Heavy Tails and Anomalous Diffusion in Human Online Dynamics

Xiangwen Wang

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Physics

Michel J. Pleimling, Chair
Giti A. Khodaparast
Vito W. Scarola
Uwe C. Täuber

February 22, 2019
Blacksburg, Virginia

Keywords: Human Dynamics, Data-Driven Modeling, Heavy Tails, Random Walks,
Anomalous Diffusion

Heavy Tails and Anomalous Diffusion in Human Online Dynamics

Xiangwen Wang

(ABSTRACT)

In this dissertation, I extend the analysis of human dynamics to human movements in online activities. My work starts with a discussion of the human information foraging process based on three large collections of empirical search click-through logs collected in different time periods. With the analogy of viewing the click-through on search engine result pages as a random walk, a variety of quantities like the distributions of step length and waiting time as well as mean-squared displacements, correlations and entropies are discussed. Notable differences between the different logs reveal an increased efficiency of the search engines, which is found to be related to the vanishing of the heavy-tailed characteristics of step lengths in newer logs as well as the switch from superdiffusion to normal diffusion in the diffusive processes of the random walks. In the language of foraging, the newer logs indicate that online searches overwhelmingly yield local searches, whereas for the older logs the foraging processes are a combination of local searches and relocation phases that are power-law distributed. The investigation highlights the presence of intermittent search processes in online searches, where phases of local explorations are separated by power-law distributed relocation jumps. In the second part of this dissertation I focus on an in-depth analysis of online gambling behaviors. For this analysis the collected empirical gambling logs reveal the wide existence of heavy-tailed statistics in various quantities in different online gambling games. For example, when players are allowed to choose arbitrary bet values, the bet values present log-normal distributions, meanwhile if they are restricted to use items as wagers, the distribution becomes truncated power laws. Under the analogy of viewing the net change of income of each player as a random walk, the mean-squared displacement and

first-passage time distribution of these net income random walks both exhibit anomalous diffusion. In particular, in an online lottery game the mean-squared displacement presents a crossover from a superdiffusive to a normal diffusive regime, which is reproduced using simulations and explained analytically. This investigation also reveals the scaling characteristics and probability reweighting in risk attitude of online gamblers, which may help to interpret behaviors in economic systems. This work was supported by the US National Science Foundation through grants DMR-1205309 and DMR-1606814.

Heavy Tails and Anomalous Diffusion in Human Online Dynamics

Xiangwen Wang

(GENERAL AUDIENCE ABSTRACT)

Humans are complex, meanwhile understanding the complex human behaviors is of crucial importance in solving many social problems. In recent years, sociophysicists have made substantial progress in human dynamics research. In this dissertation, I extend this type of analysis to human movements in online activities. My work starts with a discussion of the human information foraging process. This investigation is based on empirical search logs and an analogy of viewing the click-through on search engine result pages as a random walk. With an increased efficiency of the search engines, the heavy-tailed characteristics of step lengths disappear, and the diffusive processes of the random walkers switch from superdiffusion to normal diffusion. In the language of foraging, the newer logs indicate that online searches overwhelmingly yield local searches, whereas for the older logs the foraging processes are a combination of local searches and relocation phases that are power-law distributed. The investigation highlights the presence of intermittent search processes in online searches, where phases of local explorations are separated by power-law distributed relocation jumps. In the second part of this dissertation I focus on an in-depth analysis of online gambling behaviors, where the collected empirical gambling logs reveal the wide existence of heavy-tailed statistics in various quantities. Using an analogy of viewing the net change of income of each player as a random walk, the mean-squared displacement and first-passage time distribution of these net income random walks exhibit anomalous diffusion. This investigation also reveals the scaling characteristics and probability reweighting in risk attitude of online gamblers, which may help to interpret behaviors in economic systems. This work was supported by the US National Science Foundation through grants DMR-1205309 and DMR-1606814.

Dedication

To

my wife, Dr. Linjun Li

my mother, Huifang Li

and my father, Sheng Wang

Acknowledgments

I wish to express my sincere gratitude to my advisor, Dr. Michel Pleimling, for his sustained help and guidance during my Ph.D. study, especially for his supporting of my interdisciplinary research. I benefited a lot from his insightful feedback and our discussions over the years. With his endorsement, I attended two related master programs, one in statistics and one in computer science, from which I learned the necessary knowledge for performing my research. Also, it is in his team that I met my wife Linjun.

I would like to extend my gratitude to Dr. Eric Sharpe and my committee members, Dr. Uwe Täuber, Dr. Vito Scarola, and Dr. Giti Khodaparast, for their valuable comments on my work.

I want to thank Department of Physics of Virginia Tech and Dr. Pleimling for supporting my Ph.D. study and the other M.S. programs, and Betty Wilkins and the Graduate School of Virginia Tech for their help on administrative issues.

I would like to express my thankfulness and love to my parents and my wife for their constant encouragement and their endless support throughout the years. They mean everything to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sociophysics

Human behaviors and their interactions form our human society, but humans are complex. Understanding these complex human behaviors is of crucial importance to mankind. Providing models that precisely describe human beings not only contributes to better predictions of human activities and solutions for social problems, but also helps to interpret human intelligence and explain social evolution. For such reasons, the study of human behaviors has led to the emergence of research areas including economics, sociology, politics, as well as psychology, anthropology, and behavioral science. Traditionally, those different research areas investigate different aspects of humans, but scientists begin to realize the importance of cooperation among different disciplines. Yet subject barriers still exist. Most of the studies in these areas focus on qualitative descriptions of human behaviors. However, qualitative descriptions cannot meet the requirements for precise predictions. On the other hand, statistical physicists have proposed theories which provide satisfying quantitative descriptions of systems that consist of large numbers of particles. Despite the similarities with systems that consist of large numbers of elements, usually those theories cannot be directly applied to model social problems, since compared to particles, humans have emotions and intelligence, and there are strong interactions between individuals and their peers, as well as strong interactions between individuals and the environment. Therefore new approaches based on

modifications of the theories in statistical physics are required.

Sociophysics is a field in which physicists and scientists from other disciplines work together to use the conceptual theories and tools of statistical physics to describe and interpret the complex systems resulting from human activities. Compared to traditional research, the advantage of sociophysics research resides in its possibilities to provide quantitative analysis and analytic solutions for social problems. Generally speaking, sociophysics aims to extract the characteristics of the abstract concepts and rules in social phenomena and summarize them into quantitative laws. With that said, a lot of current sociophysics studies are still exploring the qualitative features of human behaviors. For example, in social systems there is a co-existence of order and disorder. How and why transitions happen between the two states is a question needed to be addressed by sociophysicists [1]. Another example is the study of the emergence of scaling and universality in human behaviors. With the introduction of fresh physics concepts and analogies, mathematical tools and topology, nowadays sociophysics has become a large and vastly growing area, in which sociophysicists have already achieved fruitful research results addressing different social applications [1], especially in the past decade. Once again we should realize that sociophysics is not restricted to physicists, the knowledge and prior work from other disciplines are essential for building realistic individual and social models, leading to the critical needs for persistent inter-disciplinary cooperation [2].

1.2 Human dynamics

From a high-level overview, there are in general two sub-branches in sociophysics research: one is to capture the underlying mechanism behind the dynamics of the individual human behaviors, known as Human Dynamics; whereas the other one focuses more on the outcome of the interactions among individuals, known as Social Dynamics. This dissertation particularly

focuses on the analysis of human dynamics. Statistical physics is built upon the fact that the behaviors of the fundamental elements in physical systems, particles, are well understood. On the other hand, the fundamental elements in society are human beings. However, the physiological and psychological complexities of human individuals lead to erratic nature of individual behaviors, as well as to a huge diversity among individuals [1]. The complex patterns behind the behaviors still remain largely uninvestigated. Laws quantifying the activity patterns of human beings are crucial for applying statistical physics analogies and models to social problems; they therefore become an important research topic in sociophysics. There are mainly two paths in the studies of human dynamics: one is dedicated to the empirical analysis of real-world human activities for the purpose of uncovering unknown spatiotemporal patterns, based on which new assumptions and models are proposed; the other one aims at modifying existing assumptions and models, and relies on computational simulations to reproduce social phenomena. My work mainly follows the first path.

For empirical analysis in human dynamics research, many studies stick with the modeling of the patterns at the aggregate level. Studies have shown signs of the existence of solid regularities across different human activities [3], meanwhile there are also studies showing that those regularities might not hold at the individual level. Statistical physics relies on the law of large numbers [4]. Often, good estimations of the mean behaviors of elements can provide satisfying predictions on global behaviors, therefore analyses at the population level can be good starting points for the studies of human dynamics.

1.3 Data-driven research

Information technology has made tremendous progress in the past three decades, which includes the advancements of information capturing, storing, and analyzing methods and

tools. As a result, extensive amounts of information and data have been collected and digitized. For example, ubiquitous portable sensors and online loggers enable large collections of human-related data. The abundant data leads to a new methodology for performing scientific studies, data-driven research, which follows three steps: first, use automated programs to clean and parse the massive and usually unstructured raw data; second, obtain statistics, including features and relations, from the prepared data using advanced statistical methods; third, explore reasons and infer insights from the statistics obtained. Jim Gray described data-driven research as the fourth paradigm of science, after the empirical, theoretical, and computational branches, and his view has been widely accepted [5]. With this new paradigm which focuses on data-intensive systems, countless new findings emerged in different research areas. Nonetheless, there are still lots of discussions between the different methodologies for performing research (such as the comparison between data-driven research and problem-driven or hypothesis-driven research [6, 7]).

On the other hand, data-driven research also brings new challenges. One usually finds cleaning and processing large amounts of data to be difficult, not only because of the fact that empirical data usually comes with noise, correlations, and heterogeneity [8], but also due to the incomplete development of big-data analytic techniques. Luckily physicists have long been working on analyzing large and complex datasets, such as those collected in particle physics or astrophysics studies, and when physicists expand their research to sociophysics, some techniques and intuitions can still be applied. However, through data analysis one only obtains statistics or evidence, not insights. To address this, we should realize that disciplinary knowledge and theories are important for exploring reasons and inferring insights from the statistics of the data [2, 9], therefore are essential for data-driven research.

Scientific experiments and observations usually cost considerable amounts of time and resources, meanwhile the data obtained in one experiment/observation sometimes can be used

for different research topics. In recent years, there has been an increasing trend towards the open-data movements [10, 11, 12]. More and more research facilities, companies, and governments begin to make their collected data publicly available, which reduces the overall data collection workload of academic research. Sociophysics research can also take advantage of open data. The massive online activity and social statistics data, usually published by major companies or research facilities, enable the studies of human activity pattern to a large but detailed scale. With more open data being published online, more new findings related to sociophysics will emerge in the future.

1.4 Structure of this dissertation

This dissertation is structured as follows. In Chapter 1, I present high-level overviews of the research topics in this dissertation, including introductions to sociophysics, human dynamics, and data-driven research. In Chapter 2, I give literature reviews of the history and recent progress in human dynamics, which also includes the discussions of candidate distribution models and appropriate statistical methods for performing analysis. Then, three projects about human dynamics will be discussed separately, in all of which I consider certain human activities as random walk processes. This analogue enables the connection between human activities and the random walk models as well as the diffusive theories discussed in statistical physics. In Chapter 3, we analyze the click-through behaviors on search engine result pages, from which we extract the human information foraging pattern. We consider the searching process as a random walk on a one-dimensional half line and observe a switch from a superdiffusive searching pattern to a normal diffusive one when the search engine ranking algorithm is improved. In Chapter 4, we collect online gambling logs from an online lottery game, where we treat the change of gamblers' net income as a random walk, and observe a

crossover from superdiffusion to normal diffusion. This crossover can be reproduced using simulations with two conditions: finite individual wealth and conserved total wealth. We further provide analytic explanations for this crossover. In Chapter 5, we extend the analysis of gambling dynamics to more general online gambling games, in which we report some commonalities of online gambling behaviors, including the log-normal wager distribution, the scaling characteristics of risk attitude, and the anomalous diffusive pattern. Finally, in Chapter 6, I summarize the findings and potential applications of the work presented in this dissertation.

Chapter 2

Literature Review

2.1 A quick review of the history of sociophysics

Starting from the 19th century, sustained efforts have been made in studying and understanding complex human behaviors. In the 1830s, Quetelet adopted mathematical tools, including probability theory, to describe some of the basic phenomena in human society with empirical data [13], and he pointed out the wide existence of normal distributions in social problems. Around 1860, Carey published three volumes [14, 15, 16], in which he considered humans as the molecules in society, and proposed a gravity law to describe the demographic interactions in society, which involves the concentration of individuals (mass) and a "distance." During the first half of the 20th century, with the increasing development of physics theories and related mathematical tools, scientists began to realize the importance of research related to sociophysics. Typical results include Pareto's law for the social income distribution [17], and the demographic/retail gravitation theories [18, 19], etc. Around 1950, there were a series of introductory articles published by Stewart focusing on the subject "social physics," in which he listed the bases [20], suggested principles [21], concerns [22], and developments [23] of this new area. Sociophysics research began to accelerate after 1960s with the emergence of cellular automata models [24] which provide a practical tool for performing simulations of social dynamics, and the adoption of new topology, for example complex networks [25, 26], which renders the structures of social interactive models more realistic compared to those

built on lattice structures. In addition, the introduction of advanced theories and concepts in statistical physics, such as self-organization [27] and critical phenomena (order-disorder phase transition) [28], helps to better interpret complex social behaviors, therefore further boosts the research of sociophysics. Today, sociophysics has become a widely accepted science subject, and it has flourished into a wide range of research topics, which covers human dynamics [29, 30], opinion dynamics [31, 32], language dynamics [33, 34], traffic flow modeling [35, 36], etc. As this dissertation focuses mainly on human dynamics, I refer those who are interested in the other research topics to articles/books which provide more comprehensive sociophysics reviews [1, 3].

2.2 Recent findings in human dynamics

2.2.1 Temporal patterns

The burst of human dynamics research began with Barabási's *Nature* publication in 2005 [37], in which he studied the temporal patterns of email communications, and pointed out that the response time of replying emails and waiting time between sending emails both follow power-law distributions with exponents close to 1.5, meaning that the email sending/responding behaviors present short-time bursts and long-time silences. In the same paper Barabási also proposed a queuing model with a fixed and finite queue length as the generative mechanism for these power-law distributions. Later Vázquez [38] provided an exact solution for this model. There are dozens of follow-up studies confirming the wide existence of similar power-law waiting time distributions in other interactive human activities, such as in mail correspondences among famous scientists [29], message communications [39], online communications [40], etc. However, these studies also showed that the exponents are

not restricted to 1.5. At the same time, power-law distributed inter-event waiting times are also observed in non-interactive human activities, such as in web browsing [41], rating movies [30], computer usage [42], etc. Vázquez summarized the above results into interactive and non-interactive groups, respectively with an exponent of 1.5 and 1, and proposed two queuing models with different queue lengths to explain the exponents [43]. However, as mentioned earlier the exponents are not fixed for one group of activities, therefore Vázquez's conclusion is faulty. Apart from the queuing models, sociophysicists also proposed other explanations based on characteristics of human behaviors for the power-law distributions of waiting time, including memory effects [44], seasonality [45], and human interactions [46]. On the other hand, whether the waiting times indeed follow power-law distributions is still debated. Several studies [47, 48, 49] suggested that other types of heavy-tailed distributions can better describe the waiting times mentioned above, and Malmgren [50] proposed a cascading non-homogeneous Poisson model for explaining the heavy tails observed.

2.2.2 Spatial patterns

The investigation of human mobility patterns started with the publication [51] by Brockmann et al. in *Nature* in 2006 in which they indirectly studied the trajectories of individuals by tracking bank notes and observed that the jump-length distribution at the population level follows a power-law distribution. They further pointed out that the movements of bank notes follow an ultra-slow diffusion, and proposed a continuous-time random walk (CTRW) model to describe the spatial-temporal pattern of human movements. Using cellphone location data, which provides better and consistent tracking, González [52] observed that the displacements in human movements follow a power-law distribution with an exponential cutoff. The paper also revealed the bounded nature and temporal regularities in human movements. More fine-grained results were obtained by analyzing the GPS logs of human daily movements [53, 54],

and similar scaling properties were reported. Rhee [53] further observed that as the time scale increases, there is a crossover from superdiffusion to subdiffusion in human movements. Aside from the daily movements, the foraging pattern of hunters and food gatherers in Africa [55] and marine fishers [56] also presented power-law distributed displacements, similar to those observed in animal foraging processes [57]. To explain the generative mechanism of the power laws of the displacement distributions, Song proposed an exploration and preferential return model [58], Han gave an alternative explanation based on the hierarchical structure of the traffic systems [59, 60], and Zhao suggested that the cause is due to the mixture of different transportation modes [61]. On the other hand, studies have shown that the scaling characteristics no longer hold at the individual level [62]. And even at the population level, certain movements (such as taxi trips [63] and driving [64]) only provide exponentially distributed jump lengths. The inconsistency here reveals the fact that human mobility patterns can hardly be captured by simple generative mechanisms.

The above analysis focuses on human movements in real space. Sociophysicists also looked into the movement patterns in mental/cognitive spaces [65]. Rhodes [66] studied human's lag time when recalling information, and found that it follows a power-law distribution. Considering the lag time as the traveling distance in mental space, Rhodes concludes that human memory retrieval is similar to Lévy flights. Another study by Radicchi [67] focused on an online auction game called "Lowest Unique Bid," in which players try to find the unmatched lowest price for a product. The researchers investigated how players adjust their proposed prices (bids), and found that the change of the prices follow a power-law distribution, which suggests that the mental searching process of the "correct" price can be described by a Lévy flight. Radicchi further proposed an evolutionary model for explaining the price searching strategies [68].

