^2 está sendo feita em um plano infinito e, dessa forma, podemos empregar uma mudança de variáveis para coordenadas polares (r, θ):

$$I^2 = \int_0^{\infty} \int_0^{2\pi} e^{-r^2} r d\theta dr = 2\pi \int_0^{\infty} e^{-r^2} r dr. \quad (\text{A.15})$$

Fazendo $R = r^2$, constatamos que

$$I^2 = \pi \int_0^{\infty} e^{-R} dR = \pi \quad (\text{A.16})$$

e, assim, $I = \sqrt{\pi}$. Voltando a Equação (A.12), chegamos a

$$\rho(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \quad (\text{A.17})$$

e finalmente obtemos a função da distribuição gaussiana padrão.

Apenas lembrando o que já foi dito, para a validade, desse teorema é necessário garantir a existência da média a , da variância b e ausência de correlações entre as variáveis x_j . Como deve ter ficado claro ao longo da demonstração, para o cálculo da $G(k)$ interessa a convergência dos momentos da distribuição e à medida que n cresce apenas os dois primeiros momentos tornam-se realmente relevantes.

A única parte subjetiva é a determinação do quão grande deve ser n para haver uma boa convergência do Teorema central do limite. Na Figura A.1, apresentamos a evolução da soma de algumas variáveis aleatórias. Observando a figura, percebemos que a convergência da distribuição uniforme (em azul) ocorre muito rapidamente, e a soma de $n = 3$ variáveis aleatórias já apresenta uma distribuição normal bem visível. No caso da distribuição exponencial (em verde), $n = 3$ não é suficiente para determinar a normal, porém $n = 10$ já apresenta boa convergência. Por outro lado, a distribuição log-normal (em laranja) tem uma convergência bem mais complicada, para $n = 3$ e $n = 10$ ela tem uma assimetria bem visível. Uma boa convergência para a variável log-normal começa parecer razoável para $n = 100$. Vale dizer que devemos tomar cuidado com a distribuição log-normal e com distribuições de cauda longa em geral. Para alguns valores dos parâmetros da log-normal, por exemplo, a distribuição da soma dessas variáveis aleatórias falha em convergir para uma normal padrão. Quando um caso assim ocorre ou a distribuição não garante a existência da média e da variância, generalizações do Teorema central do limite devem ser consideradas [215].

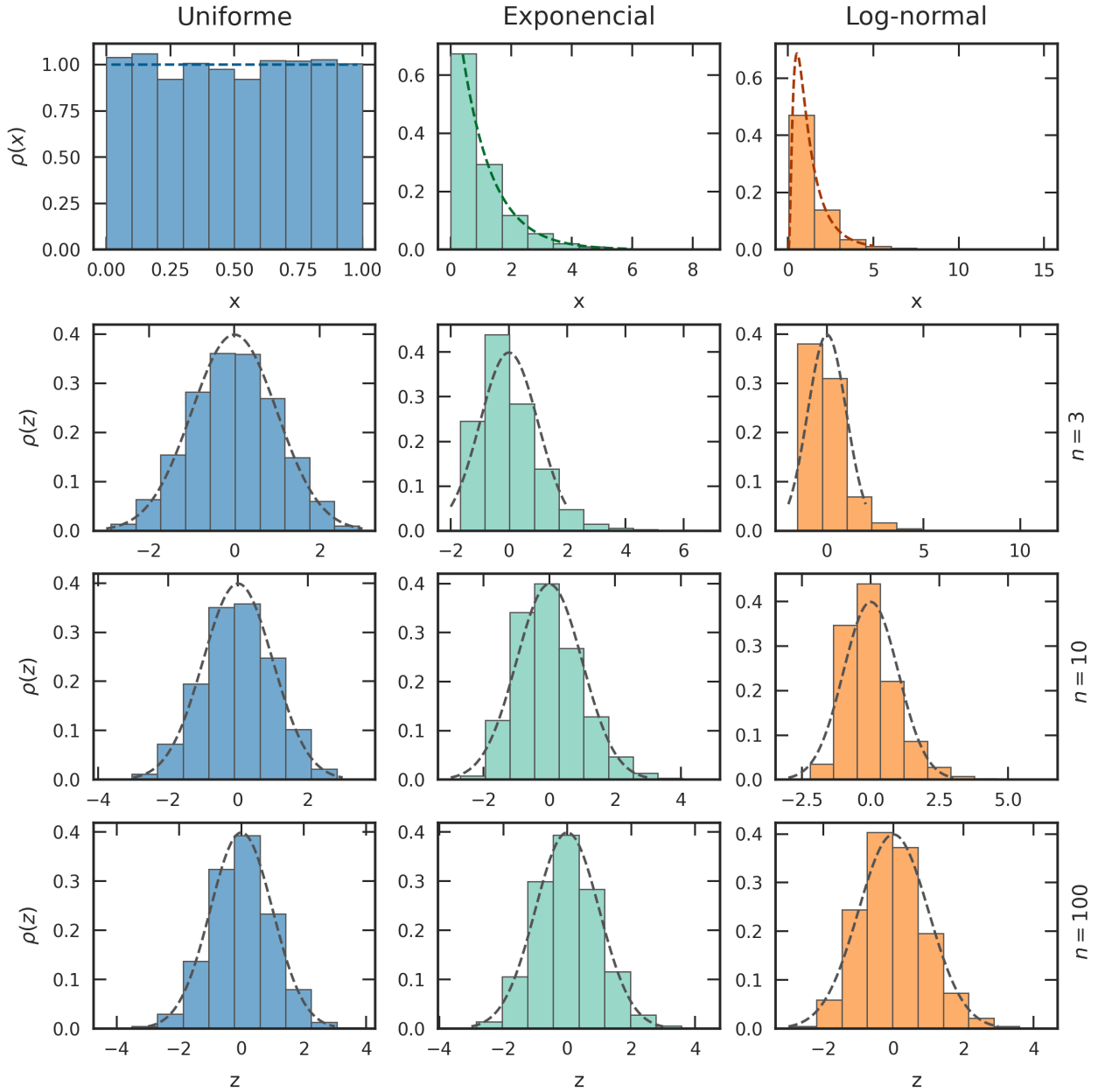


Figura A.1: Convergência da soma de variáveis aleatórias. Nos painéis da primeira coluna, apresentamos uma variável uniformemente distribuída ($a = 0$, $b = 1$) na primeira linha, e nas linhas seguintes a variável z , como definida na Equação (A.1), para a soma de $n = 3$, $n = 10$ e $n = 100$ variáveis aleatórias idênticas e independentes. Na segunda coluna, mostramos uma variável de distribuição exponencial ($\beta = 1$) e a soma de variáveis exponenciais. Na terceira coluna apresentamos o mesmo processo para uma distribuição log-normal ($\mu = 0$, $\sigma = 0,8$). A linha tracejada azul representa a distribuição normal ($a = 0$, $b = 1$), a tracejada verde a exponencial ($\beta = 1$), a laranja a log-normal ($\mu = 0$, $\sigma = 0,8$), e a preta a normal ($\mu = 0$, $\sigma = 1$).

APÊNDICE B

Dados gerados aleatoriamente

Neste apêndice, apresentamos a tabela de uma amostra hipotética (gerada aleatoriamente) utilizada nos exemplos das Seções 1.9 e 1.10.

Altura (m)									
1,74	1,90	1,89	1,76	1,68	1,71	1,65	1,65	1,82	1,76
1,66	1,93	1,88	1,57	1,74	1,62	1,61	1,66	1,65	1,74
1,72	1,76	1,69	1,69	1,69	1,93	1,58	1,78	1,64	1,97
1,71	1,65	1,68	1,73	1,75	1,81	1,78	1,64	1,78	1,71
1,77	1,69	1,72	1,88	1,97	1,66	1,74	1,73	1,73	1,84
1,74	1,77	1,64	1,79	1,97	1,76	1,56	1,76	2,08	1,84
1,80	1,85	1,88	1,55	1,56	1,66	1,61	1,81	1,68	1,78
1,83	1,63	1,67	1,80	1,79	1,87	1,71	1,76	1,71	1,77
1,62	1,57	1,69	1,90	1,64	1,87	1,67	1,69	1,50	1,60
1,72	1,63	1,75	1,71	1,60	1,73	1,77	1,73	1,57	1,77
1,60	1,74	1,68	1,68	1,82	1,76	1,81	1,65	1,81	1,81
1,66	1,68	1,69	1,72	1,60	1,85	1,81	1,55	2,00	1,79
1,88	1,90	1,61	1,70	1,70	1,75	1,65	1,75	1,64	1,74
1,86	1,75	1,62	1,62	1,80	1,85	1,70	1,65	1,69	1,72
1,79	1,70	1,58	1,60	1,67	1,87	1,85	1,80	1,90	1,59
1,46	1,69	1,69	1,80	1,63	1,66	1,73	1,69	1,75	1,88
1,96	1,74	1,85	1,87	1,72	1,87	1,76	1,79	1,57	1,84
1,79	1,82	1,76	1,64	1,71	1,78	1,65	1,60	1,75	1,71
1,64	1,68	1,61	1,74	1,99	1,81	1,91	1,60	1,97	1,89
1,64	1,87	1,77	1,70	1,69	1,68	1,77	1,70	1,69	1,63

Tabela B.1: **Amostra gerada aleatoriamente.** Altura em metros gerada para os exemplos das Seções 1.9 e 1.10.