2.3 Heavy-tailed distributions and proper statistical methods for distribution analysis

I have repeatedly mentioned power-law distributions in the above sections, which is one of the key findings in recent human dynamics research. However, one can easily confuse a power-law distribution with other types of heavy-tailed distributions [69]. Heavy-tailed distributions are those distributions which have tails decaying slower than exponential functions $e^{-\mu x}$ [70]. A power-law distribution $P(x) \sim x^{-\alpha}$ is a typical example, which has been seen in many physical, biological and social systems, such as the wealth distribution in a society [71]. One characteristic of a power-law distribution is scale invariance, therefore it is also referred to as a scaling law. In practical, empirical data usually comes with an upper boundary, and due to the finite-size effect, the power-law distribution sometimes is truncated with an exponential tail. Although a power-law distribution with exponential cutoff $P(x) \sim x^{-\alpha}e^{-\mu x}$ is not strictly a heavy-tailed distribution according to the mathematical definition, it still presents the scaling property within a certain range and the data following that distribution still present a large inequality. For that reason, we often keep using the term "heavy-tailed" for that distribution. Another example of heavy-tailed distribution is the log-normal distribution [72], in which the logarithm of the data follows a normal distribution. It is widely found in biological, economic, and physical systems, such as the rainfall size distribution [73]. Due to its connection with the normal distribution, a log-normal distribution can be generated with a multiplicative process [74]. Some other examples of heavy-tailed distributions include the Weibull distribution, stretched exponential distribution, t-distribution, etc.

A commonly used method for identifying a power-law distribution and the corresponding parameters is through a graphical approach. The most naive graphical method is to draw the probability distribution curve under a double logarithmic scale, and a power-law distribution

will appear as a straight line. However, empirical data usually are affected by noise, which leads to noisy distribution curves, from which it is hard to identify lines. To remove the noise, one can adopt logarithmic binning techniques, in which the points on the curves are averaged within equal logarithmic intervals [75]. Instead of directly drawing the probability distribution curves, an improved method relies on using the complementary cumulative distribution function (CCDF) curves $\bar{F}(x) = P(X > x)$, which vastly reduces the influence of noise [76]. Yet it is not perfect, as a graphical approach always relies on visual judgment, which is inaccurate [77].

To address this, Clauset [69] proposed a standard procedure for performing the distribution analysis for heavy-tailed distributed data. The first step is to estimate parameters for candidate distributions using maximum likelihood estimation (MLE)

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{x}) \ , \tag{2.1}$$

where $\hat{\theta}$ are the estimated parameters, $\mathcal{L}(\theta|\mathbf{x})$ is the likelihood of the parameters θ given observed data $\mathbf{x}$, and $\arg\max_{y} G(y)$ represents finding the y that maximizes the value of $G(y)$. The likelihood is a function of the parameters θ of a distribution. When assuming the observed data are independent, the likelihood can be calculated as the probability of the observed data under that distribution which uses the parameters θ

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{n} P_{\theta}\left(x_i\right) \ . \tag{2.2}$$

Once we obtain the estimated parameter, we calculate the Akaike information criterion (AIC) of the distribution,

$$\mathrm{AIC} = 2n_p - 2\ln\hat{\mathcal{L}} \ , \tag{2.3}$$

where n_p is the number of parameters in that distribution. AIC is widely used for model

comparison. Among the candidate distribution models, the one which provides the lowest AIC is preferred. AIC rewards good fittings to the observed data (with $-\ln\hat{\mathcal{L}}$), at the same time it penalizes the complexity of the distribution (with n_p). Then, we can calculate the Akaike weights for each of the distributions

$$w_i = \frac{\exp\left((\mathrm{AIC}_{\min} - \mathrm{AIC}_i)/2\right)}{\sum_j \exp\left((\mathrm{AIC}_{\min} - \mathrm{AIC}_j)/2\right)} \ . \tag{2.4}$$

Obtained Akaike weight is the relative probability that the data follows that distribution. The most likely distribution is the one with the greatest Akaike weight. The last issue needed to be addressed is the determination of the start of the "tails." The suggested method is to choose the value that minimizes the Kolmogorov–Smirnov (K-S) distance between the CCDFs of the fitted distribution and the empirical distribution [69]. I will come back to these procedures in Ch. 3.

Chapter 3

Foraging Patterns in Online Searches

The contents in this chapter were copied with permission from our publication [78]

X. Wang and M. Pleimling, Physical Review E 95, 032145 (2017). APS-Copyrighted.

Under Dr. Michel Pleimling's supervision, I contributed all the contents in this chapter.

3.1 Introduction

An increasingly large part of our day is devoted to online activities. It is therefore not surprising that in recent years online mobility patterns have emerged as a new interdisciplinary research area. Much attention has been given to scaling and non-Markovian features of web browsing [79, 80, 81, 82, 83, 84], to the features of mobility in online games [85, 86] as well as to emerging scaling properties in e-commerce [87].

Each day, tens of billions of clicks are generated on search engines. Understanding human online search click-through behavior can therefore be of central importance to improve ranking algorithms, rearrange page layout for search engines, and reduce advertisement spending for enterprises. Click-through data, which are extracted from the click logs of search engines, contain information on the links clicked by a user as a result of a query submitted to a search engine. These data have been exploited in a variety of studies that aimed at optimizing web searches and at improving retrieval quality [88, 89, 90, 91].

Our daily experience with web searches shows that a fully deterministic search strategy usually does not optimize the search outcome. Instead, there is often some degree of randomness involved in choosing the links to click on among those provided by a search engine. It is therefore tempting to investigate web searches from the point of view of random search strategies.

Studies of search strategies [92] in animal foraging [57, 93, 94, 95, 96, 97] hint at intriguing connections between movement patterns and availability of prey. Roughly speaking, one can distinguish between two different cases. If prey are abundant, then the predator tends to perform a random walk and only explores a rather restricted territory. On the other hand, if prey are scarce, evidence has been found that the pattern changes to a Lévy walk (or, alternatively, to an intermittent search process that includes Lévy movements [98, 99]) that

allows one to cover much larger areas or volumes and to optimize search efficiency when resources are sparsely distributed. It has been claimed that Lévy movement patterns also show up in human foraging [55, 100] as well as in the migration of bacteria [101, 102] and T cells [103]. Of course, the simple relationship between displacement pattern and availability of prey only prevails on large scales, and a much more complex and subtle picture emerges when going beyond such a coarse-grained description [104, 105, 106, 107].

Analyzing extensive click-through data sets from different search engines collected in different years, we consider in this chapter online searches as foraging processes on a straight semi-infinite line and find a transition in the search patterns from a behavior that includes long-range relocations to a purely local Brownian-type motion with increasing efficiency of the search engines. A more detailed analysis reveals a behavior that is more complex than simple Lévy flights or Brownian motions.

In the next Section we describe the click-through data as trajectories on a semi-infinite line. Section 3.3 analyzes these data as foraging processes on the semi-infinite line through the study of numerous quantities, ranging from probability distributions and complementary cumulative distribution functions of displacement and waiting time to mean-squared displacements and entropies. Our analysis indicates that the character of online searches changes with increasing efficiency of search engines, shifting from processes that include power-law distributions to a Brownian-type behavior.

3.2 Click-through data sets

Our study of human online search patterns is based on three click-through data sets collected by different commercial search engine providers. The sets Sogou-08 and Sogou-11 were collected on the Chinese search engine Sogou in 2008 and 2011 respectively, whereas the set

Yahoo-10 was collected on Yahoo in 2010. For a detailed description of the data sets and of the data preparation we refer the reader to Appendix A.

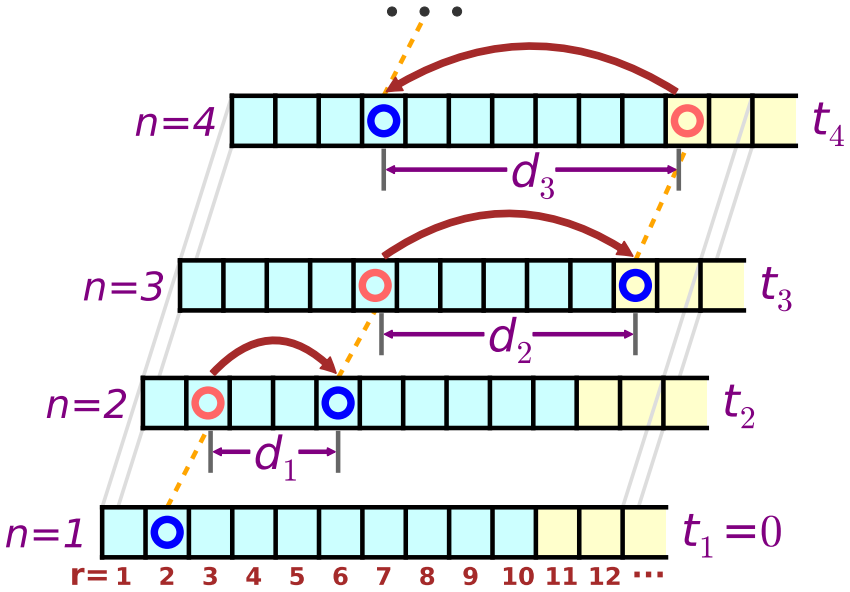

Figure 3.1: Illustration of the terms used in this study: clicking order n, rank r, time of nth click t_n, and step-length d_n between successive clicks on results provided by the search engine.

As explained in Appendix A and illustrated in Fig. 3.1, for each query submitted to one of the search engines we assign to every click on a search result a pair of "space"-"time" coordinates where the "time" is the time in seconds passed since the first click (i.e., for the first click, $t_1 = 0$), whereas the "space" coordinate is the rank of the search result, i.e., its position when treating all search results as points on a semi-infinite line, where the top result is assigned the rank $r = 1$. The nth click is therefore characterized by the pair (t_n, r_n). Subsequent clicks for a given search then correspond to subsequent steps along the semi-infinite line where the rank difference $\Delta r_n = |r_{n+1} - r_n|$ is the step-length $d_n = \Delta r_n$. The sign of $r_{n+1} - r_n$ provides us with the direction of the steps. As we will see later, the data show a strong bias in the forward direction. In a similar way we define as waiting time τ_n

the time interval between two successive clicks [1]: $\tau_n = t_{n+1} - t_n$.

Usually a query ends when the user finds the relevant information. In the language of foraging the relevant information is therefore the resource. We know from our own experience that users occasionally terminate a search early once they are convinced that the search term is unlikely to yield the expected result. The limitations of the data bases used in this study do not allow to identify these instances. We therefore do not try to distinguish between these cases and treat all queries in the same way, with the resource being located at the site $r = r_f$ that corresponds to the rank of the last search result clicked by the user. Of interest is also the number of clicks (steps) needed to reach the resource. As we discuss below, the resource location r_f and the clicking number N_c provide simple ways of evaluating the efficiency of a search engine. Both r_f and N_c would presumably change slightly if we would be able to identify those searches that resulted in the user finding the information they were looking for.

In the following we focus on probability distributions and on complementary cumulative distribution functions (CCDF(x) is the probability $P(X > x)$ that the random variable X has a value larger than x) in order to analyze the motion patterns emerging from online searches. We take a population level approach and base our analysis on population-averaged distributions. As we are dealing with millions of queries for every search engine, see Table 3.1, we expect that distribution functions provide a very reliable characterization of the foraging patterns in online searches.

[1] The waiting time consists of two parts: the time for viewing the previously clicked result and the time needed to select the next link on the search result pages. Our data do not allow us to distinguish between these two contributions.

3.3 Model selection and parameter estimation

In the following we model the distributions of various quantities in the large-value limit with a variety of models and select the best model using the Akaike information criterion (AIC). As the quantities derived from the click-through data sets take on positive integers only, we consider only discrete versions of the models.

As a power-law model we use the discrete power-law (DPL) distribution

$$P(k) = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})} \sim k^{-\alpha}, \quad k \geq k_{\min},\ \alpha > 1, \tag{3.1}$$

where the normalizing factor

$$\zeta(\alpha, k_{\min}) = \sum_{m=k_{\min}}^{\infty} m^{-\alpha} \tag{3.2}$$

is the incomplete ζ-function [108]. For the exponential model we use the "shifted" geometric (SG) distribution

$$\begin{aligned} P(k) &= p(1-p)^{k-k_{\min}}, \quad k \geq k_{\min},\ 1 \geq p > 0, & (3.3)\\ &= \left(1 - e^{-\lambda}\right) e^{-\lambda(k-k_{\min})} \sim e^{-\lambda k}, \quad \lambda > 0, & (3.4) \end{aligned}$$

where $\lambda = -\ln(1-p)$.

We also included in the model selection the power law with exponential cutoff (PEC) model:

$$P(k) = \frac{1}{Li_{\alpha}(e^{-\lambda}) - \sum_{i=1}^{k_{\min-1}} i^{-\alpha} e^{-\lambda i}} k^{-\alpha} e^{-\lambda k} \sim k^{-\alpha} e^{-\lambda k}, \quad k \geq k_{\min},\ \lambda > 0,\ \alpha > 0, \tag{3.5}$$

where $Li_{\alpha}(z)$ is the polylogarithm function, the discrete log-normal model (DLN) [109]:

$$P(k) = \frac{\Phi\left(\frac{\ln(k+1)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(k)-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\ln(k_{\min})-\mu}{\sigma}\right)}, \quad k \geq k_{\min},\ \sigma > 0, \tag{3.6}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, the Yule-Simon (YS) distribution [69, 110]

$$P(k) = (\alpha-1)\frac{\Gamma\left(k_{\min}+\alpha+1\right)}{\Gamma\left(k_{\min}\right)}\frac{\Gamma(k)}{\Gamma(k+\alpha)}, \quad k \geq k_{\min},\ \alpha > 1, \tag{3.7}$$

which for k sufficiently large yields $P(k) \sim k^{-\alpha}$, and the conditional Poisson (CP) distribution [69]

$$P(k) = \left[e^{\mu} - \sum_{m=0}^{k_{\min}-1}\frac{\mu^m}{m!}\right]^{-1}\frac{\mu^k}{k!}, \quad k \geq k_{\min},\ \mu > 0. \tag{3.8}$$

Finally, we also considered a pairwise power-law (PPL) distribution, which consists of two power-law regions that are connected at $k = k_{\text{trans}}$:

$$P(k) = \begin{cases} C\ k^{-\alpha}, & k_{\min} \leq k < \lceil k_{\text{trans}} \rceil \\ C k_{\text{trans}}^{\beta-\alpha}\ k^{-\beta}, & \lceil k_{\text{trans}} \rceil \leq k \end{cases}, \quad \alpha, \beta > 1,\ k_{\text{trans}} > k_{\min}, \tag{3.9}$$

with the normalization factor

$$C = \left(\zeta(\alpha, k_{\min}) - \zeta(\alpha, \lceil k_{\text{trans}} \rceil) + k_{\text{trans}}^{\beta-\alpha}\zeta(\beta, \lceil k_{\text{trans}} \rceil)\right)^{-1}. \tag{3.10}$$

Due to the ceiling function $\lceil x \rceil$ this distribution does not strictly sum up to 1. Still, as we will see in the following, it does provide in many instances a good fit to our data.

Inspection of these distributions reveals the presence of a minimal value $k_{\min}$ that determines the start of the 'tail' used for the modeling. In many cases $k_{\min}$ can be determined as the value that minimizes the Kolmogorov-Smirnov statistics between the empirical distributions

and the fitted distributions [69].

Due to the large size of our data we directly use the formula $\mathrm{AIC} = -2\ln\hat{\mathcal{L}} + 2n_p$ for the Akaike information criterion. Here n_p is number of parameters in each distribution model and $\hat{\mathcal{L}}$ is the maximum likelihood of the model (see Appendix B). The Akaike weight w_i [111] for each model is

$$w_i = \frac{\exp\left((\mathrm{AIC}_{\min} - \mathrm{AIC}_i)/2\right)}{\sum_j \exp\left((\mathrm{AIC}_{\min} - \mathrm{AIC}_j)/2\right)} , \tag{3.11}$$

with the model with the largest Akaike weight being the most likely model.

Although we show in the following plots of the probability distribution functions (with logarithmic binning) and of the complementary cumulative distribution functions as illustration, we do not use them directly for parameter estimation. Instead we estimate parameters from the distribution models with the maximum likelihood method. The maximum likelihood estimators (MLE) for the different parameters are summarized in Appendix B.

3.4 Results

3.4.1 Search engine efficiency

Before delving into a detailed analysis of foraging related quantities like step-lengths and waiting times, we will first briefly characterize in a straightforward way the efficiency of the different search engines (or of the same search engine in different years) through the clicking number, i.e., the number of clicks needed before the user ends the query.

From the point of view of a user submitting a query to a search engine, what matters is the number of links they have to click on before retrieving the needed information. The clicking number N_c should therefore directly reflect the efficiency of a search engine to provide the

user with the relevant information.

Table 3.1: Comparison of the search engine efficiency based on the clicking number N_c. The third line provides the total number of queries analyzed in our study. As a default the search engines provide 10 links per page.

	Sogou-08	Sogou-11	Yahoo-10
year	2008	2011	2010
search engine provider	Sogou	Sogou	Yahoo
number of valid queries (in millions)	14.6	30.4	53.8
$\langle N_c \rangle$	1.741	1.433	1.130
$\max(N_c)$	299	23	19
$P(N_c > 10)$	0.626%	0.000033%	0.00655%

Table 3.1 summarizes some of our findings for the clicking number. Inspection of the table reveals immediately striking differences when comparing Sogou-08 with the other, newer, sets. Both the average clicking number $\langle N_c \rangle$ and the largest clicking number found in the millions of queries forming the different sets point to an impressive increase of the efficiency when going from Sogou-08 to the newer sets. The most striking difference is provided by the probability $P(N_c > 10)$ that more than 10 clicks are needed for accessing the relevant information. Whereas for Sogou-08 around 0.63% of the queries result in more than 10 clicks on links provided by the search engine, only a very small number of searches in Sogou-11 and Yahoo-10 result in the inspection of more than 10 links.