APÊNDICE C

Figuras complementares do capítulo 3

Neste apêndice apresentamos figuras que complementam as ideias do capítulo 3. Julgamos que tais figuras são necessárias para entender detalhes da dinâmica de popularidade dos jogos abordados. Optamos por compilar as figuras neste apêndice com o objetivo de deixar o texto do capítulo 3 mais claro e objetivo.

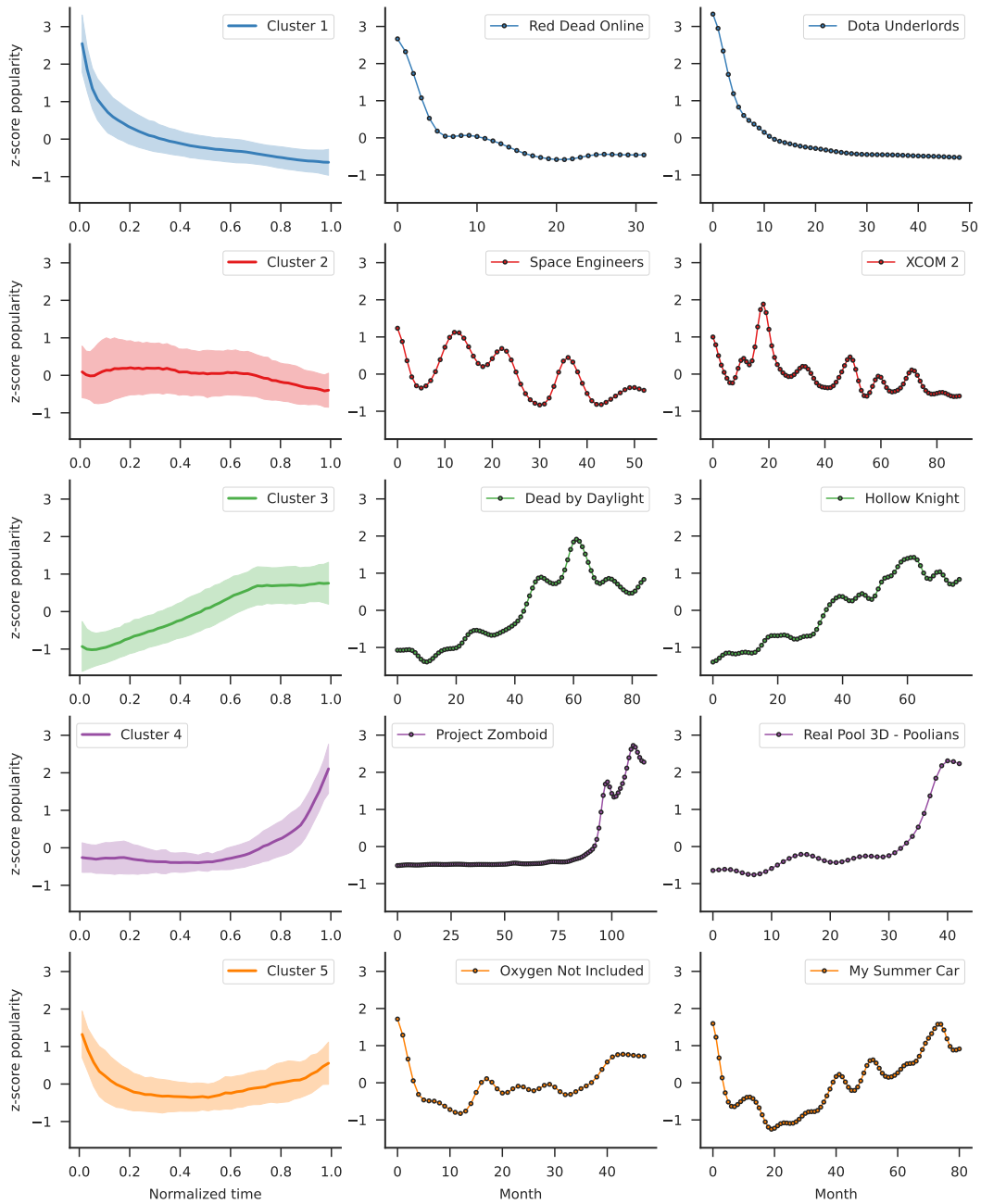


Figura C.1: Exemplos de séries de popularidade de cada *cluster* detectado. A primeira coluna mostra o padrão de popularidade de cada *cluster*: Cluster 1 (azul), Cluster 2 (vermelho), Cluster 3 (verde), Cluster 4 (roxo) e Cluster 5 (laranja). A linha sólida representa a média das séries de *z-score* suavizadas dos jogos de cada *cluster* a área sombreada representa um desvio padrão. A segunda e a terceira coluna fornece exemplos de séries de *z-score* suavizadas de alguns jogos dentro de cada *cluster*.

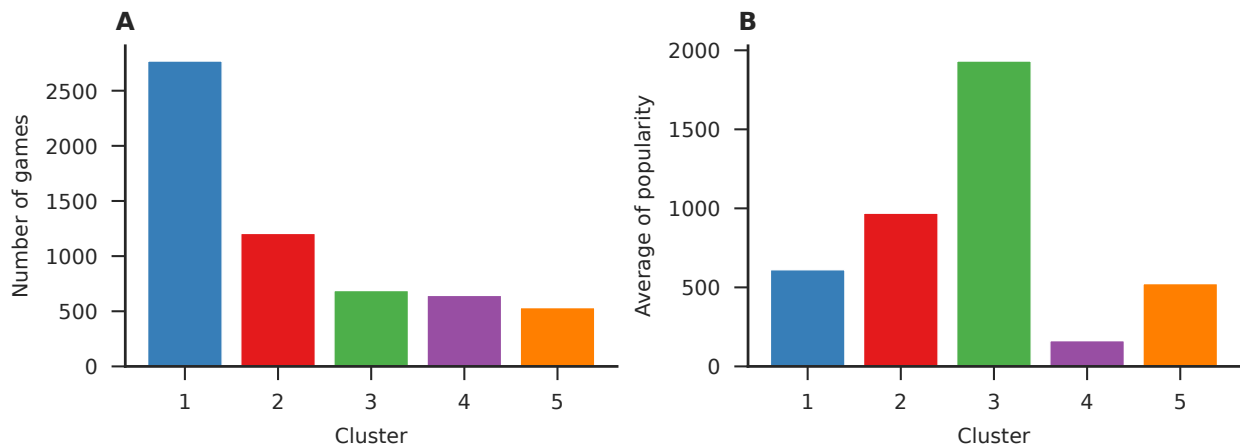


Figura C.2: Número de jogos e jogadores associados aos *clusters* detectados. (A) O número de jogos em cada *cluster*. (B) A média dos valores médios das séries de popularidade de cada *cluster*.

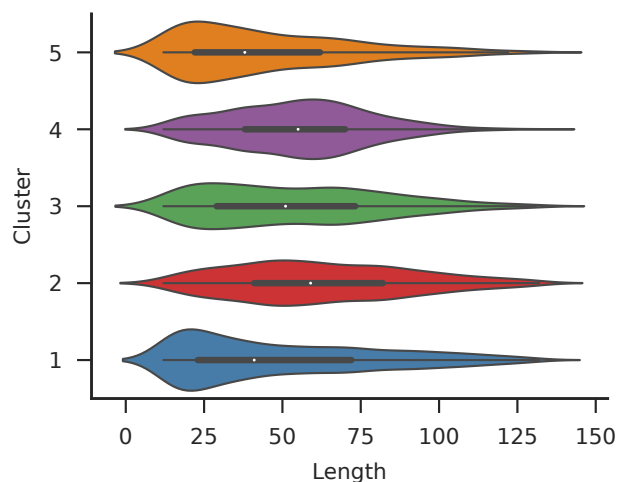


Figura C.3: A distribuição do tamanho das séries de cada *cluster*. As curvas de densidade foram construídas baseadas no dado do tamanho das séries temporais. O eixo x representa o tamanho das séries temporais, o eixo y identifica o *cluster* e a espessura da área colorida representa a densidade de jogos. Por exemplo, o Cluster 1 e o Cluster 5 têm muitos jogos com séries de popularidade com tamanho de 25 meses, enquanto o Cluster 4 tem muitos jogos com tamanho entre 50 e 75 meses. No gráfico de cada distribuição, o ponto branco denota a mediana, a barra mais espessa preta representa o intervalo interquartil e as linhas pretas mais finas se estendem entre os limites inferior e superior.

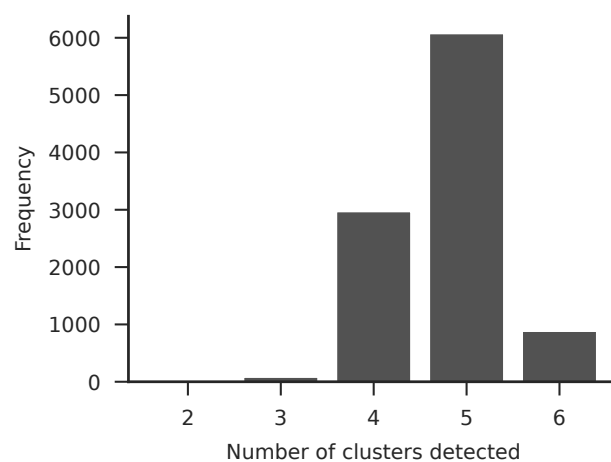


Figura C.4: Frequência do número de *clusters* detectados pelo algoritmo Infomap para os primeiros 24 meses de jogos com séries de popularidade superiores a 74 meses. O histograma mostra a frequência do número de *clusters* detectados em dez mil execuções. A detecção de 5 *clusters* foi a moda, ocorrendo 6.072 vezes.

APÊNDICE D

Figuras complementares do capítulo 4

Neste apêndice apresentamos figuras que complementam as ideias do capítulo 4. Pelos mesmos motivos apresentados no apêndice anterior, compilamos as figuras complementares do capítulo 4 neste apêndice.

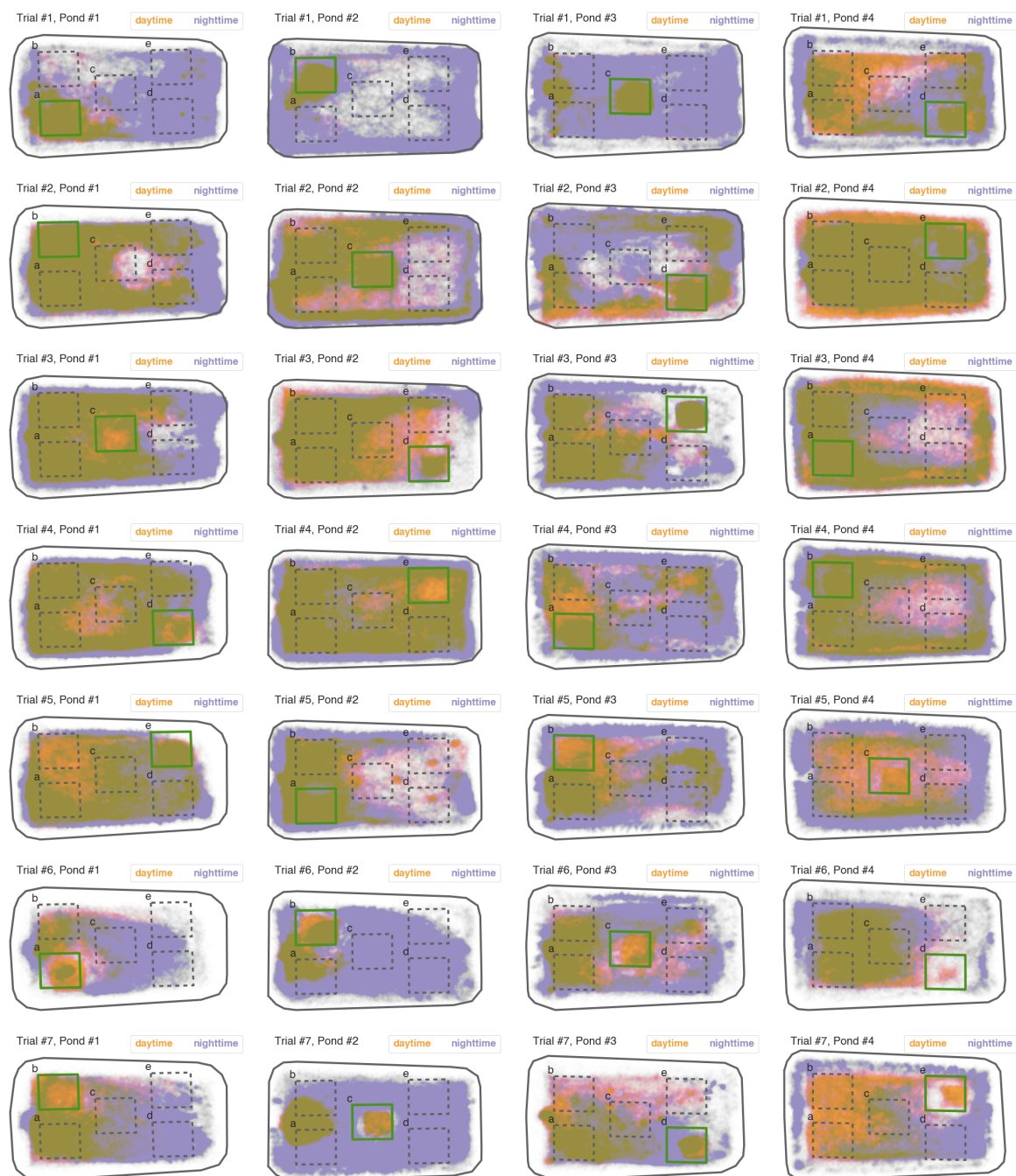


Figura D.1: Mapas de densidade das posições dos peixes de cada lagoa e ensaio usados em nosso estudo. Tons de laranja indicam posições registradas durante o dia, enquanto tons de roxo representam posições coletadas à noite. As linhas sólidas cinzas representam os limites das lagoas e os quadrados verdes espessos indicam a localização real da estrutura de sombra em cada ensaio e lagoa (como mostrado no canto superior esquerdo dos painéis). Os quadrados tracejados indicados por letras minúsculas (de “a” a “e”) representam todas as possíveis posições de sombra. As estruturas de sombra estão respectivamente localizadas nas posições “a”, “b”, “c” e “d” das lagoas 1, 2, 3 e 4 no ensaio 1. Essas estruturas de sombra são então movidas sequencialmente ao longo dos ensaios; por exemplo, as localizações de sombra ao longo dos ensaios na lagoa #2 são: “c” → “d” → “e” → “a” → “b” → “c” → “d”.