The distribution CCDF(N_c) shown in Fig. 3.2 reveals the dramatic differences between the efficiency (as measured by N_c) of the different search engines. Note that the error of the obtained probability values of heavy-tailed distributions is not normally distributed. Although there exist several statistical methods for estimating the confidence intervals of those probability values, including data tilting [112] and bootstrap [113], our adoption of logarithmic binning techniques for smoothing the distribution curves further complicated the error estimation. To simplify our analysis, we will omit the error bars from the distribution plots.

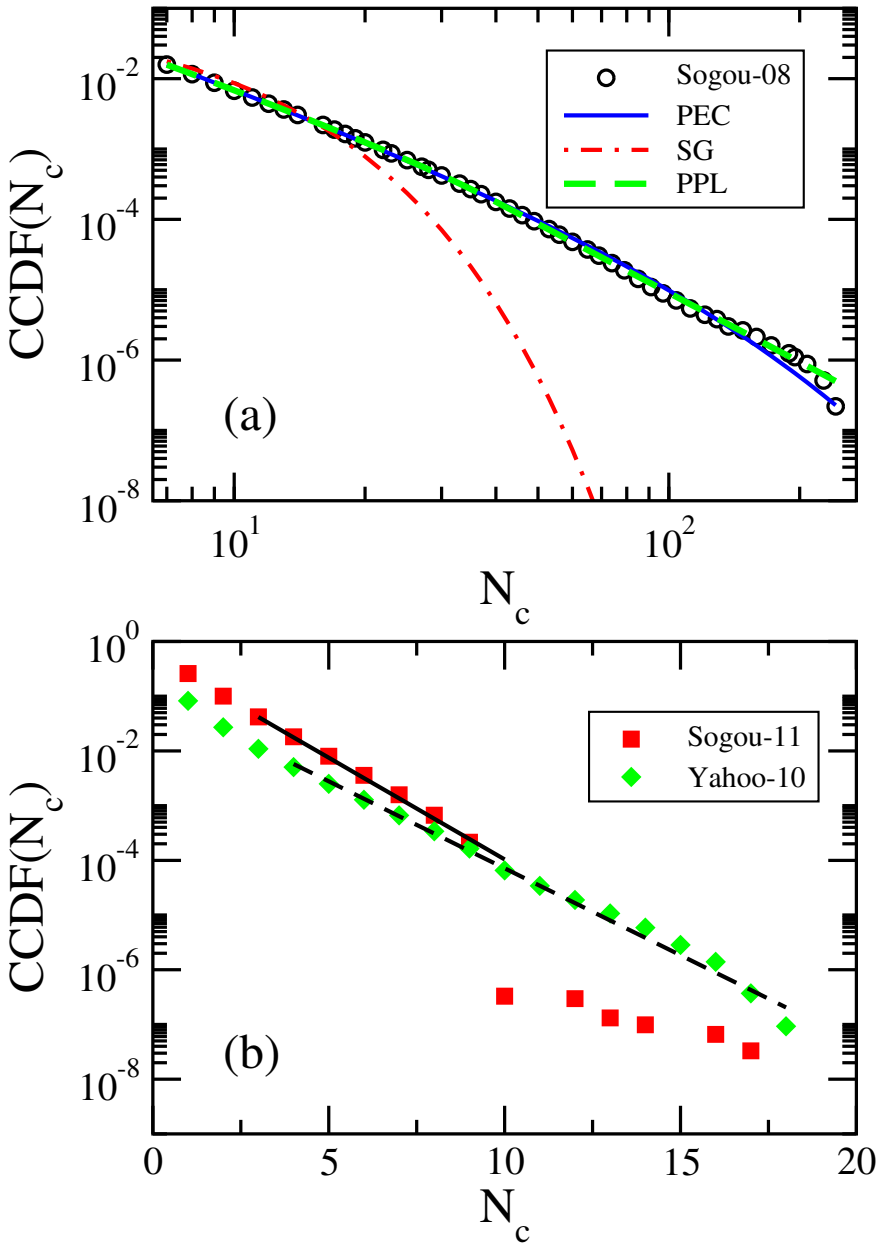

Figure 3.2: Complementary cumulative distribution function of the clicking number N_c obtained from the different sets. (a) Log-log plot of CCDF(N_c) for Sogou-08 and comparison with the distributions obtained from three different models, see Table 3.2: power law with exponential cutoff (PEC), exponential model (SG), and pairwise power law (PPL). (b) Linear-log plot of CCDF(N_c) for Sogou-11 and Yahoo-10. For Sogou-11 the probability distribution displays a dramatic drop for $N_c \geq 10$. The lines indicate an exponential decay.

Table 3.2: Model selection using AIC and maximum likelihood estimators for the parameters in the most likely models of N_c and r_f. In the table $\ln \hat{\mathcal{L}}$ and AIC are rounded to integers. See the main text for the meaning of the acronyms.

	Set	model	$\ln \hat{\mathcal{L}}$	AIC	w_i	most likely model	$k_{\min}$	MLE
N_c	Sogou-08	YS	-722852	1445707	0.000	PPL	7	$\hat{\alpha} = 3.488$ $\hat{\beta} = 4.280$ $\hat{k}_{\text{trans}} = 39.234$
		DPL	-722912	1445825	0.000			
		SG	-746099	1492200	0.000			
		CP	-1000863	2001728	0.000			
		DLN	-723126	1446255	0.000			
		PEC	-722760	1445524	0.000			
		PPL	-722739	1445484	1.000			
	Sogou-11	YS	-3637855	7275713	0.000	SG	3	$\hat{\lambda} = 0.855$
		DPL	-3671768	7343538	0.000			
		SG	-3603146	7206293	1.000			
		CP	-3659310	7318621	0.000			
	Yahoo-10	YS	-789652	1579305	0.000	SG	4	$\hat{\lambda} = 0.732$
		DPL	-794101	1588203	0.000			
		SG	-786042	1572086	1.000			
		CP	-803599	1607199	0.000			
r_f	Sogou-08	YS	-1657346	3314695	0.000	PPL	16	$\hat{\alpha} = 2.108$ $\hat{\beta} = 2.948$ $\hat{k}_{\text{trans}} = 139.580$
		DPL	-1657674	3315350	0.000			
		SG	-1734486	3468975	0.000			
		CP	-9745388	19490777	0.000			
		DPL	-1655263	3310530	0.000			
		PEC	-1654553	3309109	0.000			
		PPL	-1654359	3308723	1.000			

As reported in Table 3.2, we find using AIC that the probability distribution for Sogou-08 follows a pairwise power-law distribution with the maximum likelihood estimations $\hat{\alpha} = 3.448, \hat{\beta} = 4.280$, whereas the transition happens at $\hat{k}_{\text{trans}} = 39.234$. For Sogou-11 and Yahoo-10, however, the distributions rapidly show exponential decay. This exponential decay stops at $N_c = 10$ for Sogou-11, indicating that with a very few exceptions all queries are finished within 10 clicks. This discontinuity does not happen in Yahoo-10 which instead shows a smooth behavior. The jump in Sogou-11 reveals that with a few exceptions users found within the links provided on the very first page with results the information they were looking for.

One also expects from an efficient search engine that the relevant result is included in the very

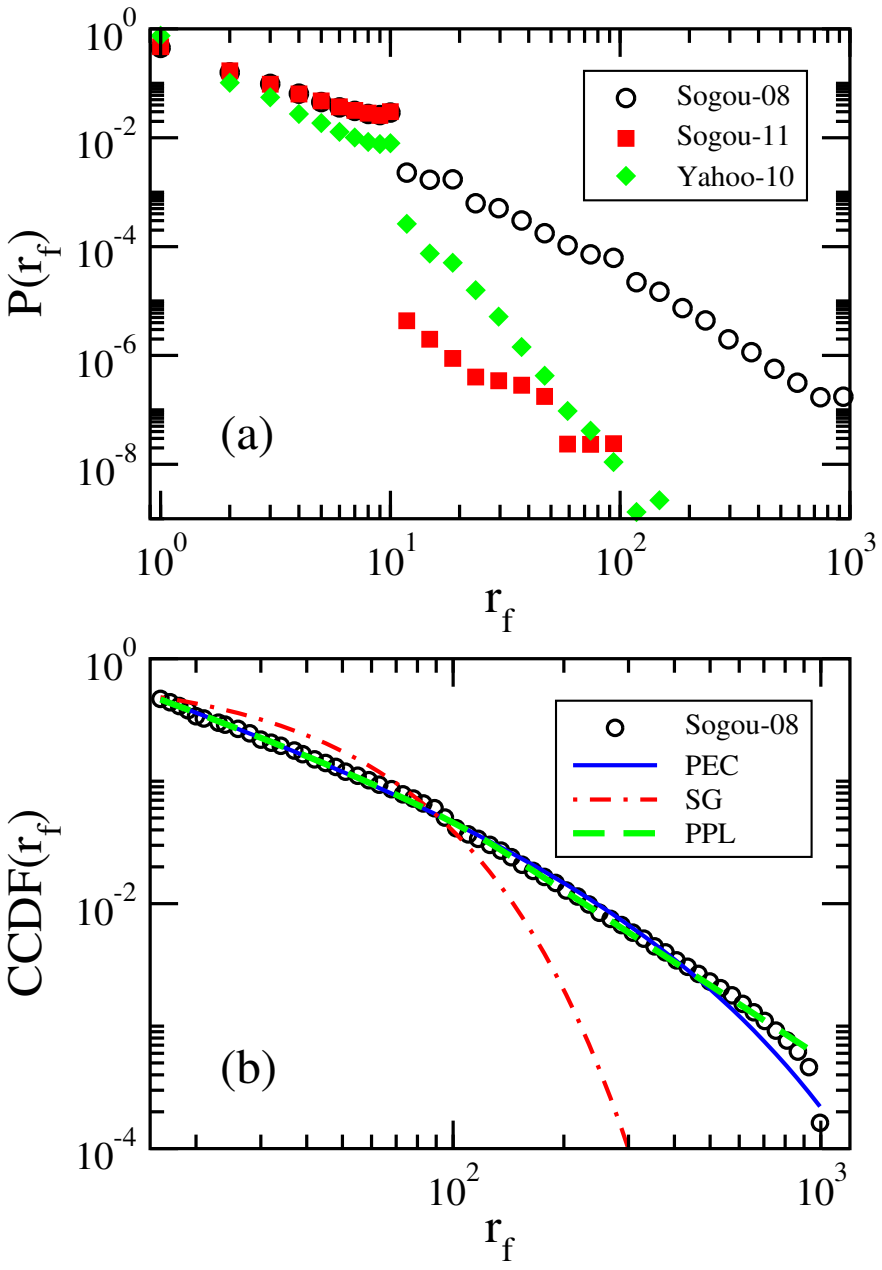

Figure 3.3: (a) Probability distribution $P(r_f)$ where r_f is the rank of the link that is clicked as last in a given search. r_f therefore corresponds to the position of the resource on the semi-infinite line. Large differences can be observed between the tails of Sogou-08 and Sogou-11. (b) Complementary cumulative distribution function CCDF(r_f) for Sogou-08 as well as the corresponding distributions obtained from three different models, see Table 3.2: power law with exponential cutoff (PEC), exponential model (SG), and pairwise power law (PPL).

first suggested links. Fig. 3.3 compares for our three data sets the probability distribution $P(r_f)$ where r_f is the rank of the final click (i.e., the position of the resource on the semi-infinite line). We first note that for Yahoo-10 75% of the searches end with a click on the very first link on the search result page provided by the search engine. For Sogou-11 this number is 47%, very similar to the 45% of searches that end with a click on the very first link for Sogou-08. For Sogou-11 and Yahoo-10 almost all resources (99.997% for Sogou-11 and 99.864% for Yahoo-10) are located on the first page with $1 \leq r_f \leq 10$. For these two cases one also observes large changes between $P(r_f = 10)$ and $P(r_f = 11)$, illustrating the fact that only for a negligible number of searches the resource is found for $r_f > 10$. This is different for Sogou-08 where $P(r_f > 10) = 3.661\%$ and $P(r_f > 100) = 0.224\%$, resulting in a much smoother shape with a pairwise power-law tail. The probability distribution $P(r_f)$ indicates that for Sogou-11 and Yahoo-10 only a local exploration on the first page is needed in order to reach the resource. As we will see in the following, this difference yields different space-time patterns during the searches.

3.4.2 Step-lengths and waiting times

In the following analysis we view as a random walk on the semi-infinite line the exploration by the user of the links provided by the search engine. Focusing on the step-length and waiting time distributions, we will see that the difference in efficiency noticed in the previous subsection yields different space-time processes.

A way to distinguish between Brownian-type motion and Lévy movement is to investigate the probability distribution $P(d)$ of the step-length d, which should display a heavy tail in the form of a power law

$$P(d) \sim d^{-\alpha} \tag{3.12}$$

Table 3.3: Model selection using AIC and maximum likelihood estimators for the parameters in the most likely model for the step-lengths. See the main text for the meaning of the acronyms.

	Set	model	$\ln \hat{\mathcal{L}}$	AIC	w_i	most likely model	$k_{\min}$	MLE
d	Sogou-08	YS	-2090805	4181612	0.000	PPL	10	$\hat{\alpha} = 2.169$ $\hat{\beta} = 3.417$ $\hat{k}_{\text{trans}} = 91$
		DPL	-2091568	4183137	0.000			
		SG	-2187409	4374821	0.000			
		CP	-7645269	15290541	0.000			
		DLN	-2088146	4176295	0.000			
		PEC	-2087035	4174075	0.000			
		PPL	-2085243	4170491	1.000			
	Sogou-11	YS	-22684073	45368149	0.000	SG	1	$\lambda = 0.544$
		DPL	-23284589	46569181	0.000			
		SG	-21326460	42652922	1.000			
		CP	-22894455	45788912	0.000			
	Yahoo-10	YS	-12128059	24256120	0.000	SG	1	$\lambda = 0.540$
		DPL	-12450282	24900567	0.000			
		SG	-11415241	22830484	1.000			
		CP	-12326037	24652076	0.000			
d_{in}	Sogou-08	YS	-16506000	33012001	0.000	SG	1	$\lambda = 0.550$
		DPL	-16924868	33849738	0.000			
		SG	-15652590	31305183	1.000			
		CP	-16985431	33970864	0.000			
d_{out}	Sogou-08	YS	-1232283	2464569	0.000	PPL	1	$\hat{\alpha} = 2.353$ $\hat{\beta} = 3.226$ $\hat{k}_{\text{trans}} = 9.000$
		DPL	-1225123	2450249	0.000			
		SG	-1435452	2870906	0.000			
		CP	-1943098	3886197	0.000			
		DLN	-1235756	2471516	0.000			
		PEC	-1223417	2446838	0.000			
		PPL	-1221825	2443655	1.000			

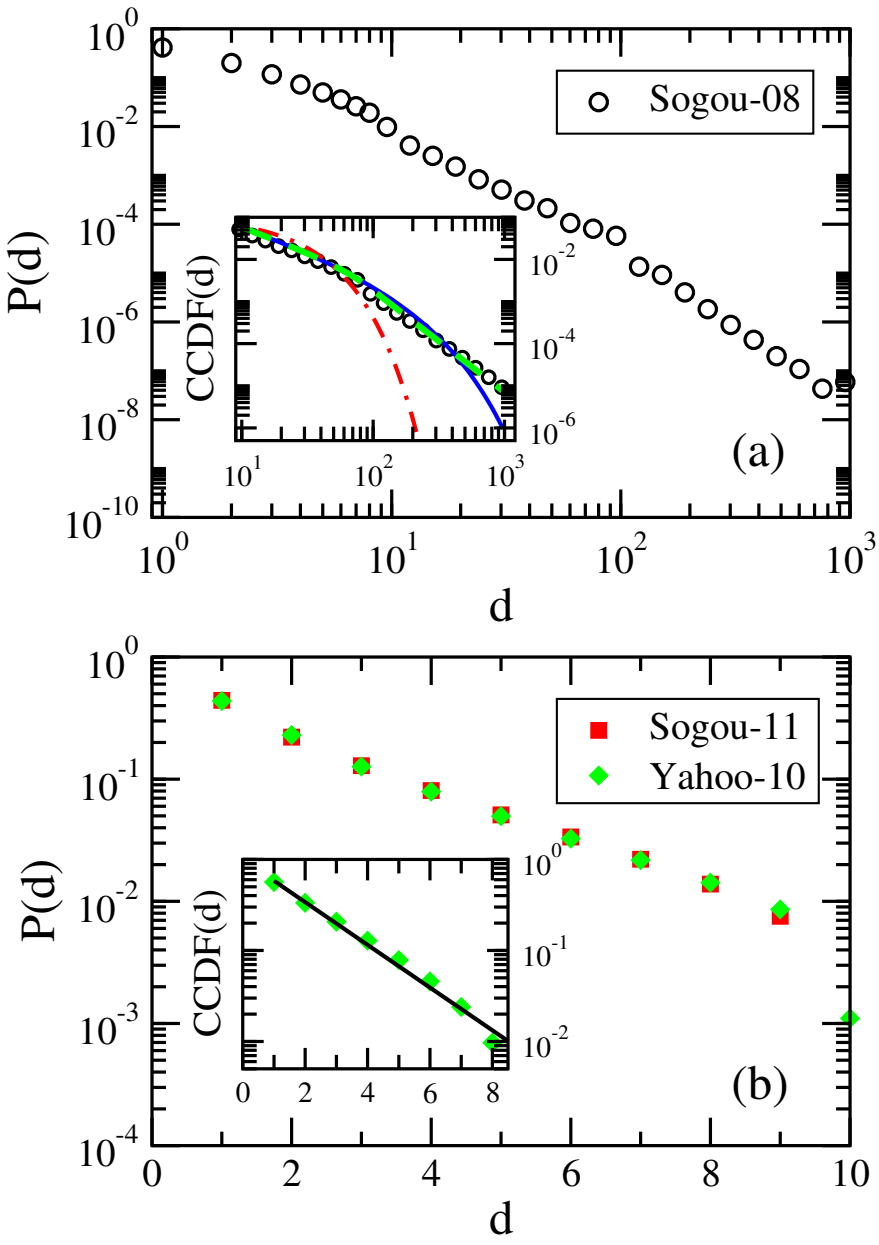

Figure 3.4: Step-length probability distribution functions and (in the inset) complementary cumulative distribution functions. (a) For Sogou-08 the distributions display a pairwise power-law tail, whereas (b) for Sogou-11 and Yahoo-10 they decay exponentially; see Table 3.3. The lines in the inset of panel (a) show the following fitted models: PEC (solid blue line), SG (dot-dashed red line), and PPL (dashed green line). The solid black line in the inset of panel (b) represents an exponentially decaying function.

with $1 < \alpha < 3$ for superdiffusive Lévy flights.