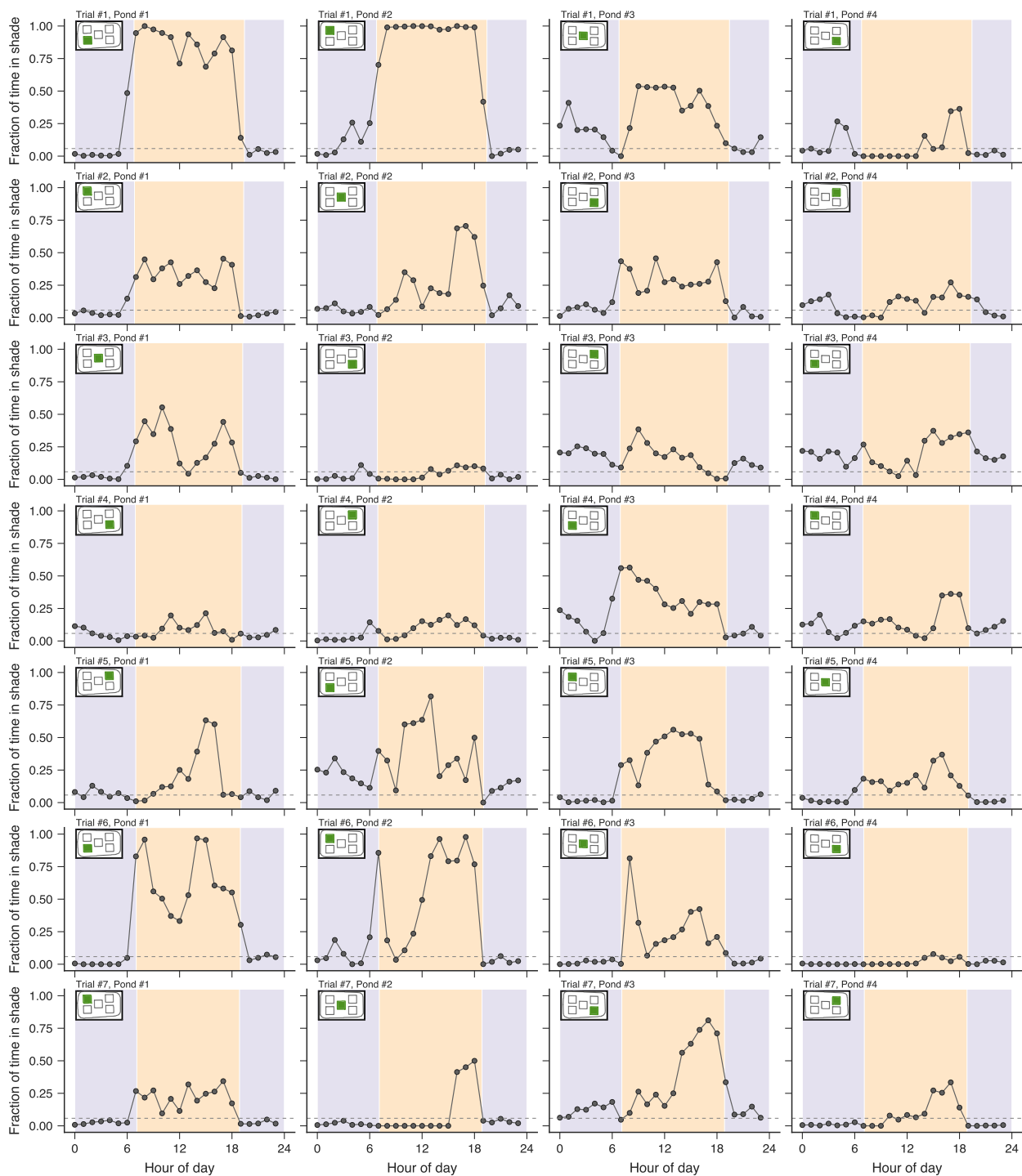


Figura D.2: Diferenças na seleção de sombra em função da hora do dia em cada lagoa e ensaio. Os círculos mostram a fração de todas as posições dos peixes registradas sob as estruturas de sombra dentro de uma janela de uma hora para cada lagoa e ensaio (como mostrado no canto superior esquerdo dos painéis). A linha tracejada representa a fração esperada de posições sob sombra se os peixes se movessem aleatoriamente pelas lagoas, e as cores de fundo indicam o dia (laranja) e a noite (roxo). Pequenos inserções em cada painel mostram a localização da lagoa e da sombra em cada ensaio e lagoa.

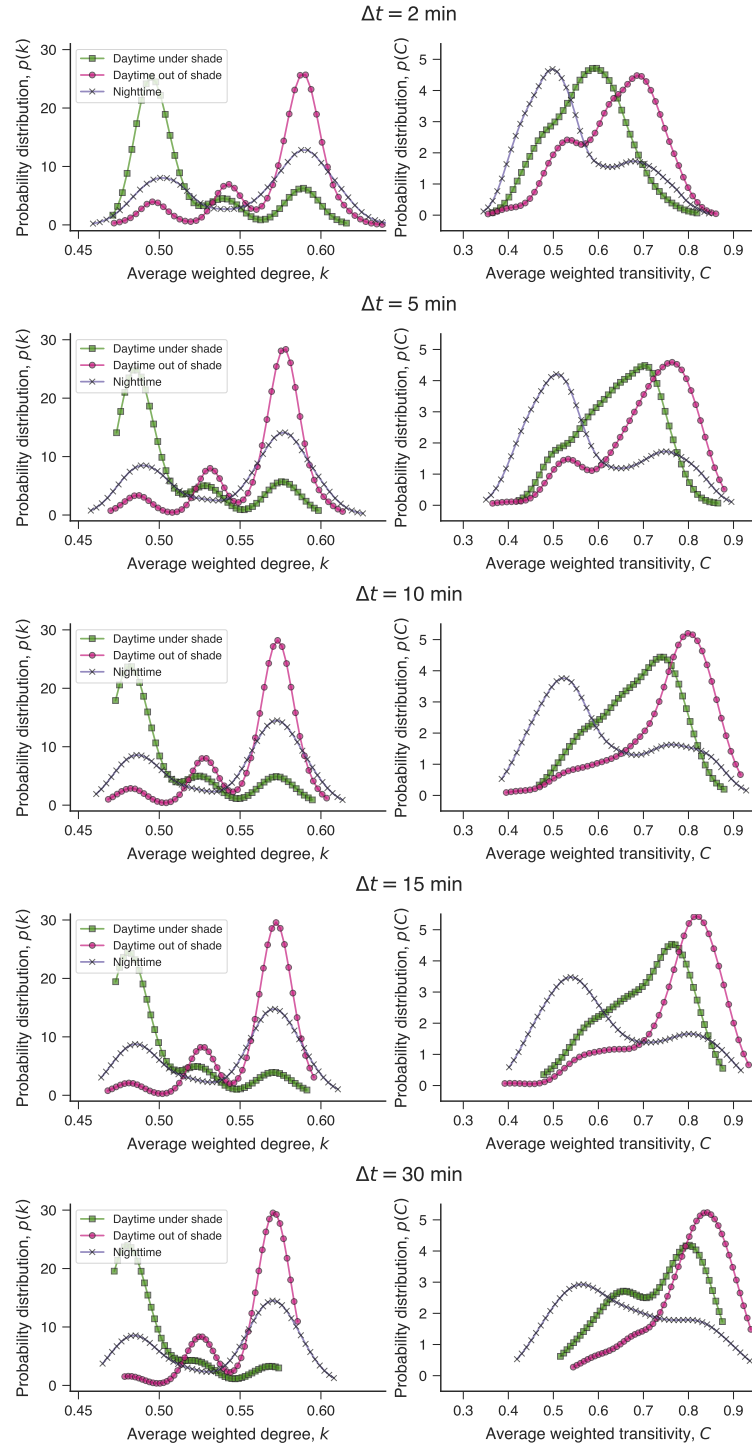


Figura D.3: Análise de sensibilidade em cinco intervalos de tempo de medidas de rede. Cada linha de gráficos mostra as distribuições de probabilidade das mesmas medidas de rede representadas na Figura 4.4 nos diversos intervalos de tempo Δt usados para criar as redes ponderadas (os valores estão indicados dentro dos gráficos). Observamos que essas distribuições não diferem significativamente entre os diferentes intervalos de tempo Δt .

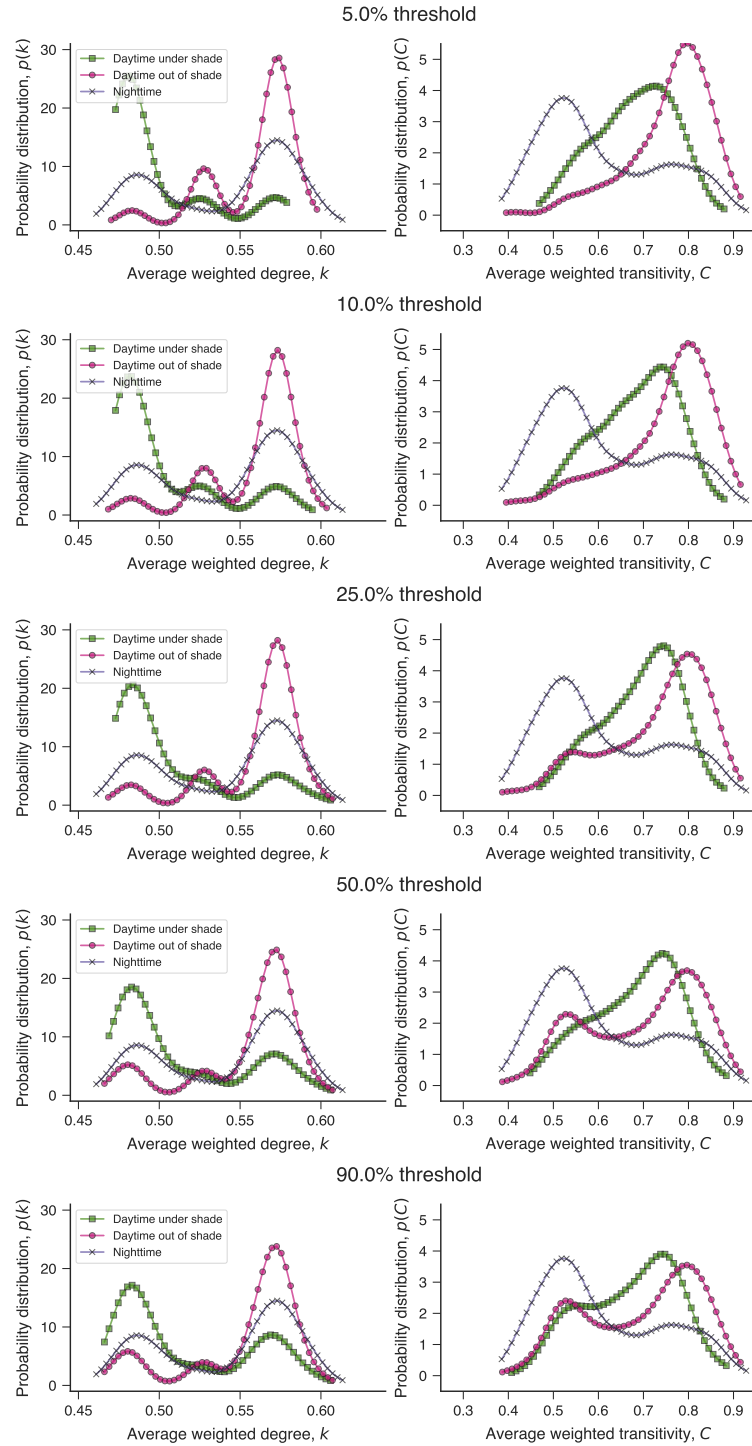


Figura D.4: Análise de sensibilidade em cinco limiares de medidas de rede usadas para definir a categoria sob-sombra na análise relacionada às redes de interação. Cada linha de gráficos mostra as distribuições de probabilidade das mesmas medidas de rede representadas na Figura 4.4 para vários limiares (os valores estão indicados dentro dos gráficos). Observamos que essas distribuições não diferem significativamente entre os diferentes limiares.