Fig. 3.4 shows the probability distributions for the step-length derived from the data at our disposal. Focusing first on Sogou-08, our model selection procedure shows that the pairwise power-law model provides a good absolute fit, see Fig. 3.4a as well as Fig. 3.5b. From the maximum likelihood estimation we obtain that the transition between the two power laws happens at $d_{\text{trans}} = 91$. The exponent in the first power-law region is given by $\alpha = 2.169$, a value that is within the Lévy-flight range $1 < \alpha < 3$, i.e., between $10 < d < d_{\text{trans}}$ we have a long-range search pattern that is consistent with the power-law distribution of a Lévy flight. This behavior does not persist for the largest values of d. Instead, the exponent in the second power-law region is $\beta = 3.417$, which is outside the Lévy-flight range. This value guarantees a finite variance for step-lengths and suggests that the very long-range movements have the properties of normal diffusion. We believe that this change in behavior around $d_{\text{trans}} = 91$ is due to the layout of the search engine result pages, since the search engines used in our study list 10 pages at the bottom of a result page. For Sogou-11 and Yahoo-10, 99.997% and 99.890% of steps have a length $d < 10$. The distributions for $d < 10$ are exponential which points to the overwhelming predominance of local searches where only a small "area" is explored.

Interestingly, even for Sogou-08 an exponential decay is hidden in the distribution shown in Fig. 3.4a. Separating the jumps within a page from those between pages, we discover in Fig. 3.5 a more complex behavior. Restricting ourselves to jumps within a given page, where we denote by d_{in} the corresponding step-length, we find for Sogou-08 an exponential decay of the step-length distribution. The difference between Sogou-08 and the other sets therefore mainly results from searches where the resource is not readily found, yielding jumps between different pages with a pairwise power-law probability distribution for the out-of-page step-length $d \geq 10$. As shown in Fig. 3.5b, the page difference d_{out} of out-of-page jumps still

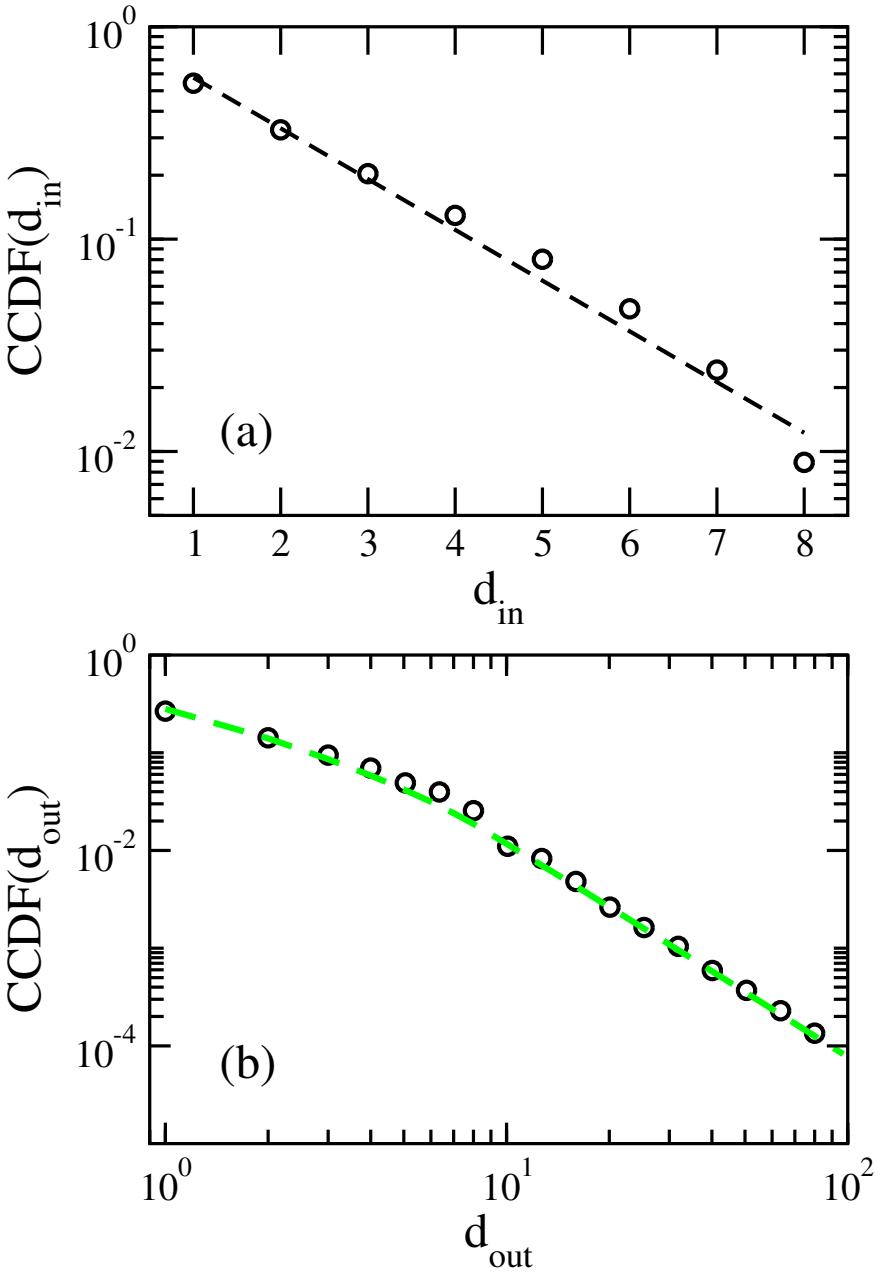

Figure 3.5: (a) Step-length distribution for jumps within a page for Sogou-08. An exponential decay is observed for local searches, similar to the distributions encountered for Sogou-11 and Yahoo-10. (b) Distribution of the page difference for jumps between pages for Sogou-08. From the model selection follows that these data are described by a pairwise power-law distribution, indicated by the green dashed line (see also Table 3.3).

yields a pairwise power-law distribution.

The fact that Sogou-08 yields a switch between a local (i.e., on one page with search results) Brownian search and a relocation phase that is power-law distributed is very reminiscent of an intermittent search process that includes Lévy strategies [98, 99]. Intermittent search processes have been proposed as search strategies in cases where the targets are hidden [114, 115, 116]. They are characterized by switches between two different phases: careful searches around one location, followed by rapid relocations to some other areas. The careful searches are usually described as Brownian searches whereas the relocations are often assumed to be either ballistic or Lévy distributed. The set-up of search engines queries has many obvious direct connections with an intermittent search process. This is especially true for Sogou-08 for which we observe relocations over large distances.

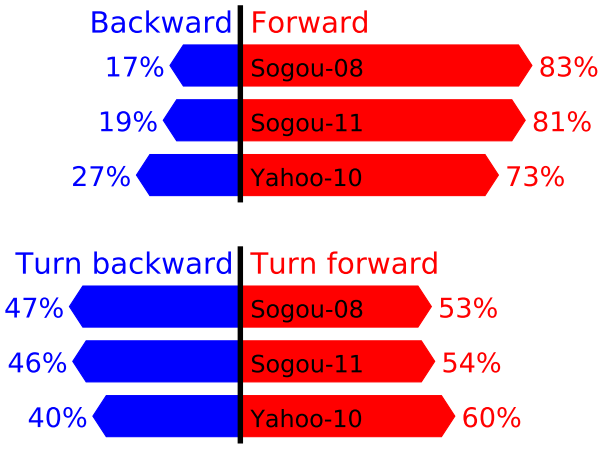

Figure 3.6: Upper part: fraction of jumps in the forward and backward directions. Lower part: fraction of turns that change the direction from forward to backward and of turns that change the direction from backward to forward.

For any of the processes usually discussed in the context of foraging, one generally assumes the movement to be unbiased, i.e., that jumps are happening with a direction independent probability distribution. As we know from our own experience and as shown in Fig. 3.6 for the different data sets, this is not the case in online searches, where users have a tendency to

start at the top of a page with research results and proceed to the bottom of the page (i.e., to move preferentially in one direction, see top of Fig. 3.6). While there is a clear directionality in how a user exploits search results, a much smaller bias is observed in the 'turning angles,' i.e., in the changes of direction from forward to backward and from backward to forward.

On the semi-infinite line, which provides the landscape for online foraging, moving forward usually means exploring new results, while moving backwards often means revisiting previously viewed results. The bias of foraging means that there is much more exploration of new results than revisitation of already viewed ones. Meanwhile, since most of the initial movements are forward when users start to view results from a search, the much weaker bias in turning angles indicates that a revisitation is usually followed by another exploration.

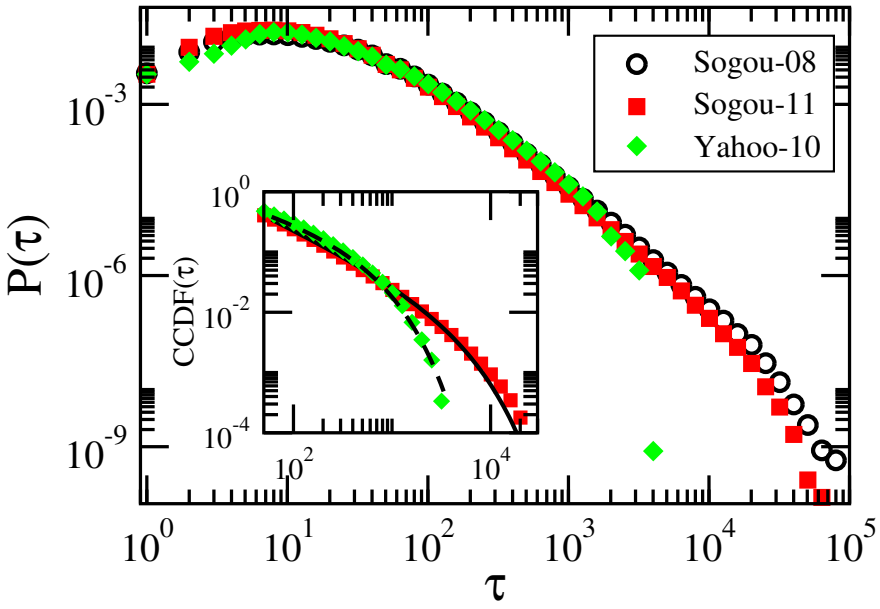

Figure 3.7: Waiting time distributions $P(\tau)$ for the different sets, with τ measured in seconds. For all three cases the distribution exhibits a power-law tail, with an exponent around 1.9, followed by an exponential cutoff. The different sets have different upper limits for the waiting times: one day for Sogou-08 and Sogou-11 and one hour for Yahoo-10. Inset: the corresponding complementary cumulative distribution functions and the fits (solid and dashed lines) obtained from power-law models with exponential cutoffs.

Other differences between online searches and models for foraging emerge when considering

the probability distribution $P(\tau)$ of the waiting time τ, i.e. the time elapsed between two consecutive clicks on links provided by the search engine. The time between two clicks on the links provided by the search engine is mainly the time spent by the user viewing a web site. The average time spent on selecting the next link on the search result pages is indeed small compared to the average time spent on a selected web site. Inspection of Fig. 3.7 reveals that $P(\tau)$ is well modeled by a power-law distribution with exponential cutoff, see inset. This is especially true for Yahoo-10 where our model selection indeed finds that a power-law distribution with exponential cutoff provides the best fit. For Sogou-08 and Sogou-11 AIC yields the log-normal distribution as the most probable one, but the log-normal distribution does not at all capture the behavior for large τ, which instead is well described by a power-law distribution with exponential cutoff, see the black lines in the inset of Fig. 3.7. This exponential cutoff in $P(\tau)$ is due to upper limits for the waiting time set by the session expiration time (one hour for Yahoo-10 and one day for Sogou-08 and Sogou-11) [2].

3.4.3 Mean-squared displacement

For Brownian motion and Lévy flights the mean-squared displacement shows a characteristic behavior, increasing linearly with the number of jumps for the first case, whereas for the second case a superdiffusive behavior is expected, with a power-law increase where the exponent is larger than one.

Regarding a query again as a motion along the semi-infinite line where the rank of a click r corresponds to the position whereas the clicking order n counts the number of jumps and

[2]Whereas for Sogou-08 and Sogou-11 the logs are divided into different days, yielding a cutoff of 86,400 seconds for the waiting time, for Yahoo-10 the longest allowed time for a search is one hour. These cutoffs are readily seen in Fig. 3.7.

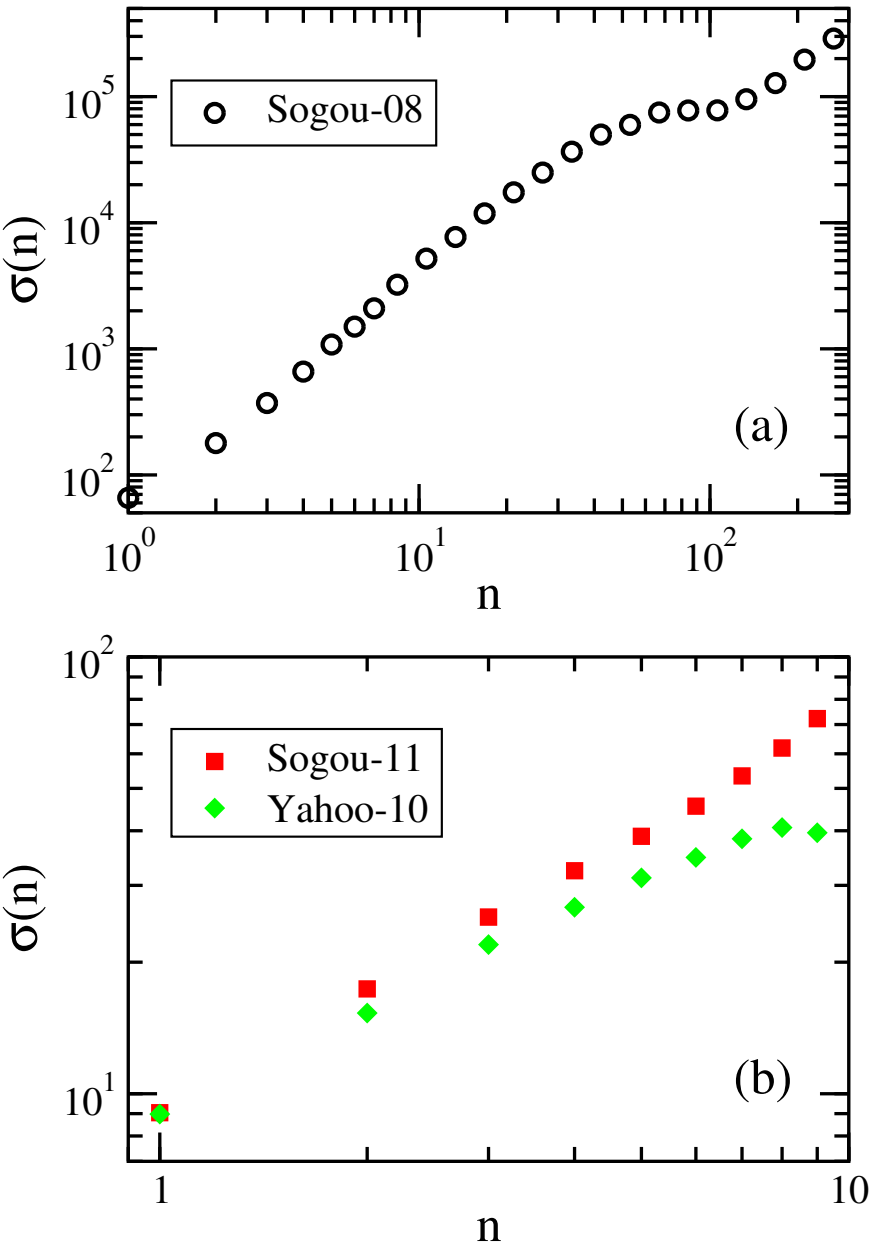

Figure 3.8: Mean-squared displacement $\sigma(n)$ as a function of the clicking order n. (a) A superdiffusive behavior is observed for Sogou-08. (b) For Sogou-11 and Yahoo-10 a slightly subdiffusive behavior is revealed by $\sigma(n)$.

therefore serves as a proxy for "time," we can calculate the mean-squared displacement as

$$\sigma(n) = \langle (r(n) - r(1))^2 \rangle \tag{3.13}$$

where $r(1)$ is the rank of the first clicked result. In (3.13) the average is over the different trajectories along the semi-infinite line.