Referências Bibliográficas

- [1] Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
- [2] Kwapień, J. & Drożdż, S. Physical approach to complex systems. *Physics Reports* **515**, 115–226 (2012).
- [3] Mitchell, M. Complex systems: Network thinking. *Artificial Intelligence* **170**, 1194–1212 (2006).
- [4] Kobayashi, N., Kuninaka, H., Wakita, J.-i. & Matsushita, M. Statistical features of complex systems toward establishing sociological physics. *Journal of the Physical Society of Japan* **80**, 072001 (2011).
- [5] Boccaro, N. *Modeling complex systems* (Springer Science & Business Media, 2010).
- [6] Arthur, W. B. Complexity and the economy. *Science* **284**, 107–109 (1999).
- [7] Lillo, F. & Mantegna, R. N. Power-law relaxation in a complex system: Omori law after a financial market crash. *Physical Review E* **68**, 016119 (2003).
- [8] Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T. & DeAngelis, D. L. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science* **310**, 987–991 (2005).
- [9] Guimera, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433**, 895 (2005).
- [10] Hamilton, M. J., Milne, B. T., Walker, R. S., Burger, O. & Brown, J. H. The complex structure of hunter–gatherer social networks. *Proceedings of the Royal Society B: Biological Sciences* **274**, 2195–2203 (2007).

- [11] Mason, O. & Verwoerd, M. Graph theory and networks in biology. *IET Systems Biology* **1**, 89–119 (2007).
- [12] May, R. M., Levin, S. A. & Sugihara, G. Complex systems: Ecology for bankers. *Nature* **451**, 893 (2008).
- [13] Picoli Jr, S. & Mendes, R. S. Universal features in the growth dynamics of religious activities. *Physical Review E* **77**, 036105 (2008).
- [14] Donges, J. F., Zou, Y., Marwan, N. & Kurths, J. Complex networks in climate dynamics. *The European Physical Journal Special Topics* **174**, 157–179 (2009).
- [15] Galea, S., Hall, C. & Kaplan, G. A. Social epidemiology and complex system dynamic modelling as applied to health behaviour and drug use research. *International Journal of Drug Policy* **20**, 209–216 (2009).
- [16] Tabak, B., Serra, T. & Cajueiro, D. Topological properties of commodities networks. *The European Physical Journal B* **74**, 243–249 (2010).
- [17] Garrett, K. A. *et al.* Complexity in climate-change impacts: An analytical framework for effects mediated by plant disease. *Plant Pathology* **60**, 15–30 (2011).
- [18] Xiong, F., Liu, Y., Zhang, Z.-j., Zhu, J. & Zhang, Y. An information diffusion model based on retweeting mechanism for online social media. *Physics Letters A* **376**, 2103–2108 (2012).
- [19] Cervantes-De la Torre, F., González-Trejo, J. I., Real-Ramirez, C. A. & Hoyos-Reyes, L. F. Fractal dimension algorithms and their application to time series associated with natural phenomena. In *Journal of Physics: Conference Series*, vol. 475, 012002 (IOP Publishing, 2013).
- [20] Bittencourt, N., Meeuwisse, W., Mendonça, L., Nettel-Aguirre, A., Ocarino, J. & Fonseca, S. Complex systems approach for sports injuries: Moving from risk factor identification to injury pattern recognition—narrative review and new concept. *British Journal of Sports Medicine* **50**, 1309–1314 (2016).
- [21] Brailsford, S., Lattimer, V., Tarnaras, P. & Turnbull, J. Emergency and on-demand healthcare: Modeling a large complex system. In *Operational Research for Emergency Planning in Healthcare: Volume 2*, 13–30 (Springer, 2016).
- [22] Fry, J. & Cheah, E.-T. Negative bubbles and shocks in cryptocurrency markets. *International Review of Financial Analysis* **47**, 343–352 (2016).

- [23] Carlson, J. M., Langer, J. S. & Shaw, B. E. Dynamics of earthquake faults. *Reviews of Modern Physics* **66**, 657 (1994).
- [24] Shaw, B. E. Generalized Omori Law for aftershocks and foreshocks from a simple dynamics. *Geophysical Research Letters* **20**, 907–910 (1993).
- [25] Bak, P., Christensen, K., Danon, L. & Scanlon, T. Unified scaling law for earthquakes. *Physical Review Letters* **88**, 178501 (2002).
- [26] Christian, R. R. & Luczkovich, J. J. Organizing and understanding a winter’s seagrass foodweb network through effective trophic levels. *Ecological Modelling* **117**, 99–124 (1999).
- [27] Bascompte, J. & Melián, C. J. Simple trophic modules for complex food webs. *Ecology* **86**, 2868–2873 (2005).
- [28] Woodward, G., Papantoniou, G., Edwards, F. & Lauridsen, R. B. Trophic trickles and cascades in a complex food web: Impacts of a keystone predator on stream community structure and ecosystem processes. *Oikos* **117**, 683–692 (2008).
- [29] Matia, K., Ashkenazy, Y. & Stanley, H. E. Multifractal properties of price fluctuations of stocks and commodities. *EPL (Europhysics Letters)* **61**, 422 (2003).
- [30] Farmer, J. D., Gallegati, M., Hommes, C., Kirman, A., Ormerod, P., Cincotti, S., Sanchez, A. & Helbing, D. A complex systems approach to constructing better models for managing financial markets and the economy. *The European Physical Journal Special Topics* **214**, 295–324 (2012).
- [31] Pessa, A. A., Perc, M. & Ribeiro, H. V. Age and market capitalization drive large price variations of cryptocurrencies. *Scientific Reports* **13**, 3351 (2023).
- [32] Bariviera, A. F., Zunino, L. & Rosso, O. A. An analysis of high-frequency cryptocurrencies prices dynamics using permutation-information-theory quantifiers. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28**, 075511 (2018).
- [33] Sigaki, H. Y., Perc, M. & Ribeiro, H. V. Clustering patterns in efficiency and the coming-of-age of the cryptocurrency market. *Scientific Reports* **9**, 1440 (2019).
- [34] Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Physical Review E* **68**, 065103 (2003).
- [35] Ding, K., Jiang, P., Leng, J. & Cao, W. Modeling and analyzing of an enterprise relationship network in the context of social manufacturing. *Proceedings of the Institution*

- of Mechanical Engineers, Part B: Journal of Engineering Manufacture* **230**, 752–769 (2016).
- [36] Cai, M., Wang, W., Cui, Y. & Stanley, H. E. Multiplex network analysis of employee performance and employee social relationships. *Physica A: Statistical Mechanics and its Applications* **490**, 1–12 (2018).
 - [37] Alves, L. G., Ribeiro, H. V., Lenzi, E. K. & Mendes, R. S. Distance to the scaling law: A useful approach for unveiling relationships between crime and urban metrics. *PLoS One* **8**, e69580 (2013).
 - [38] Alves, L. G., Ribeiro, H. V. & Rodrigues, F. A. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications* **505**, 435–443 (2018).
 - [39] Economist, T. Data, data everywhere (2010). URL <https://www.economist.com/special-report/2010/02/27/data-data-everywhere>.
 - [40] Bakker, E. M. Open and free datasets for multimedia retrieval. *International Journal of Multimedia Information Retrieval* **5**, 135–136 (2016).
 - [41] Gandomi, A. & Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* **35**, 137–144 (2015).
 - [42] Rossum, G. Python reference manual (1995).
 - [43] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
 - [44] McKinney, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython* ("O'Reilly Media, Inc.", 2012).
 - [45] Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T. & Boyle, J. M. A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education* **59**, 661–686 (2012).
 - [46] Eglesz, D., Fekete, I., Kiss, O. E. & Izso, L. Computer games are fun? On professional games and players' motivations. *Educational Media International* **42**, 117–124 (2005).
 - [47] Cunha, L. R., Mendes, L. O. & Mendes, R. S. Complexity of popularity and dynamics of within-game achievements in computer games. *arXiv e-prints* arXiv-2404 (2024).
 - [48] Cunha, L. R., Pessa, A. A. & Mendes, R. S. Shape patterns in popularity series of video games. *Chaos, Solitons & Fractals* **185**, 115081 (2024).

- [49] Duenas, M.-A., Hemming, D. J., Roberts, A. & Diaz-Soltero, H. The threat of invasive species to iucn-listed critically endangered species: A systematic review. *Global Ecology and Conservation* **26**, e01476 (2021).
- [50] Sakai, A. K. *et al.* The population biology of invasive species. *Annual Review of Ecology and Systematics* **32**, 305–332 (2001).
- [51] Lu, G., Wang, C., Zhao, J., Liao, X., Wang, J., Luo, M., Zhu, L., Bernatzhez, L. & Li, S. Evolution and genetics of bighead and silver carps: Native population conservation versus invasive species control. *Evolutionary Applications* **13**, 1351–1362 (2020).
- [52] Williamson, C. J. & Garvey, J. E. Growth, fecundity, and diets of newly established silver carp in the middle mississippi river. *Transactions of the American Fisheries Society* **134**, 1423–1430 (2005).
- [53] Ribeiro, H. V., Acre, M. R., Faulkner, J. D., da Cunha, L. R., Lawson, K. M., Wamboldt, J. J., Brey, M. K., Woodley, C. M. & Calfee, R. D. Effects of shady environments on fish collective behavior. *Scientific Reports* **12**, 17873 (2022).
- [54] Cunha, L. R., Mendes, L. O. & Mendes, R. S. Complexity of popularity and dynamics of within-game achievements in computer games. *International Journal of Modern Physics C* **0**, 2450112 (2024). <https://doi.org/10.1142/S0129183124501122>.
- [55] Tomé, T. & De Oliveira, M. J. *Stochastic dynamics and irreversibility* (Springer, 2015).
- [56] Ross, S. M. *Introduction to probability models* (Academic press, 2014).
- [57] Kalbfleisch, J. D. & Prentice, R. L. *The statistical analysis of failure time data* (John Wiley & Sons, 2011).
- [58] Zwet, W. R. v. Mean, median, mode ii. *Statistica Neerlandica* **33**, 1–5 (1979).
- [59] Blitzstein, J. & Hwang, J. *Introduction to Probability*. Chapman & Hall/CRC Texts in Statistical Science (CRC Press, 2014).
- [60] Casella, G. & Berger, R. L. *Statistical inference* (Cengage Learning, 2021).
- [61] Lehmann, E. L. & Casella, G. *Theory of point estimation* (Springer Science & Business Media, 2006).
- [62] Myung, I. J. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* **47**, 90–100 (2003).

- [63] Neyman, J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **236**, 333–380 (1937).
- [64] Hesterberg, T. Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics* **3**, 497–526 (2011).
- [65] Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap* (CRC press, 1994).
- [66] Björck, Å. Least squares methods. *Handbook of numerical analysis* **1**, 465–652 (1990).
- [67] Draper, N. R. & Smith, H. *Applied regression analysis*, vol. 326 (John Wiley & Sons, 1998).
- [68] Miles, J. R-squared, adjusted R-squared. *Encyclopedia of statistics in behavioral science* (2005).
- [69] Smith, B. T. & Hess, T. M. The impact of motivation and task difficulty on resource engagement: Differential influences on cardiovascular responses of young and older adults. *Motivation Science* **1**, 22 (2015).
- [70] Chatelain, M., Silvestrini, N. & Gendolla, G. H. Task difficulty moderates implicit fear and anger effects on effort-related cardiac response. *Biological Psychology* **115**, 94–100 (2016).
- [71] DiCerbo, K. E. Game-based assessment of persistence. *Journal of Educational Technology & Society* **17**, 17–28 (2014).
- [72] Lucas, G. M., Gratch, J., Cheng, L. & Marsella, S. When the going gets tough: Grit predicts costly perseverance. *Journal of Research in Personality* **59**, 15–22 (2015).
- [73] Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A. & Goldberg, L. R. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science* **2**, 313–345 (2007).
- [74] Poropat, A. E. A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin* **135**, 322 (2009).
- [75] Andersson, H. & Bergman, L. R. The role of task persistence in young adolescence for successful educational and occupational attainment in middle adulthood. *Developmental Psychology* **47**, 950 (2011).

- [76] Wijman, T. Global games market to generate \$175.8 billion in 2021; despite a slight decline, the market is on track to surpass \$200 billion in 2023 (2021). URL <https://newzoo.com/insights/articles/global-games-market-to-generate-175-8-billion-in-2021-despite-a-slight-decline-the-market-is-on-track-to-surpass-200-billion-in-2023/>.
- [77] Kuss, D. J., Griffiths, M. D. & Pontes, H. M. Chaos and confusion in dsm-5 diagnosis of internet gaming disorder: Issues, concerns, and recommendations for clarity in the field. *Journal of Behavioral Addictions* **6**, 103–109 (2017).
- [78] Hamari, J. & Koivisto, J. Measuring flow in gamification: Dispositional flow scale-2. *Computers in Human Behavior* **40**, 133–143 (2014).
- [79] Qian, M. & Clark, K. R. Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior* **63**, 50–58 (2016).
- [80] Kahn, A. S., Shen, C., Lu, L., Ratan, R. A., Coary, S., Hou, J., Meng, J., Osborn, J. & Williams, D. The trojan player typology: A cross-genre, cross-cultural, behaviorally validated scale of video game play motivations. *Computers in Human Behavior* **49**, 354–361 (2015).
- [81] Csikszentmihalyi, M. & LeFevre, J. Optimal experience in work and leisure. *Journal of Personality and Social Psychology* **56**, 815 (1989).
- [82] Keller, J. & Bless, H. Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality and Social Psychology Bulletin* **34**, 196–209 (2008).
- [83] Jin, S.-A. A. Toward integrative models of flow: Effects of performance, skill, challenge, playfulness, and presence on flow in video games. *Journal of Broadcasting & Electronic Media* **56**, 169–186 (2012).
- [84] Hessing, D. J., Elffers, H. & Weigel, R. H. Exploring the limits of self-reports and reasoned action: An investigation of the psychology of tax evasion behavior. *Journal of Personality and Social Psychology* **54**, 405 (1988).
- [85] Devaux, M. & Sassi, F. Social disparities in hazardous alcohol use: self-report bias may lead to incorrect estimates. *The European Journal of Public Health* **26**, 129–134 (2016).
- [86] Newman, M., Barabási, A.-L. & Watts, D. J. *The structure and dynamics of networks* (Princeton university press, 2011).

- [87] Szell, M., Lambiotte, R. & Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* **107**, 13636–13641 (2010).
- [88] Games, E. Epic games store (2023). URL <https://www.epicgames.com/store/>.
- [89] Projekt, C. Gog (2023). URL <https://www.gog.com/>.
- [90] Arts, E. Origin (2023). URL <https://www.origin.com/>.
- [91] Valve. Steam (2020). URL <https://store.steampowered.com/>.
- [92] Teng, C.-I. Strengthening loyalty of online gamers: Goal gradient perspective. *International Journal of Electronic Commerce* **21**, 128–147 (2017).
- [93] Teng, C.-I. Managing gamer relationships to enhance online gamer loyalty: The perspectives of social capital theory and self-perception theory. *Computers in Human Behavior* **79**, 59–67 (2018).
- [94] Butt, D. & Wilson, A. A study of the radioactive decay law. *Journal of Physics A: General Physics* **5**, 1248 (1972).
- [95] Norman, E. B., Gazes, S. B., Crane, S. G. & Bennett, D. A. Tests of the exponential decay law at short and long times. *Physical Review Letters* **60**, 2246 (1988).
- [96] Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C., Barabási, A.-L. & Hidalgo, C. A. The universal decay of collective memory and attention. *Nature Human Behaviour* **3**, 82–91 (2019).
- [97] Valve. Steam stats (2020). URL <https://store.steampowered.com/stats>.
- [98] Valve. Steam (2020). URL <https://steamcommunity.com/dev?l=english>.
- [99] AStats. Astats (2021). URL <https://astats.astats.nl/astats/index.php>.
- [100] Wilde, T. The life of a top steam achievement hunter. *PC Gamer* **9** (2016).
- [101] Yee, N. Motivations for play in online games. *CyberPsychology & Behavior* **9**, 772–775 (2006).
- [102] Hoffman, B. & Nadelson, L. Motivational engagement and video gaming: A mixed methods study. *Educational Technology Research and Development* **58**, 245–270 (2010).