Fig. 3.8 shows for the different sets the variation of the mean-squared displacement with increasing clicking order n. For Sogou-08, see Fig. 3.8a, we do find for $n \leq 30$ the expected superdiffusive behavior, $\sigma(n) \sim n^a$, with an exponent $a \approx 1.95$. For Sogou-11 and Yahoo-10 reliable data are only available for small values of n due to the fact that the resource is usually found after only a few clicks. For $n < 10$ we find for Sogou-11 an exponent $a = 0.92$, whereas for Yahoo-10 the value of the exponent is $a = 0.75$. These values, which indicate a slight subdiffusive behavior, are rather close to the value $a = 1$ of normal diffusion.

As we know from the click-through logs the time elapsed between any two consecutive clicks, we can also calculate the mean-squared displacement as a function of the real time measured since the very first click:

$$\sigma(t) = \langle (r(t) - r_0))^2 \rangle \tag{3.14}$$

where r_0 is the rank of the first click at time $t = 0$. We know from Fig. 3.7 that for all three sets the distributions of waiting times, which are composed by the times spent on selecting the next link on the search result pages and the times spent viewing the previously selected website, are rather complicated. This will of course impact the time dependence of $\sigma(t)$. Our result for Sogou-08 shown in Fig. 3.9a indicates that for that case the time-dependent mean-squared displacement varies in the time interval 100 $s < t < 2000$ s like a power law with an exponent close to 1.30. For Sogou-11 and Yahoo-10, however, $\sigma(t)$ roughly varies logarithmically with time in the interval 10 $s < t < 500$ s, see Fig. 3.9b.

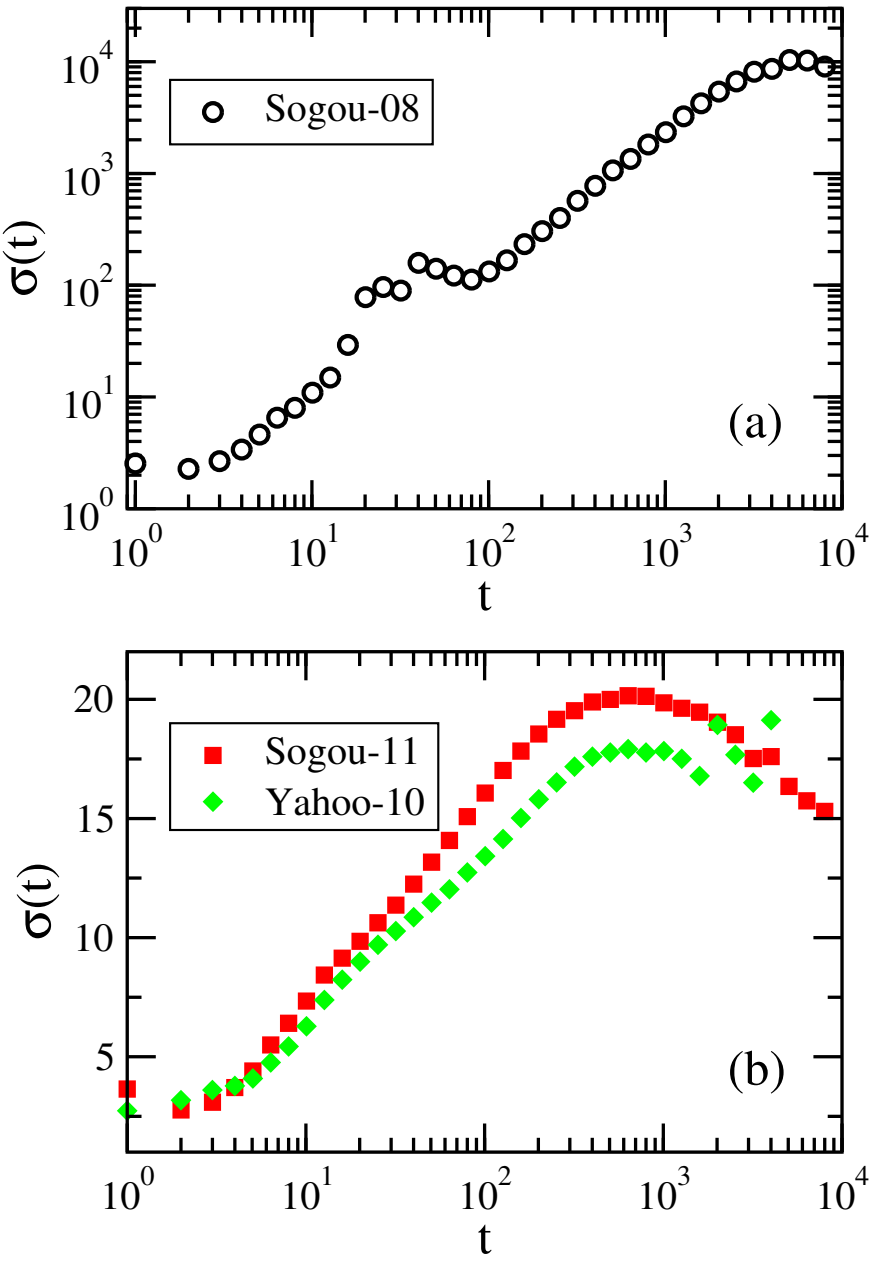

Figure 3.9: Mean-squared displacement $\sigma(t)$ as a function of time t (measured in seconds) elapsed since the very first click on a link provided by the search engine. Whereas for Sogou-08 a power law with an exponent close to 1.30 is observed in the time interval 100 $s < t <$ 2000 s, for Sogou-11 and Yahoo-10 $\sigma(t)$ is found to vary logarithmically with time for 10 s $< t <$ 500 s.

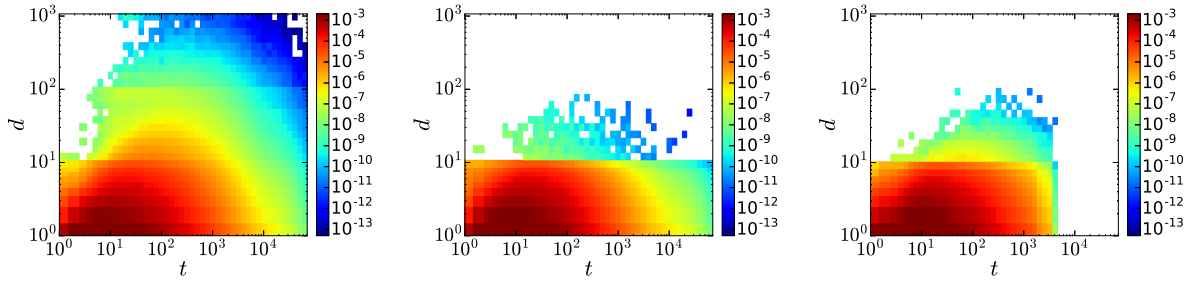

Figure 3.10: Joint probability $P(t, d)$ for Sogou-08 (left panel), Sogou-11 (middle panel), and Yahoo-10 (right panel). Note the sharp transition at $d = 10$ for Sogou-11 and Yahoo-10, due to the rather few jumps between different pages. t is the time measured in seconds since the first click on a link and d is the step-length between two consecutive clicks.

3.4.4 Entropy

The differences between the different sets can also be highlighted through the study of time-dependent entropies. We can for example start from the conditional probability $P(d|t) = P(t, d)/P(t)$, which is the probability that the step-length is d given that the jump takes place at time t. Here $P(t, d)$ is the joint probability of the time t elapsed since the first click and the step-length d, whereas $P(t)$ is the probability that the event happens at time t. This allows us to define the time-dependent entropy

$$S_d(t) = -\sum_{d=1}^{\infty} P(d|t) \ln P(d|t) \tag{3.15}$$

We could also start from a different conditional probability distribution, e.g., $P(d|n)$ where n is the clicking order,

$$S_d(n) = -\sum_{d=1}^{\infty} P(d|n) \ln P(d|n) \ . \tag{3.16}$$

Fig. 3.10 compares the joint probabilities $P(t, d)$ for the different sets. We note again that for Sogou-11 and Yahoo-10 almost all displacements are local with $d < 10$, which is reflected by the very low or even vanishing joint probability $P(t, d)$ for $d \geq 10$ for these cases. We also

note that for $d \leq 9$ the probabilities are very similar for the different sets, which indicates that the properties of local searches are rather set independent. The probability for Sogou-08 reveals the emergence and distribution of long-range relocations. Finally, for all cases $P(t, d)$ changes as a function of time, as expected for a search process that is taking place far from equilibrium.

The time dependence of the entropy $S_d(t)$ defined in Eq. (3.15) is shown in Fig. 3.11a. We first note that for all three sets the entropy shows a strong increase at early times. For Sogou-11 and Yahoo-10, characterized by mostly local searches on the first page with links, this increase rapidly weakens and S_d reaches a plateau for times $t > 50$ seconds. For Sogou-08, where the motion is formed by a combination of local searches and long relocation jumps, the entropy keeps increasing up to $t \approx 2000$ seconds before reaching a plateau. $S_d(t)$ is therefore another quantity that allows to easily distinguish between Brownian-like searches and searches that are characterized by power-law distributions.

Fig. 3.11b shows the entropy $S_d(n)$ defined in Eq. (3.16). For Sogou-08, $S_d(n)$, after an initial increase for small n, rapidly reaches a plateau for $n \geq 11$. The process is therefore stationary for $n \geq 11$. In contrast to this, for Sogou-11 and Yahoo-10 the entropy does not reach a well defined plateau, and the processes therefore do not reach a stationary state.

3.4.5 Correlations

Correlation coefficients allow us to gain additional insights into the relationships between different quantities (see Appendix C for the definitions of the correlation coefficients discussed in the following). This is of interest as Brownian motions and Lévy flights, which are commonly used to model human dynamics, assume that space and time are uncorrelated and that the random walks are memoryless. Studying correlation coefficients will allow us

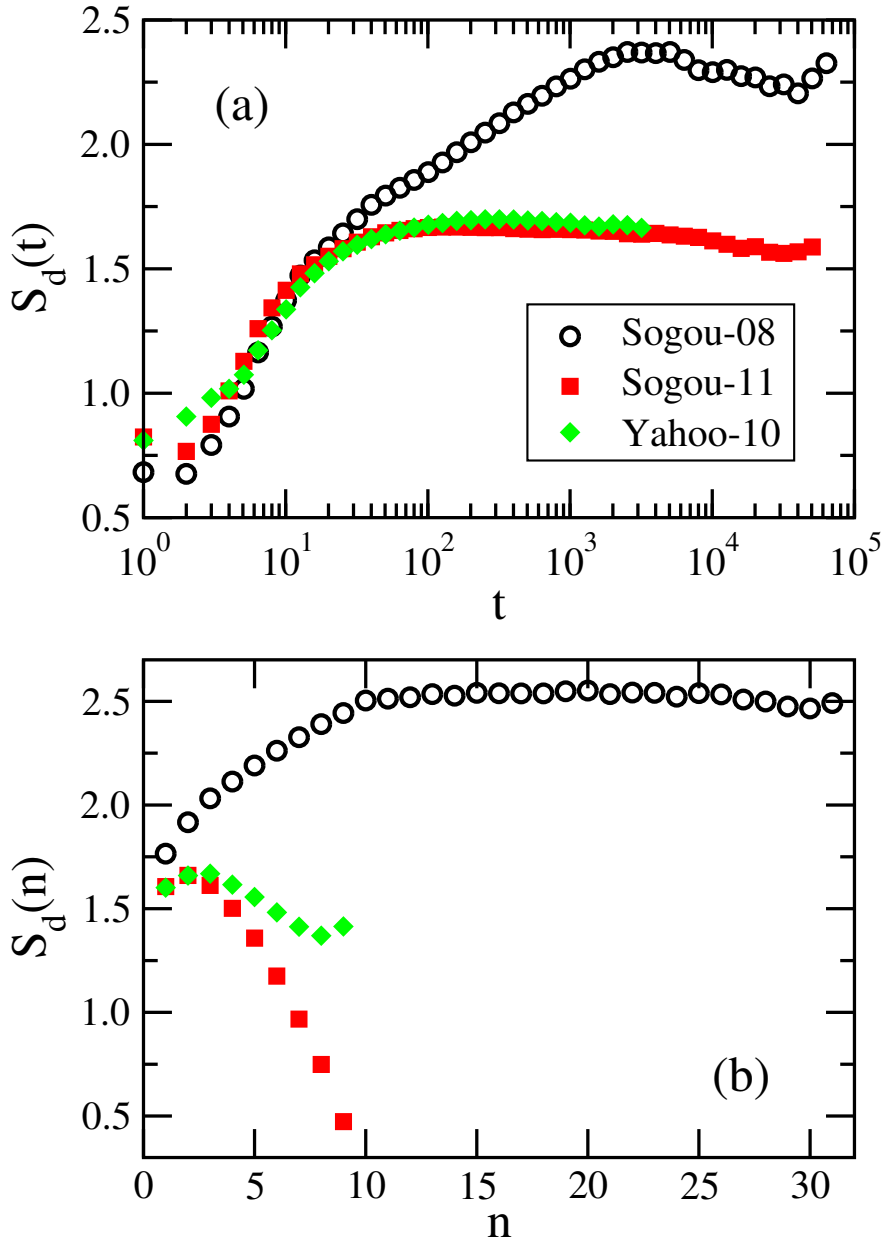

Figure 3.11: (a) Time dependence of the entropy $S_d(t)$ for the different sets. The early time regime is characterized by a strong increase of $S_d(t)$. For Sogou-11 and Yahoo-10 a plateau is reached much earlier than for Sogou-08 (of the order of a minute for Sogou-11 and Yahoo-10 and of the order of an hour for Sogou-08). For Sogou-08 the increase of the entropy at intermediate times is due to the predominance of long relocation jumps in this regime. (b) Entropy $S_d(n)$ as a function of the clicking order n. For Sogou-08, $S_d(n)$ reaches a plateau for $n \geq 11$.

to see to what extent these assumptions are fulfilled in human online search processes.

Well suited correlation coefficients for our purpose are Kendall's tau [117], τ_K, and Spearman's rho [118], ρ_S, as these two non-parametric measures are distribution-free and can handle power-law distributed quantities. Both coefficients are rank-based and measure the correspondence of two series of ordinal numbers. Series for waiting times and displacements are of course readily obtained from the original search engine click-through logs.

Table 3.4: Correlations between step-length and waiting time for the different data sets. The positive correlation coefficients indicate that for human online search processes spatial and temporal activities are not independent, with stronger correlations emerging for Sogou-08 than for Sogou-11 and Yahoo-10.

	Sogou-08	Sogou-11	Yahoo-10
$\tau_K(d,\tau)$	0.1273	0.0914	0.0878
$\rho_S(d,\tau)$	0.1721	0.1224	0.1175

Table 3.4 shows our results for the correlations between step-length d and waiting time τ. Both Kendall's tau and Spearman's rho provide correlation coefficients close to or larger than 0.10, indicating a weak positive correlation between step-length and waiting time. We also note that the correlations for Sogou-08 are larger than for Sogou-11 and Yahoo-10, illustrating again that different mechanisms underlie the different click-through logs. The positive correlation indicates that the assumption of independence between spatial and temporal activities, valid for both Brownian motion and Lévy flights, does not hold in a strict sense for human online search processes.

We can also check whether a process is memory-less or not. In order to do so we define for every step-length series $\{d_i\} = \{d_1, d_2, \cdots\}$ sets $\{d_{i+m}\} = \{d_{1+m}, d_{2+m}, \cdots\}$ with $m > 0$ and calculate the correlation coefficient $\tau_K(\{d_i\}, \{d_{i+m}\})$ and $\rho_S(\{d_i\}, \{d_{i+m}\})$. For a completely memory-less process these two coefficients should be zero. As we have previously seen that $S_d(n)$ exhibits a plateau, characteristic of a stationary process, only for Sogou-08 and $n \geq 11$,

Table 3.5: Correlations between successive displacements. The positive correlation coefficients indicate the presence of long-term memory effects in human online searches. We only calculated correlations for $\{d_i\}$ with $i \geq 11$.

correlation	Sogou-08	
m	$\tau_K(\{d_i\},\{d_{i+m}\})$	$\rho_s(\{d_i\},\{d_{i+m}\})$
1	0.2763	0.3511
2	0.2761	0.3496
3	0.2653	0.3358
4	0.2608	0.3297
5	0.2568	0.3243
6	0.2519	0.3176
7	0.2509	0.3157
8	0.2506	0.3149
9	0.2476	0.3105
10	0.2477	0.3102
20	0.2433	0.3001
30	0.2343	0.2852
40	0.2321	0.2784
50	0.2227	0.2638

we performed this analysis only for Sogou-08 and series $\{d_i\}$ with $i \geq 11$. The results shown in Table 3.5 and Fig. 3.12 reveal positive correlations even for large values of m. Long-term memory effects therefore permeate online human searches.

3.5 Discussion and conclusion

Online activities are an integral part of our daily lives. In many cases these activities involve search queries submitted to one of the search engines. In this chapter we proposed to view the exploration of the results (i.e., links) provided by the search engine as a foraging process on a semi-infinite line where the rank of a link corresponds to a coordinate on that line. Using a variety of space- and time-dependent quantities we investigated three different publicly available click-through logs.