- [103] Boyle, E. A., Connolly, T. M., Hailey, T. & Boyle, J. M. Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior* **28**, 771–780 (2012).
- [104] Billieux, J., Van der Linden, M., Achab, S., Khazaal, Y., Paraskevopoulos, L., Zullino, D. & Thorens, G. Why do you play world of warcraft? An in-depth exploration of self-reported motivations to play online and in-game behaviours in the virtual world of Azeroth. *Computers in Human Behavior* **29**, 103–109 (2013).
- [105] Dindar, M. Do people play mmorpgs for extrinsic or intrinsic rewards? *Telematics and Informatics* **35**, 1877–1886 (2018).
- [106] Johnson, D., Gardner, M. J. & Perry, R. Validation of two game experience scales: the player experience of need satisfaction (pens) and game experience questionnaire (geq). *International Journal of Human-Computer Studies* **118**, 38–46 (2018).
- [107] Sole, R. Valve will limit to a maximum of 100 achievements in games published on steam (2018). URL <https://hardwaresfera.com/en/noticias/videojuegos/valve-\protect\@normalcr\relaxlimitara-a-un-maximo-de-100-logros-en-los-juegos-publicados-\protect\@normalcr\relaxen-steam/>.
- [108] Barten, A. P. The coefficient of determination for regression without a constant term. In *The Practice of Econometrics*, 181–189 (Springer, 1987).
- [109] Crow, E. L. & Shimizu, K. *Lognormal distributions* (Marcel Dekker New York, 1987).
- [110] Fortunato, S. & Castellano, C. Scaling and universality in proportional elections. *Physical Review Letters* **99**, 138701 (2007).
- [111] Bauckhage, C. Insights into internet memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, 42–49 (2011).
- [112] Van Mieghem, P., Blenn, N. & Doerr, C. Lognormal distribution in the digg online social network. *The European Physical Journal B* **83**, 251–261 (2011).
- [113] Yip, M., Shadbolt, N. & Webber, C. Structural analysis of online criminal social networks. In *2012 IEEE International Conference on Intelligence and Security Informatics*, 60–65 (IEEE, 2012).
- [114] Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Raubal, M. & Giannopoulos, I. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13 (2018).

- [115] Strojny, P., Strojny, A. & Rebilas, K. Player involvement as a result of difficulty: An introductory study to test the suitability of the motivational intensity approach to video game research. *PLoS One* **18**, e0282966 (2023).
- [116] Csiszár, I. & Körner, J. *Information theory: coding theorems for discrete memoryless systems* (Cambridge University Press, 2011).
- [117] Schultheiss, D. Long-term motivations to play mmogs: A longitudinal study on motivations, experience and behavior. In *Proceedings of DiGRA 2007 Conference: Situated Play* (2007).
- [118] Reinhard, C. D. & Dervin, B. Situational and gender comparisons of digital game players' preferences for game features and gratifications. In *In the annual Meeting of the International Communication Association Meeting* (2009).
- [119] Shute, V. & Ventura, M. *Stealth assessment: Measuring and supporting learning in video games* (The MIT Press, 2013).
- [120] Ventura, M., Shute, V. & Zhao, W. The relationship between video game use and a performance-based measure of persistence. *Computers & Education* **60**, 52–58 (2013).
- [121] Turp, D. *La contribution du droit international au maintien de la diversité culturelle* (Martinus Nijhoff, 2013).
- [122] Ferdman, B. Literacy and cultural identity. *Harvard Educational Review* **60**, 181–205 (1990).
- [123] Fernández-Llamazares, Á. & Cabeza, M. Rediscovering the potential of indigenous storytelling for conservation practice. *Conservation Letters* **11**, e12398 (2018).
- [124] Griswold, W. *Cultures and societies in a changing world* (Sage, 2012).
- [125] Lorenz-Spreen, P., Mønsted, B. M., Hövel, P. & Lehmann, S. Accelerating dynamics of collective attention. *Nature Communications* **10**, 1759 (2019).
- [126] Sorensen, A. T. Bestseller lists and product variety. *The journal of industrial economics* **55**, 715–738 (2007).
- [127] Berger, J. & Milkman, K. L. What makes online content viral? *Journal of marketing research* **49**, 192–205 (2012).
- [128] Gleeson, J. P., O'Sullivan, K. P., Baños, R. A. & Moreno, Y. Effects of network structure, competition and memory time on social spreading phenomena. *Physical Review X* **6**, 021019 (2016).

- [129] D O’Brien, J., Dassios, I. K. & Gleeson, J. P. Spreading of memes on multiplex networks. *New Journal of Physics* **21**, 025001 (2019).
- [130] Lazer, D. M. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
- [131] Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *science* **359**, 1146–1151 (2018).
- [132] Paek, H.-J. & Pan, Z. Spreading global consumerism: Effects of mass media and advertising on consumerist values in China. *Mass Communication & Society* **7**, 491–515 (2004).
- [133] Salganik, M. J. *Bit by bit: Social research in the digital age* (Princeton University Press, 2019).
- [134] Chang, H.-C. A new perspective on twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology* **47**, 1–4 (2010).
- [135] Bruns, A. & Stieglitz, S. Towards more systematic twitter analysis: metrics for tweeting activities. *International journal of social research methodology* **16**, 91–108 (2013).
- [136] Eom, Y.-H. & Fortunato, S. Characterizing and modeling citation dynamics. *PLoS One* **6**, e24926 (2011).
- [137] Bello, P. & Garcia, D. Cultural divergence in popular music: the increasing diversity of music consumption on spotify across countries. *Humanities and Social Sciences Communications* **8**, 1–8 (2021).
- [138] Miz, V., Benzi, K., Ricaud, B. & Vandergheynst, P. Wikipedia graph mining: dynamic structure of collective memory. *arXiv preprint arXiv:1710.00398* (2017).
- [139] Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E. & Preis, T. Quantifying wikipedia usage patterns before stock market moves. *Scientific reports* **3**, 1801 (2013).
- [140] Crane, R. & Sornette, D. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* **105**, 15649–15653 (2008).
- [141] Muriel, D. & Crawford, G. *Video games as culture: Considering the role and importance of video games in contemporary society* (Routledge, 2018).
- [142] Jovanovic, B. Gamer demographics: Facts and stats about the most popular hobby in the world. *Retrieved from DataProt. August* **2** (2022).

- [143] Sifa, R., Drachen, A. & Bauckhage, C. Large-scale cross-game player behavior analysis on Steam. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* **11**, 198–204 (2015).
- [144] Zendle, D., Flick, C., Halgarth, D., Ballou, N., Demediuk, S. & Drachen, A. Cross-cultural patterns in mobile playtime: an analysis of 118 billion hours of human data. *Scientific reports* **13**, 386 (2023).
- [145] Lin, D., Bezemer, C.-P. & Hassan, A. E. An empirical study of early access games on the Steam platform. *Empirical Software Engineering* **23**, 771–799 (2018).
- [146] Lin, D., Bezemer, C.-P., Zou, Y. & Hassan, A. E. An empirical study of game reviews on the Steam platform. *Empirical Software Engineering* **24**, 170–207 (2019).
- [147] Ahn, S., Kang, J. & Park, S. What makes the difference between popular games and unpopular games? Analysis of online game reviews from steam platform using word2vec and bass model. *ICIC Express Letters* **11**, 1729–1737 (2017).
- [148] Mendes, L. O., Cunha, L. R. & Mendes, R. S. Popularity of video games and collective memory. *Entropy* **24**, 860 (2022).
- [149] Gidea, M., Goldsmith, D., Katz, Y., Roldan, P. & Shmalo, Y. Topological recognition of critical transitions in time series of cryptocurrencies. *Physica A: Statistical Mechanics and its Applications* **548**, 123843 (2020).
- [150] Sunahara, A. S., Perc, M. c. v. & Ribeiro, H. V. Universal productivity patterns in research careers. *Physical Review Research* **5**, 043203 (2023).
- [151] Pessa, A. A., Perc, M. & Ribeiro, H. V. Clustering free-falling paper motion with complexity and entropy. *Europhysics Letters* **138**, 30003 (2022).
- [152] Zhou, S., Wang, X., Zhou, W. & Zhang, C. Recognition of the scale-free interval for calculating the correlation dimension using machine learning from chaotic time series. *Physica A: Statistical Mechanics and its Applications* **588**, 126563 (2022).
- [153] Gray, J. Steam charts-tracking what’s played (2012). URL <https://steamcharts.com/>.
- [154] Massey Jr, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* **46**, 68–78 (1951).
- [155] Wagenmakers, E.-J., Grasman, R. P. & Molenaar, P. C. On the relation between the mean and the variance of a diffusion model response time distribution. *Journal of Mathematical Psychology* **49**, 195–204 (2005).

- [156] Ito, K. & Xiong, K. Gaussian filters for nonlinear filtering problems. *IEEE transactions on automatic control* **45**, 910–927 (2000).
- [157] Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* **26**, 43–49 (1978).
- [158] Meert, W., Hendrickx, K., Van Craenendonck, T., Robberechts, P., Blockeel, H. & Davis, J. Dtaidistance (2020). URL <https://doi.org/10.5281/zenodo.7158824>.
- [159] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- [160] McInnes, L., Healy, J., Saul, N. & Grossberger, L. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software* **3**, 861 (2018).
- [161] Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences* **105**, 1118–1123 (2008).
- [162] Rosvall, M., Axelsson, D. & Bergstrom, C. T. The map equation. *The European Physical Journal Special Topics* **178**, 13–23 (2009).
- [163] Edler, D., Holmgren, A. & Rosvall, M. The MapEquation software package (2022). URL <https://mapequation.org>.
- [164] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987).
- [165] Lee, E. K., Balasubramanian, H., Tsolias, A., Anakwe, S. U., Medalla, M., Shenoy, K. V. & Chandrasekaran, C. Non-linear dimensionality reduction on extracellular waveforms reveals cell type diversity in premotor cortex. *Elife* **10**, e67490 (2021).
- [166] Wu, F. & Huberman, B. A. Novelty and collective attention. *Proceedings of the National Academy of Sciences* **104**, 17599–17601 (2007).
- [167] Yook, S.-H. & Kim, Y. Origin of the log-normal popularity distribution of trending memes in social networks. *Physical Review E* **101**, 012312 (2020).
- [168] Dudley, R. M. Central limit theorems for empirical measures. *The Annals of Probability* 899–929 (1978).
- [169] Downey, A. B. *Probably Overthinking It: How to Use Data to Answer Questions, Avoid Statistical Traps, and Make Better Decisions* (University of Chicago Press, 2023).