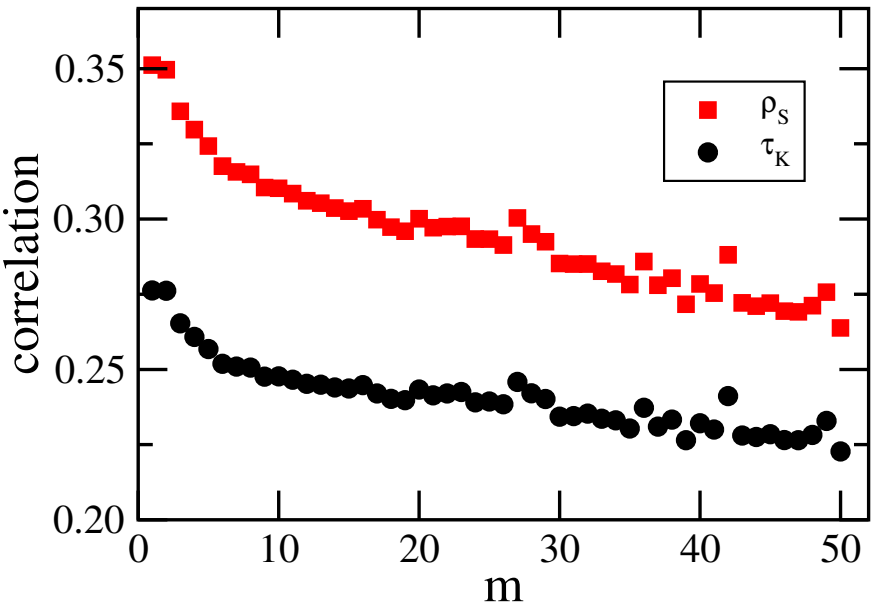

Figure 3.12: τ_K and ρ_S as a function of m for Sogou-08, see main text. These correlations decrease for increasing m but remain positive even for large m, revealing the presence of long-term memory effects in online human searches.

Our study reveals a sharp contrast between the oldest log and the two logs obtained more recently. For the two newer sets almost all queries can be understood as local searches restricted to a single page. These local searches have many characteristics of a Brownian random walk, as for example an exponential decay of the step-length distribution or a mean-square displacement that increases almost linearly with the clicking order. However, there are also marked differences, as a power-law behavior of the waiting time distribution or the directionality of the jumps along the semi-infinite line. A very interesting behavior emerges for the oldest data set, where both local searches as well as long-range power-law distributed jumps are found. This behavior is reminiscent of intermittent processes with Lévy strategies that have been proposed to describe searches with hidden targets. Interestingly, the waiting time distributions, i.e., the distributions of the times elapsed between successive clicks, are very similar for all three click-through logs and reveal a power-law behavior with an exponential cutoff as that encountered in other areas of human activities.

The different properties of the different data sets point to the evolution of the search engines over the years. Until recently search engines were of limited efficiency and as a result a sizable number of queries ended up with the user jumping from one page with links to another. These searches are therefore best characterized as a combination of local explorations and power-law distributed relocations. The more recent logs reveal an increased efficiency of the search engines where the overwhelming part of the queries yield only local searches on the first page of results. These local searches have some characteristics of Brownian motion.

The online search behavior clearly changes as a function of the efficiency of a search engine. Whereas efficient search engines result in an overwhelmingly large number of local searches, earlier engines prompted searches where local explorations and long-range, power-law distributed, relocations processes are combined. This scenario shows many common features with those of intermittent search processes which have been proposed as search strategies for finding hidden resources.

The description of online searches as foraging processes yields some interesting insights, but our results also reveal notable deviations from the simple models used to describe foraging processes. It remains a challenge to come up with a more realistic foraging model that is capable of reproducing the results we obtain from our analysis of the different click-through logs.

Appendix

3.A Data description and preparation

For our study we analyzed three different click-through data sets. In the following we briefly describe these sets as well as how we prepared the data and used them in our study.

Sogou-08 and Sogou-11 refer to the data sets *Search Engine Click-through Log Version 2008* and *Search Engine Click-through Log Version 2011* [119] which are parts of the "Sogou Lab Data." These two data sets provide millions of users' search queries and click-through activities on Sogou (www.sogou.com), which is one of the largest Chinese search engines. Sogou-08 has been collected in June 2008, whereas Sogou-11 contains queries submitted from 12/30/2011 to 01/01/2012. For Sogou-08 respectively Sogou-11 we get $51,537,388$ respectively $43,545,440$ lines of record, each corresponding to an individual click and consisting of the following components [119]:

Time of click | User ID | User query | Ranking of clicked result | Order of click | URL of clicked result

where "click" refers to the event of the user clicking on a link on the search engine result pages, whereas "time" is the calendar time of the click event.

Yahoo (YahooL18 [120]) is the data set *Anonymized Yahoo! Search Logs with Relevance Judgments version 1.0* provided by Yahoo! Labs [121] as part of the Yahoo! Webscope program [122] ("Approval for Access" granted October 7, 2013). It provides users' search click-through logs on Yahoo Search (search.yahoo.com) collected in July 2010. Yahoo-10 contains $80,779,266$ lines, where each line corresponds to an individual search and contains the following information (separated by "\t") [120]:

Query | Cookie | Timestamp | List of URLs | Number of "clicks" | List of time and click position/type pairs

where the "List of URLs" is the list of web links on the first result page. "Click" refers to any type of click on the result pages, including clicking on a search result, clicking at the top of a result page (unspecified; this can include clicking on the "also-try" button, on a spell correction suggestion, on an advertisement located at the top of the page, etc.), clicking at the bottom of a result page (unspecified, this can include clicking on next page button, on the bottom "also-try" button, on an advertisement located at the bottom of the page, etc.). For the "time and click position/type pairs," the "time" is the time (in seconds) since the beginning of the search. Since we do not know the specific activities done when clicking at the top or the bottom of a result page, we ignore the clicks at the top of result pages but assume that a click at the bottom of a result page is on the next page button.

From all these sets we removed entries with missing values for the first click as well as (for Sogou-08) unusual cases where a click was on a link of rank 1000 or above. For entries with successive clicks on the same link, we only kept the first click and jumped to the next click on a different link. The total number of queries retained after this procedure are listed in Table 3.1.

After this data cleaning all clicks belonging to the same search were grouped together and the corresponding (time, rank) pairs were calculated based on the order of the clicks. Time is the time passed in seconds since the click on the first result. In this way we end up for each search with a series of (time, rank) pairs

$$(t_1, r_1) \mid (t_2, r_2) \mid (t_3, r_3) \mid (t_4, r_4) \mid \cdots$$

with $t_1 = 0$. These series were then used as the starting point for our study.

3.B Maximum likelihood estimators

Let us consider first the discrete power-law distribution

$$P(k) = \frac{k^{-\alpha}}{\zeta(\alpha, k_{\min})}. \tag{3.17}$$

For a given data set $\{k_i\}$, we have the likelihood

$$\mathcal{L}(\alpha|\mathbf{k}) = \prod_{i=1}^{n} P(k_i) = \frac{\left(\prod_{i=1}^{n} k_i\right)^{-\alpha}}{\zeta\left(\alpha, k_{\min}\right)^n}, \tag{3.18}$$

where n is the number of data points. The log-likelihood is then given by

$$\ln \mathcal{L}(\alpha|\mathbf{k}) = -\alpha \sum_{i=1}^{n} \ln k_i - n \ln \zeta(\alpha, k_{\min}) \ , \tag{3.19}$$

and the maximum likelihood estimator (MLE) for α is obtained numerically from [123]

$$\hat{\alpha} = \arg\max_{\alpha} \ln \mathcal{L}(\alpha|\mathbf{k}) = \arg\max_{\alpha} \left(-\alpha \sum_{i=1}^{n} \ln k_i - n \ln \zeta(\alpha, k_{\min}) \right) \ . \tag{3.20}$$

We used the L-BFGS-B method for parameter optimization, see for example Ref. [124].

For the "shifted" geometric distribution

$$P(k) = p(1-p)^{k-k_{\min}}, \quad k \geq k_{\min}, \tag{3.21}$$

the likelihood is given by

$$\mathcal{L}(p|\mathbf{k}) = \prod_{i=1}^{n} P(k_i) = (1-p)^{\sum_{i=1}^{n} k_i - n k_{\min}} p^n, \tag{3.22}$$

where n is again the size of the data. The maximum likelihood estimator for p is

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} k_i - nk_{\min} + n} = \frac{1}{\bar{k} - (k_{\min} - 1)}, \tag{3.23}$$

where $\bar{k}$ is the mean of k_i's. Finally, for the exponential form

$$P(k) = \left(1 - e^{-\lambda}\right) e^{-\lambda(k - k_{\min})}, \tag{3.24}$$

we obtain

$$\hat{\lambda} = -\ln(1 - \hat{p}) = -\ln\left(1 - \frac{1}{\bar{k} - (k_{\min} - 1)}\right), \tag{3.25}$$

when $\bar{k} > k_{\min}$, since $\lambda = -\ln(1 - p)$ and MLE is invariant to this transformation.

The MLE for parameters in the other distributions are obtained in similar ways. For the Yule-Simon distribution the MLE for the shape parameter α is

$$\hat{\alpha} = \arg\max_{\alpha} \left(n \ln(\alpha - 1) + n \ln \Gamma\left(k_{\min} + \alpha + 1\right) - \sum_{i=1}^{n} \ln \Gamma\left(k_i + \alpha\right) \right); \tag{3.26}$$

whereas for the conditional Poisson distribution the MLE for μ is

$$\hat{\mu} = \arg\max_{\mu} \left(-n\mu - n \ln\left(1 - F_{\mu}(k_{\min} - 1)\right) + \sum_{i=1}^{n} k_i \ln \mu \right), \tag{3.27}$$

where $F_{\mu}(\cdot)$ is the cumulative distribution function of a Poisson distribution with rate parameter μ.

Finally for the distributions with more than one parameter, one has the following:

power-law with exponential cutoff

$$\{\hat{\alpha},\hat{\lambda}\} = \arg\max_{\alpha,\lambda}\left(-n\ln\left(Li_{\alpha}(e^{-\lambda}) - \sum_{i=1}^{k_{\min-1}} i^{-\alpha}e^{-\lambda i}\right) - \alpha\sum_{i=1}^{n}\ln k_i - \lambda\sum_{i=1}^{n}k_i\right); \quad (3.28)$$

discrete log-normal

$$\{\hat{\mu},\hat{\sigma}\} = \arg\max_{\mu,\sigma}\left(\sum_{i=1}^{n}\ln\left(\Phi\left(\frac{\ln(k_i+1)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(k_i)-\mu}{\sigma}\right)\right) - n\ln\left(1-\Phi\left(\frac{\ln(k_{\min})-\mu}{\sigma}\right)\right)\right); \quad (3.29)$$

and pairwise power law,

$$\{\hat{\alpha},\hat{\beta},\hat{k}_{\text{trans}}\} = \arg\max_{\alpha,\beta,k_{\text{trans}}}\left(n\ln C - \alpha\sum_{i=1}^{n}\ln k_i - (\beta-\alpha)\sum_{k_i\geq\lceil k_{\text{trans}}\rceil}(\ln k_i - \ln k_{\text{trans}})\right). \quad (3.30)$$

3.C Correlation coefficients

3.C.1 Kendall's tau

Kendall's tau provides a measure of rank correlation. Assuming a set of observations (x_1, y_1), (x_2, y_2), $\cdots$ of two joint random variables x and y (step-length and waiting time, for example), Kendall's tau compares the number P of concordant pairs with the number Q of discordant pairs:

$$\tau_K(x,y) = \frac{P-Q}{P+Q}\ . \quad (3.31)$$

Two pairs (x_l, y_l) and (x_m, y_m) are concordant if (1) $x_l < x_m$ and $y_l < y_m$ or (2) $x_l > x_m$ and $y_l > y_m$. If, however, $x_l < x_m$ and $y_l > y_m$ or $x_l > x_m$ and $y_l < y_m$, then they are discordant. In the case of data with tied ranks (which is the case for our series), Kendall's tau can be

calculated as

$$\tau_K(x,y) = \frac{\sum_{i<j} \operatorname{sgn}\left[(\mathrm{x_i} - \mathrm{x_j})(\mathrm{y_i} - \mathrm{y_j})\right]}{\sqrt{\frac{1}{2}n(n-1) - U}\sqrt{\frac{1}{2}n(n-1) - V}}, \tag{3.32}$$

where sgn is the signum function, whereas U and V are the numbers of x-tied pairs and y-tied pairs.

3.C.2 Spearman's rho

Spearman's rho is Pearson's correlation coefficient between ranked variables. Denoting by u_i the rank of x_i and by v_i the rank of y_i, then Spearman's rho can be expresses as:

$$\rho_S(x,y) = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}\sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2}}, \tag{3.33}$$

where

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i, \bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i. \tag{3.34}$$

While Spearman's rho and Kendall's tau usually yield different numbers, they work the same way.

Chapter 4

Behavior Analysis of Virtual-Item Gambling

The contents in this chapter were copied with permission from our publication [125]

X. Wang and M. Pleimling, Physical Review E 98, 012126 (2018). APS-Copyrighted.

Under Dr. Michel Pleimling's supervision, I contributed all the contents in this chapter.

4.1 Introduction

Recent years have seen a tremendous increase in online gambling, as witnessed by the emergence of numerous online gambling sites. This surge has yielded numerous recent scientific studies, with a focus on legal, social and psychological aspects, see Refs. [126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141] for some recent references. In parallel to this, the quick expansion of the video gaming industry has resulted in the formation of a huge market for virtual (in-game) items. Due to its easy accessibility, low barriers to entry, and immediate outcomes, virtual item gambling has become popular among game players. In virtual item gambling, instead of directly using cash, gamblers place bets with virtual items as virtual currencies [138, 139, 140]. The virtual items here particularly refer to in-game cosmetic skins from video games like Counter-Strike: Global Offensive, Team Fortress 2, DOTA 2, etc., which can be obtained through regular gameplay, in-game purchase, community market purchase, or trading among players. Based on current estimates, the virtual item gambling industry has reached the multi-billion-dollar level [142] and is expected to continue increasing. For such a booming industry, it becomes important to be able to model the complex virtual item gambling behaviors at both the individual and the aggregate levels. Indeed, understanding online gambling patterns is quickly becoming a pressing need for adolescent gambling prevention, virtual gambling regulation, and online irrationality research.

In this chapter we apply the methods of statistical physics in order to develop an understanding of the behavior of online gamblers. This is supplemented by the study of different random walk models that allow to recover some of the features extracted from the empirical data. While we are not aware of any previous similar attempts to investigate online gambling, we point out that related approaches have been used in the past in the study of horse race betting [143, 144]. More recently, online lowest unique bid auctions have been the

subject of different studies that successfully applied the toolbox of statistical and nonlinear physics [65, 67, 68, 145, 146, 147].

In the following we focus on a specific type of virtual item gambling, namely jackpot, a lottery style game which occupies about half of the virtual item gambling market [142]. Our analysis is based on the publicly available gambling logs from a medium-sized skin gambling site [148]. The rules of jackpot gambling are simple: players purchase lottery tickets with skins, there will be only one winning ticket, and the winner takes it all. In another way of speaking, this is a parimutuel betting type of gambling, where players place wagers in a pool, whereas only one player is chosen as the winner and wins all the wagers in the pool. The chance of winning equals the share of the player's wagers to the total wager pool.

In the next Section we provide a more in-depth discussion of jackpot gambling and of the data used in our analysis. We also discuss the models used for describing the distributions of different quantities as well as the model selection and parameter estimation. Section 4.3 summarizes results that we obtain from a statistical analysis of the gambling logs. In Section 4.4 we view the net income of players as random walks, whereas in Section 4.5 we discuss some random walk models that allow us to understand some of the behavioral data at the aggregate level. We conclude in Section 4.6.

4.2 Data and methods

4.2.1 Online jackpot game and gambling logs

The rules of the jackpot game are very simple. The gambling site constantly hosts a single jackpot game that any player can attend. A round can last from a few seconds to several minutes. To take part in the game, a player needs to place a bet with lottery tickets

purchased with one or several in-game skins deposited to the gambling site. Each ticket is usually equivalent to 1 US cent, and the values of the skins are calculated based on their prices listed in the community market. There is only one winning ticket in each round of the game. This winning ticket is drawn when the total number of skins deposited as wagers in that round exceeds a certain threshold. The draw is based on a uniformly distributed random number with a range equal to the total number of tickets purchased in that round. The player who holds the winning ticket will be the winner. The winner wins all the wagers, which are the deposited skins in that round, after a site cut (percentage cut) has been subtracted.

From the rules follows that in each round a player's winning chance is determined by the fraction their bet contributes to the total wager value of that round. With a site cut c the expected payoff η for one player with bet value b in a round with total wager j is then

$$\eta = (1-c)\, j \times \frac{b}{j} - b = -c\, b, \tag{4.1}$$

which is always negative due to the site cut. If the random number generator is well designed, then winning or losing a game is totally chance based, with no skill effort, similar to roulette in casinos. It is interesting to explore the players' gambling behaviors knowing that the expected net income is always negative. Fig. 4.1 provides an example of the total net income for a typical gambler. The movement consists of a large amount of small steps and a few large jumps which suggest the use of a random walk based model to describe the change of net income.