- [170] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
- [171] Hawkins, D. M. *Identification of outliers*, vol. 11 (Springer, 1980).
- [172] Ke, Q., Ferrara, E., Radicchi, F. & Flammini, A. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences* **112**, 7426–7431 (2015).
- [173] Ozer, M., Sapienza, A., Abeliuk, A., Muric, G. & Ferrara, E. Discovering patterns of online popularity from time series. *Expert Systems with Applications* **151**, 113337 (2020).
- [174] Olsson, B. & Sidenblom, L. *Business models for video games*. Master’s thesis, Department of Informatics, Lund University, Sweden (2010).
- [175] Camazine, S., Deneubourg, J.-L., Franks, N. R., Sneyd, J., Theraula, G. & Bonabeau, E. *Self-Organization in Biological Systems* (Princeton University Press, Princeton, 2001).
- [176] Sumpter, D. J. *Collective Animal Behavior* (Princeton University Press, Princeton, 2010).
- [177] Vicsek, T. & Zafeiris, A. Collective motion. *Physics Reports* **517**, 71–140 (2012).
- [178] Toner, J. & Tu, Y. Long-range order in a two-dimensional dynamical XY model: How birds fly together. *Physical Review Letters* **75**, 4326 (1995).
- [179] Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M. & Walczak, A. M. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences* **109**, 4786–4791 (2012).
- [180] Gautrais, J., Ginelli, F., Fournier, R., Blanco, S., Soria, M., Chaté, H. & Theraulaz, G. Deciphering interactions in moving animal groups. *PLOS Computational Biology* **8**, e1002678 (2012).
- [181] Mateo, D., Kuan, Y. K. & Bouffanais, R. Effect of correlations in swarms on collective response. *Scientific Reports* **7**, 10388 (2017).
- [182] Charlesworth, H. J. & Turner, M. S. Intrinsically motivated collective motion. *Proceedings of the National Academy of Sciences* **116**, 15362–15367 (2019).
- [183] Heupel, M., Semmens, J. M. & Hobday, A. Automated acoustic tracking of aquatic animals: Scales, design and deployment of listening station arrays. *Marine and Freshwater Research* **57**, 1–13 (2006).

- [184] Straw, A. D., Branson, K., Neumann, T. R. & Dickinson, M. H. Multi-camera real-time three-dimensional tracking of multiple flying animals. *Journal of The Royal Society Interface* **8**, 395–409 (2011).
- [185] Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S. & de Polavieja, G. G. idTracker: Tracking individuals in a group by automatic identification of unmarked animals. *Nature Methods* **11**, 743–748 (2014).
- [186] Walter, T. & Couzin, I. D. TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *eLife* **10**, e64000 (2021).
- [187] Tuia, D. *et al.* Perspectives in machine learning for wildlife conservation. *Nature Communications* **13**, 792 (2022).
- [188] Jetz, W. *et al.* Biological earth observation with animal sensors. *Trends in Ecology & Evolution* **37**, 293–298 (2022).
- [189] Jiang, L., Giuggioli, L., Perna, A., Escobedo, R., Lecheval, V., Sire, C., Han, Z. & The-raulaz, G. Identifying influential neighbors in animal flocking. *PLOS Computational Biology* **13**, e1005822 (2017).
- [190] Herbert-Read, J. E. *et al.* How predation shapes the social interaction rules of shoaling fish. *Proceedings of the Royal Society B: Biological Sciences* **284**, 20171126 (2017).
- [191] Papadopoulou, M., Hildenbrandt, H., Sankey, D. W., Portugal, S. J. & Hemelrijk, C. K. Self-organization of collective escape in pigeon flocks. *PLOS Computational Biology* **18**, e1009772 (2022).
- [192] Nuzhin, E. E., Panov, M. E. & Brilliantov, N. V. Why animals swirl and how they group. *Scientific Reports* **11**, 1–15 (2021).
- [193] Wild, B., Dormagen, D. M., Zachariae, A., Smith, M. L., Traynor, K. S., Brockmann, D., Couzin, I. D. & Landgraf, T. Social networks predict the life and death of honey bees. *Nature Communications* **12**, 1–12 (2021).
- [194] Schaerf, T. M., Dillingham, P. W. & Ward, A. J. The effects of external cues on individual and collective behavior of shoaling fish. *Science Advances* **3**, e1603201 (2017).
- [195] Crook, D. & Robertson, A. Relationships between riverine fish and woody debris: Implications for lowland rivers. *Marine and Freshwater Research* **50**, 941–953 (1999).
- [196] McCartt, A. L., Lynch, W. E. & Johnson, D. L. How light, a predator, and experience influence bluegill use of shade and schooling. *Environmental Biology of Fishes* **49**, 79–87 (1997).

- [197] Fritts, A. K. *et al.* Lock operations influence upstream passages of invasive and native fishes at a Mississippi River high-head dam. *Biological Invasions* **23**, 771–794 (2021).
- [198] Cocheret De La Morinière, E., Nagelkerken, I., Van Der Meij, H. & Van Der Velde, G. What attracts juvenile coral reef fish to mangroves: Habitat complexity or shade? *Marine Biology* **144**, 139–145 (2004).
- [199] McMahon, T. E. & Hartman, G. F. Influence of cover complexity and current velocity on winter habitat use by juvenile coho salmon (*Oncorhynchus kisutch*). *Canadian Journal of Fisheries and Aquatic Sciences* **46**, 1551–1557 (1989).
- [200] Jones, N. A., Spence, R., Jones, F. A. & Spence-Jones, H. C. Shade as enrichment: Testing preferences for shelter in two model fish species. *Journal of Fish Biology* **95**, 1161–1165 (2019).
- [201] Scherer, E. & McNicol, R. Preference-avoidance responses of lake whitefish (*Coregonus clupeaformis*) to competing gradients of light and copper, lead, and zinc. *Water Research* **32**, 924–929 (1998).
- [202] Fernández-Juricic, E. & Tran, E. Changes in vigilance and foraging behaviour with light intensity and their effects on food intake and predator detection in house finches. *Animal Behaviour* **74**, 1381–1390 (2007).
- [203] Mandelik, Y., Jones, M. & Dayan, T. Structurally complex habitat and sensory adaptations mediate the behavioural responses of a desert rodent to an indirect cue for increased predation risk. *Evolutionary Ecology Research* **5**, 501–515 (2003).
- [204] Carr, J. M. & Lima, S. L. Wintering birds avoid warm sunshine: predation and the costs of foraging in sunlight. *Oecologia* **174**, 713–721 (2014).
- [205] Sabal, M. C., Workman, M. L., Merz, J. E. & Palkovacs, E. P. Shade affects magnitude and tactics of juvenile Chinook salmon antipredator behavior in the migration corridor. *Oecologia* **197**, 89–100 (2021).
- [206] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. & Goldberger, A. L. Mosaic organization of DNA nucleotides. *Physical Review E* **49**, 1685 (1994).
- [207] Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H. A., Havlin, S. & Bunde, A. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications* **295**, 441–454 (2001).
- [208] Kolpas, A., Moehlis, J. & Kevrekidis, I. G. Coarse-grained analysis of stochasticity-induced switching between collective motion states. *Proceedings of the National Academy of Sciences* **104**, 5931–5935 (2007).

- [209] Tunstrøm, K., Katz, Y., Ioannou, C. C., Huepe, C., Lutz, M. J. & Couzin, I. D. Collective states, multistability and transitional behavior in schooling fish. *PLOS Computational Biology* **9**, e1002915 (2013).
- [210] Newman, M. *Networks: An Introduction* (Orford University Press, New York, 2010).
- [211] Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K. & Kertesz, J. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E* **75**, 027105 (2007).
- [212] Onnela, J.-P., Saramäki, J., Kertész, J. & Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Physical Review E* **71**, 065103 (2005).
- [213] Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference*, 11–15 (Pasadena, CA USA, 2008).
- [214] Strandburg-Peshkin, A. *et al.* Visual sensory networks and effective information transfer in animal groups. *Current Biology* **23**, R709–R711 (2013).
- [215] Tsallis, C., Levy, S. V., Souza, A. M. & Maynard, R. Statistical-mechanical foundation of the ubiquity of Lévy distributions in nature. *Physical Review Letters* **75**, 3589 (1995).