The publicly available gambling logs used in the following are published in the history page of the gambling site [148]. We collected the logs of 118590 gambling rounds, containing 943216 bets placed by 105307 players in 232 days, from March 10, 2015, the date the site was

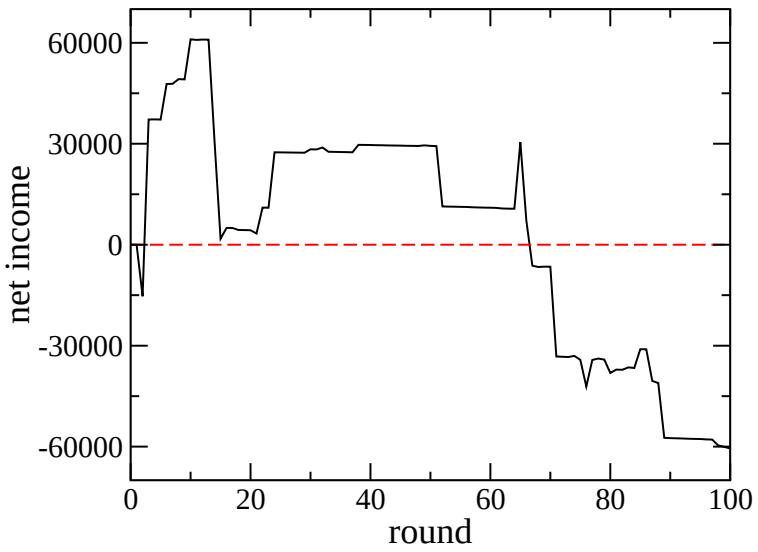

Figure 4.1: Net income vs the number of rounds played by an online gambler. Typically, these curves exhibit a large amount of small steps and a small amount of large steps.

established, to October 28, 2015. The total wager in our study sums to 2029835330 tickets, which is equivalent to about 20 million U.S. dollars, as calculated based on the players' deposited skin values. The competition is exclusively among players: The gambling site only takes cuts (3% of the total wager in each round), but is not directly involved in gambling, except through the drawing of the winning tickets. In each round, the winning ticket will be drawn when there are more than 50 skins placed as wagers. The dataset contains information on bet ID, round index, player ID, time stamp, number of tickets purchased, and winner ID. Various other quantities, such as current total number and final total number of purchased tickets, winning chance, net gain or net loss with and without site cut, can be calculated from these data.

The gamblers' wealth data have been collected in June 2017 from the game statistics site CS:GO BACKPACK [149], which provides the gamblers' inventory values based on the item prices listed in the community market in June 2017. The wealth data therefore have been collected two years after the gambling activities. In this way we obtained information on the

wealth data of 83249 out of the 105307 players that gambled in the time frame given above.

4.2.2 Ethics of data analysis

The data we analyzed in our study only contains publicly available information of gambling logs and in-game inventories, with no personally identifiable information included. On each data set, we performed a passive analysis with completely no interaction with any human subject. Before using the data, we acquired consent from the website administrators who host the data. We are not associated with any of those websites in any way. The purpose of our study is to help future researchers better understand human gambling behaviors in order to prevent adolescent gambling and problematic gambling.

4.2.3 Distributions and fitting models

Our analysis focuses on the probability distribution functions as well as on the complementary cumulative distribution functions (CCDF) of various quantities extracted from the empirical data. Whereas $P(X = x)$ is the probability that a random variable X takes on the value x, the corresponding complementary cumulative distribution function is given by

$$F(x) = 1 - P(X \leq x) = P(X > x) \ . \tag{4.2}$$

Power-law distributions and their variants have been found in previous studies of very different human activities [37, 52, 67, 78]. In online gambling quantities of interest often take on discrete values, which needs to be taken into account when selecting possible fitting models.

We consider six different fitting models in our distribution analysis. The discrete version of

a power-law distribution is given by [69]

$$P_1(x) = \frac{1}{\zeta(\alpha, x_{\min})} x^{-\alpha}, \tag{4.3}$$

with $x \geq x_{\min}$, $\alpha > 1$, and $\zeta(\cdot, \cdot)$ is the incomplete Zeta function. Here and in the following x is a positive integer value taken on by a random variable X. For some data sets a fat tail is terminated by an exponential decay, which can be taken into account by the discrete power-law distribution with exponential cutoff [78]

$$P_2(x) = \frac{1}{Li_\alpha\left(e^{-\lambda}\right) - \sum_{k=1}^{x_{\min}-1} k^{-\alpha} e^{-\lambda k}} x^{-\alpha} e^{-\lambda x}, \tag{4.4}$$

where $x \geq x_{\min}$, $\lambda > 0$, $\alpha > 0$, and $Li_\alpha(\cdot)$ is the polylogarithm function. Another heavy-tailed distribution is the log-normal distribution with the discrete version [78]

$$P_3(x) = \frac{\Phi\left(\frac{\ln(x+1)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)}{\Phi\left(\frac{\ln(x_{\min})-\mu}{\sigma}\right)}, \tag{4.5}$$

where $x \geq x_{\min}$, $\sigma > 0$, and $\Phi(\cdot)$ is the normal cumulative distribution. A forth basic model is the discrete exponential function [69]

$$P_4(x) = (1 - e^{-\lambda}) e^{\lambda x_{\min}} e^{-\lambda x}, \tag{4.6}$$

where $x \geq x_{\min}$ and $\lambda > 0$. Finally, we also consider two more complex models, namely the discrete shifted power-law distribution with exponential cutoff

$$P_5(x) = C \frac{(x-\delta)^{-\alpha}}{1 + e^{\lambda(x-\beta)}}, \tag{4.7}$$

where $x \geq x_{\min}$, $\lambda > 0$, $\delta < x_{\min}$, $\beta > x_{\min}$, and

$$C = \left(\sum_{k=x_{\min}}^{\infty} \frac{(k-\delta)^{-\alpha}}{1+e^{\lambda(k-\beta)}} \right)^{-1} \tag{4.8}$$

is the normalization factor, and the discrete pairwise power-law model [78]

$$P_6(x) = \begin{cases} C\ x^{-\alpha}, & x_{\min} \leq x < x_{\text{trans}}, \\ C x_{\text{trans}}^{\beta-\alpha}\ x^{-\beta}, & x_{\text{trans}} \leq x, \end{cases} \tag{4.9}$$

where $\alpha > 0$, $\beta > 1$, $x_{\text{trans}} > x_{\min}$, and the normalizing factor

$$C = \Big(\zeta(\alpha, x_{\min}) - \zeta(\alpha, x_{\text{trans}}) + x_{\text{trans}}^{\beta-\alpha} \zeta(\beta, x_{\text{trans}}) \Big)^{-1}. \tag{4.10}$$

We note that all these probability distributions contain a minimal value $x_{\min}$ that defines the range of values used for the modeling. For most quantities we choose as $x_{\min}$ the value of x that minimizes the Kolmogorov-Smirnov statistics between the empirical and fitted distributions [69].

For a given data set we estimate for each distribution the model parameters with the maximum likelihood method. The best fitting model is then selected using the Akaike Information Criterion (AIC). We refer the interested reader to Appendix B in Ref. [78] for a detailed discussion.

4.3 Behavioral analysis

4.3.1 Some basic statistics

In Table 4.1 we provide some basic statistics for the data used in our study. The huge diversity of the data is obvious from the very large values of the standard deviations. A meaningful analysis of the gambling data needs to consider probability distributions (or, equivalently, complementary cumulative distribution functions).

Table 4.1: Basic statistics for the gambling data used in this study.

	mean	minimum	maximum	standard deviation	50% percentile
bet value	2309.86	2	278247	8429.46	91
total net income	−578.88	−773524	751635	15513.36	−150
number of rounds a player attended	8.34	1	1931	31.94	2
number of players in a round	7.41	1	25	2.15	7
jackpot value	17116.41	100	396760	24399.50	7548

4.3.2 Distributions

A fundamental quantity for our analysis is the bet value, and the distribution of bet values allows one to gain a quick understanding of betting patterns. As shown in Fig. 4.2, the complementary cumulative distribution function for the bet value at the aggregate level is described by a shifted power law with an exponential cutoff: Bet values smaller than $\beta \sim 4.6\times10^4$ follow a power-law distribution, whereas very large bets are distributed exponentially (such guaranteeing a finite variance). The heavy-tail property of the bet distribution is also readily identified when studying the bet value distributions of individual gamblers. Fig. 4.3 shows the wager distribution for the nine players which played the largest numbers of rounds (between 1931 and 1286). While there is some variability in these distributions, they all exhibit heavy tails in the form of power laws with exponents typically in the range $[1.1, 1.7]$.

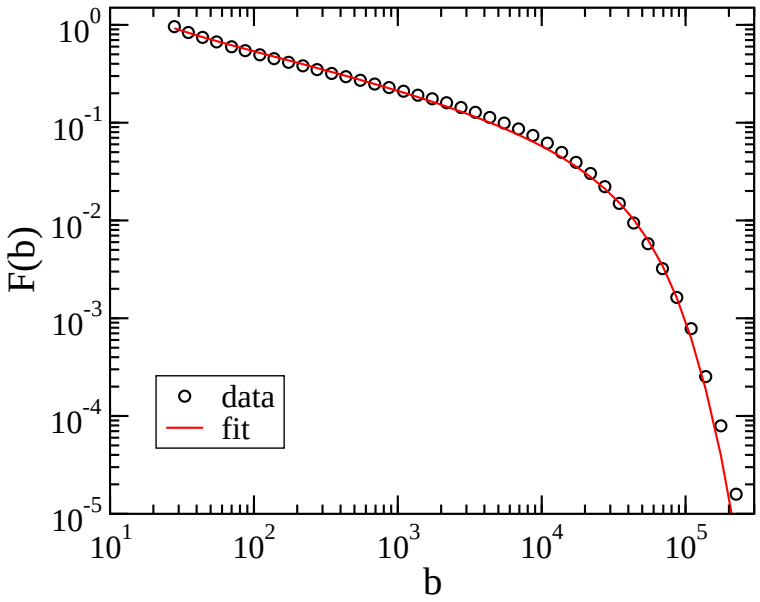

Figure 4.2: The complementary cumulative distribution function for bet values. The best fit is obtained for a shifted power law with an exponential cutoff, see Eq. (4.7), with $b_{min} = 25$ and the maximum likelihood estimators $\alpha = 1.297$, $\lambda = 3.429 \times 10^{-5}$, $\delta = 9.905$, and $\beta = 4.629 \times 10^{4}$.

In gambling a player's wealth provides a natural upper limit for possible bet values. Studies have shown that the net wealth distribution in human society follows a distribution that combines an exponential decay for small values and a power-law tail for large values [150]. For the online gamblers' wealth, this is different, see Fig. 4.4. We still have a power-law tail for large values (with an exponent $\beta = 2.442$), but for small values the exponential decay is replaced by a power-law decay with an exponent $\alpha = 1.128$. For this figure we computed the wealth of each player by taking the sum of the values (community market price) of the skins in each player's inventory.

In each round a gambler either loses their wager or wins the whole pool (minus the site cut), resulting in the random walk like behavior of the net income shown in Fig. 4.1. The probability distribution of the pool size is described by a power-law distribution with an exponent $a = 0.650$ that ends in an exponential cutoff, see Fig. 4.5. The same functional

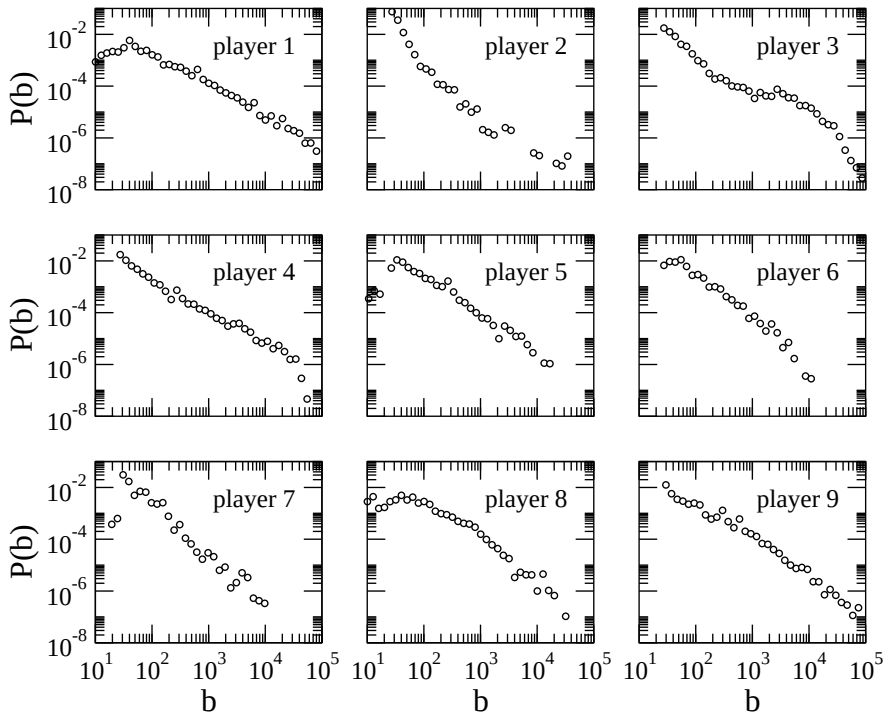

Figure 4.3: Wager probability distributions for the nine players with the largest numbers of bets (ranging from 1931 bets for player 1 to 1286 for player 9). Heavy tails are present in all nine distributions.

form is found if we consider the wins instead of the pool sizes (see Section 4.4).

The available logs also allow to discuss time dependent quantities, as for example the waiting time t_w, defined as the time measured in seconds between successive bets by the same user, or the number of rounds r played by individual gamblers. The waiting time probability distribution shown in Fig. 4.6 has some interesting features. The plateau for $P(t_w)$ close to $t_w = 10^5$ indicates that a sizable portion of gamblers play bets day after day (24 hours correspond to 86,400 seconds). The heavy tail of the distribution reveals that some persons restart gambling after a month-long hiatus, which illustrates some of the challenges gambling prevention faces. Fig. 4.7 shows that the number of rounds played by individual players during the 232 days covered by the gambling logs is well described by a log-normal distribution. Remarkably, a sizable number of gamblers placed a thousand and more bets

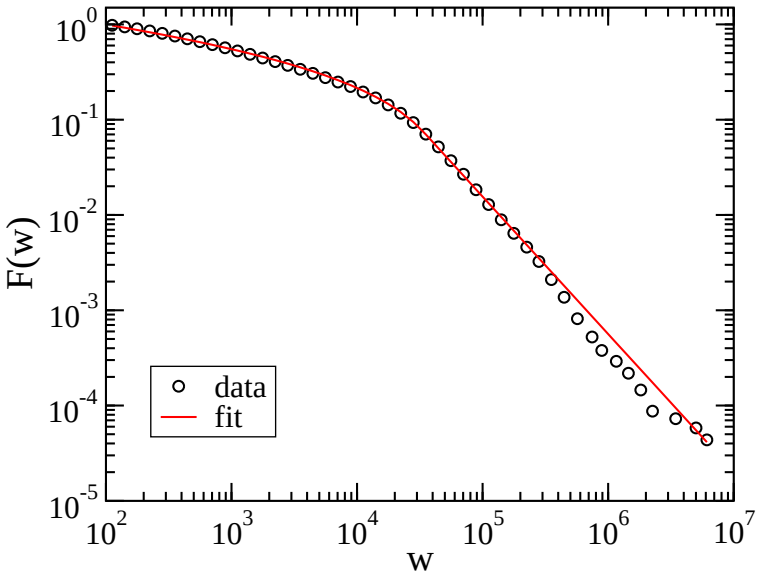

Figure 4.4: The complementary cumulative distribution function of the gambler's wealth w, where one unit corresponds to 1 US Cent. These data have been collected in June 2017 from the game statistics site CS:GO BACKPACK [149]. The best fit of the data is achieved with a pairwise power-law distribution (4.9) with the maximum likelihood estimator $\alpha = 1.128$ and $\beta = 2.442$ as well as with the parameters $w_{\text{min}} = 100$ and $w_{\text{trans}} = 33928$.

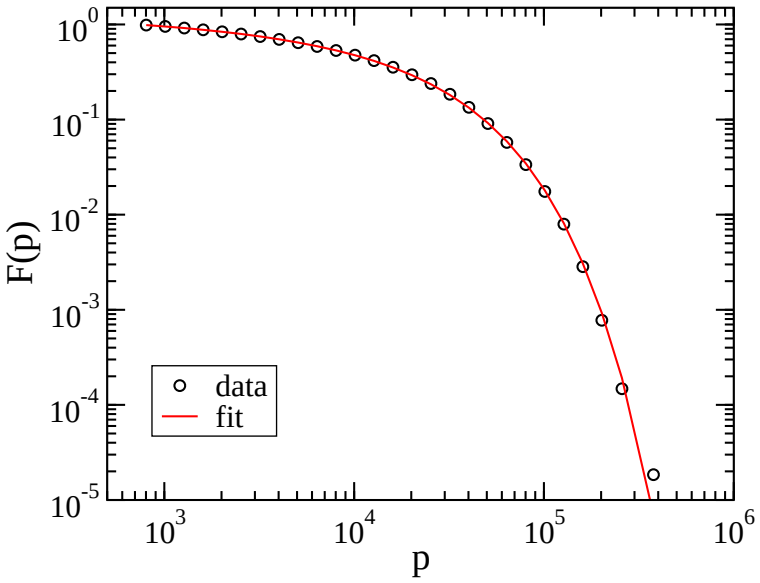

Figure 4.5: The complementary cumulative distribution function of the pool size (i.e., the total wager in one round) p. The fitting curve is a power law with exponential cutoff (4.4) with the maximum likelihood estimators $\alpha = 0.650$ and $\lambda = 2.577 \times 10^{-5}$.

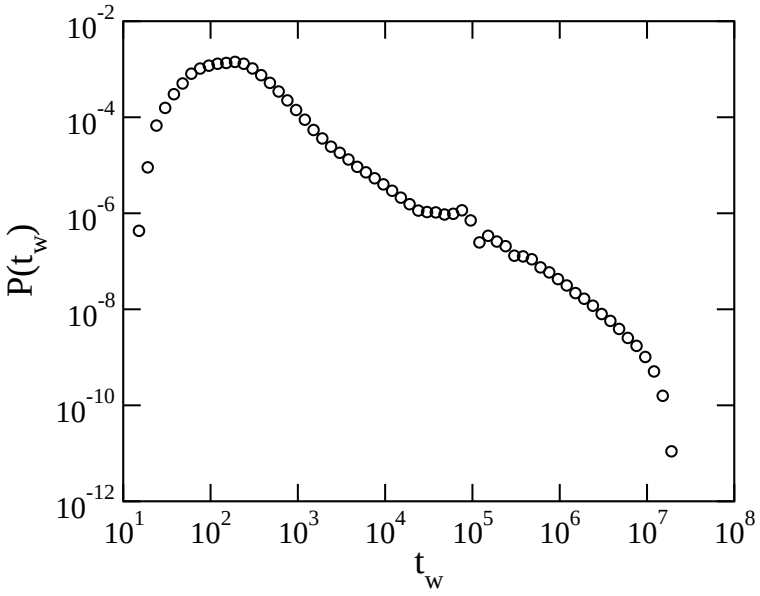

Figure 4.6: The probability distribution of the waiting time between successive gambles. The waiting time is measured in seconds.

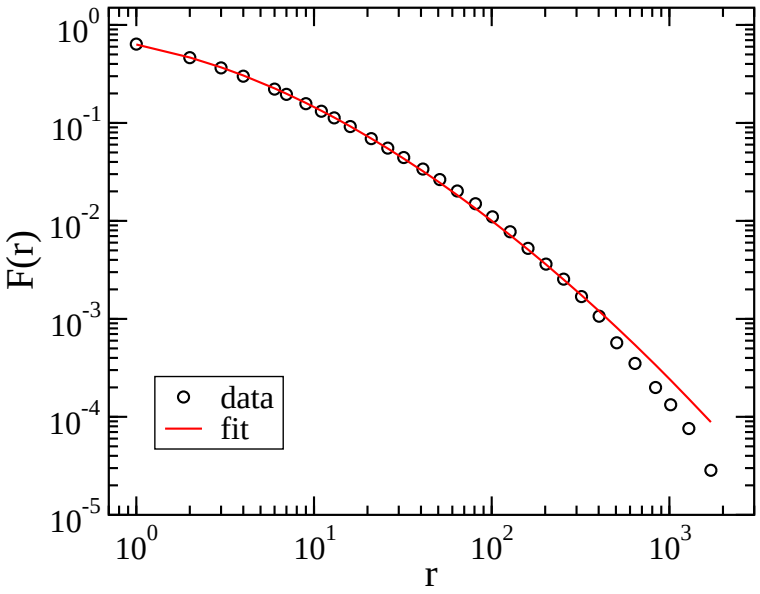

Figure 4.7: The complementary cumulative distribution function of the number of rounds r played by individual players. The data are best fitted by a log-normal distribution (4.5) with the maximum likelihood estimators $\mu = -1.777$ and $\sigma = 2.238$.

during the time frame covered by the logs.

Besides discussing data at the population level, we can also identify different sub-groups of

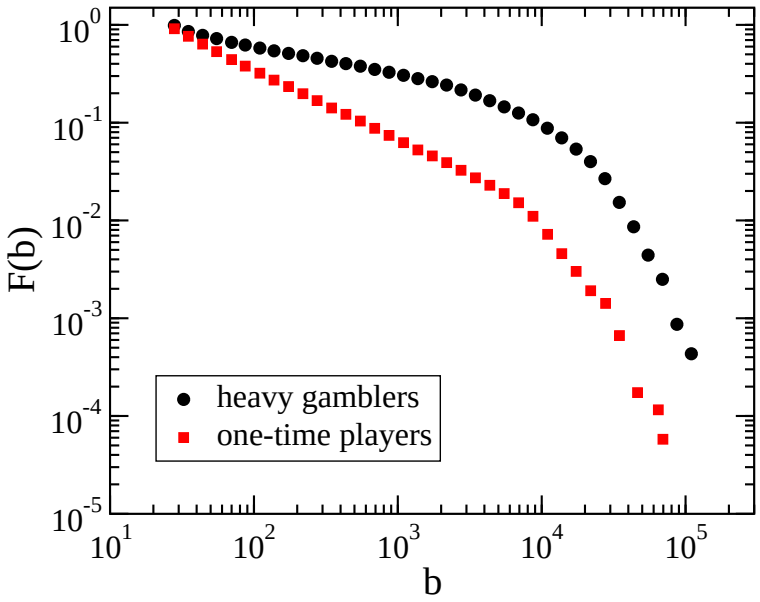

Figure 4.8: Comparison of the betting patterns of heavy gamblers and one-time players. Shown is the complementary cumulative distribution function for bet values.

gamblers and discuss differences between these groups. Fig. 4.8 provides one example where we confront the distribution of bets of one-time players with that of heavy gamblers (defined as having played at least 600 rounds). Obviously one-time players are much more risk-averse and are therefore unlikely to bet large amounts.

4.3.3 Correlations

Correlation coefficients help to understand the relationships between the different quantities. As our quantities, be it outcomes, bets, and profits, all follow heavy-tailed distributions, the standard Pearson's product-moment correlation coefficient may provide erroneous results. More appropriate are rank-based correlation coefficients, such as Kendall's tau [117] or Spearman's rho [118]. We verified that the same conclusions are obtained from these two coefficients. For that reason we will only discuss Kendall's tau in the following. Assuming a set of observations $\{(x_i, y_i)\}$ of two joint variables x and y, Kendall's tau can be calculated

as

$$\tau_K(x,y) = \frac{\sum_{i<j} \mathrm{sgn}\left[(x_i - x_j)(y_i - y_j)\right]}{\sqrt{\frac{1}{2}n(n-1) - U}\sqrt{\frac{1}{2}n(n-1) - V}}, \tag{4.11}$$

where sgn is the signum function, whereas U and V are the numbers of x-tied pairs and y-tied pairs.

For each player the gambling history can be summarized as a sequence $\{(b_i, o_i)\}$, where b_i is the value of the i-th bet and o_i is the outcome of that round. When losing the round, then the outcome is the negative of the bet value, whereas for a winning round o_i is the total bet value minus the winner's wager and the site cut. Focusing on the 2,318 players that attended more than 60 rounds, we can obtain from these data different correlation coefficients.

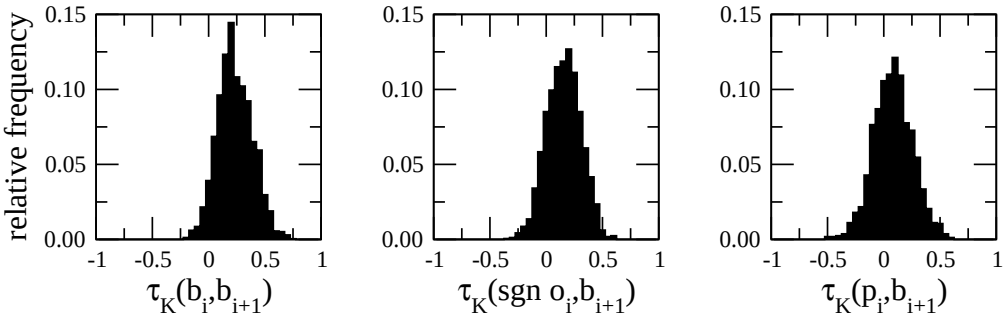

Figure 4.9: The relative frequencies for the three correlation coefficients discussed in the text. Left panel: Correlation between successive bets, with the mean value 0.260. Center panel: Correlation between the sign of a bet outcome and the next bet, with the mean value 0.181. Right panel: Correlation between the profit and the subsequent bet, with the mean value 0.107.

The correlation between successive bets $\tau_K(b_i, b_{i+1})$ is positive for most players, with an average value $\tau_K = 0.260$. The relative frequency of a given value of $\tau_K(b_i, b_{i+1})$ is displayed in the left panel of Fig. 4.9. In order to understand this graph we remark that a negative value is obtained when a gambler places larger and smaller bets in turn, whereas placing bets randomly yields a value close to zero. From the graph follows that only a few gamblers have these types of gambling behaviors. Instead, for most gamblers bets are not independent but

indicate some level of memory. Indeed, positive correlation indicates a consistent betting behavior without dramatic changes from bet to bet.

Also shown in Fig. 4.9 are the relative frequencies for the correlation between the sign of a bet outcome and the next bet, $\tau_K(\mathrm{sgn\ o_i}, \mathrm{b_{i+1}})$, and the correlation between the profit p_i (i.e., the value of the outcome in case it is positive) and the subsequent bet, $\tau_K(p_i, b_{i+1})$. The first correlation coefficient helps us to understand how gaining or losing money affects the next bet, whereas the second one shows whether a bet value is affected by the value of the previous profit. Profit corresponds to positive outcome, so that for the computation of $\tau_K(p_i, b_{i+1})$ we remove all bets with a negative outcome o_i. For both correlations we restrict ourselves to players who made profit in at least 15 rounds and had negative outcomes in also at least 15 rounds. This yields 1,608 eligible players. The relative frequencies shown in the center and right panels of Fig. 4.9 reveal for most players a weak positive correlation between the betting value and the outcome or profit. There is a tendency for gamblers to place larger (respectively, smaller) bets in case the outcome in the preceding round was positive (respectively, negative).

4.4 Net income viewed as a random walk

As we have already seen in Fig. 4.1, the net income of a player changes at each round where they place a bet, due to winning or losing that round. This then generates a time series where "time" is increased by one at each round played by the gambler and suggests a description as a random walk in the one-dimensional space of net income. Of course the random walkers are not independent as the loss of one gambler will be part of the gain of another one. Also, the fact that every gambler has a finite wealth will put constraints on the random walk.

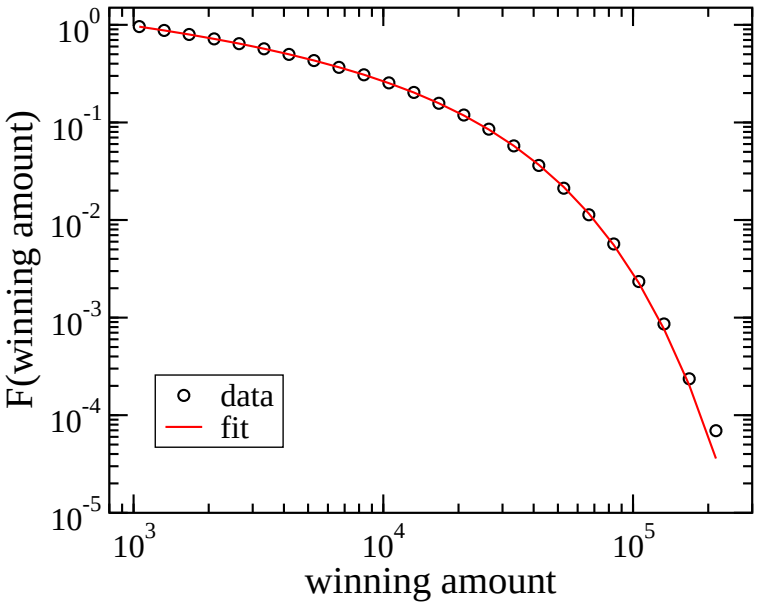

Figure 4.10: The complementary cumulative distribution function of the winning amounts. The fitting curve is a power law with exponential cutoff (4.4) with the maximum likelihood estimators $\alpha = 1.063$ and $\lambda = 3.192 \times 10^{-5}$.

The jumps done by our random walkers have the peculiarity that they follow different distributions depending on whether they jump "left" (net income decreases after losing a round) or "right" (net income increases after winning a round). "Left" and "right" indicate the relative decrease or increase with respect to the value of the net income before the round is played. The distribution of losses is very similar to the distribution of bet values (as in a given round all bets result in losses with the exception of the winning bet). As shown in Fig. 4.2, this distribution is described at the aggregate level by a shifted power law with an exponential cutoff. Power laws are also observed in Fig. 3 for individual gamblers. The distribution of winning amounts shown in Fig. 4.10 is well described by a power-law distribution with an exponential cutoff, albeit with a different power-law exponent α. The fact that the distributions for jumps in both directions, albeit not identical, are power-law distributions indicates that the random walk of the net income should follow a truncated Lévy flight pattern.

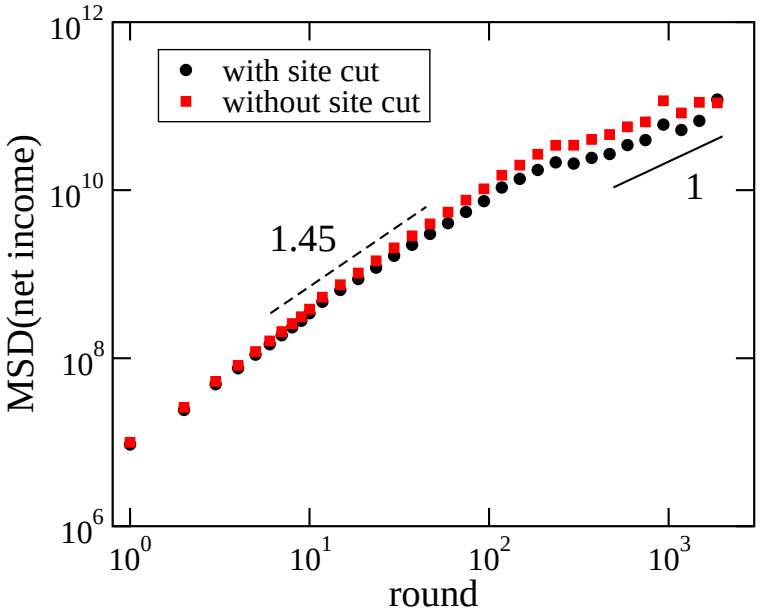

Figure 4.11: Mean-squared displacement when viewing the net income of the gamblers as a random walk, with time measured in numbers of rounds played. Independent on whether the site cut is considered or not, two different regimes are observed, with the early one being superdiffusive with an exponent close to 1.45, whereas the later one is close to normal diffusion.

Fig. 4.11 shows that the mean-squared displacement of the net income random walk displays a first regime that is superdiffusive with an exponent close to 1.45. We show two curves in that figure, one where we consider as winning amount the total pool size in a round and one where we subtract the site cut and take the remaining amount as the length of the jump. At very late times this first regime goes over into a normal diffusion regime, with the measured slope close to 1 in the log-log plot. This crossover from superdiffusion to normal diffusion is in fact expected for truncated power-law distributions [151, 152] and has been observed in a variety of systems (see, e.g., Refs. [153, 154, 155]).

A quantity of much interest is the first-passage time [156], i.e., the time needed for a stochastic variable (in our case the net income viewed as a random walker) to take on for the first time a given value. Indeed, the first-passage time distribution can help to determine the diffusive behavior of a stochastic process [151, 152]. For our stochastic process N_r, $r = 1, \cdots, R$,

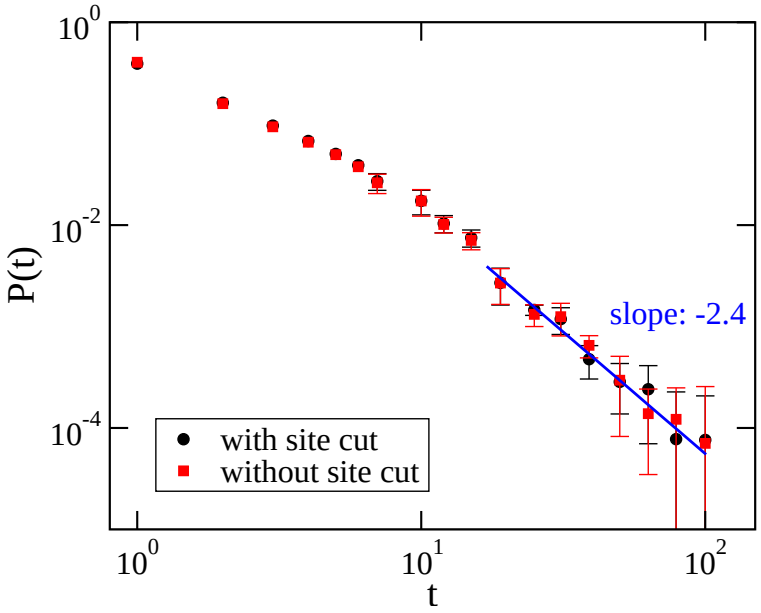

Figure 4.12: First-passage time distribution obtained from the data of 387 players that gambled in more than 200 rounds. The superdiffusive regime is revealed by a power-law decay with an exponent larger than 3/2. Error bars result from log-binning averaging and indicate 95% confidence intervals.

representing the net income with R being the maximum number of rounds played, the first-passage time is defined by $t = \min\{r > r_0; X_k = \pm N_{fp}\}$, where N_{fp} is the target value. As shown in Ref. [152], the first passage time distribution $P(t)$, defined as the survival probability that, starting from $r = r_0$, the series N_r stays within the range $[N_{r_0} - N_{fp}, N_{r_0} + N_{fp}]$ up to the round $r = r_0 + t$, is given by the expression

$$P(t) = \lim_{R\longrightarrow\infty} \frac{1}{R} \sum_{r=1}^{R} \Theta\left(|N_{r+t} - N_r| - N_{fp}\right) - \lim_{R\longrightarrow\infty} \frac{1}{R} \sum_{r=1}^{R} \Theta\left(|N_{r+t-1} - N_r| - N_{fp}\right) \ , \quad (4.12)$$

where $\Theta(x)$ is the Heaviside step function.

As Eq. (4.12) requires sufficiently long time series, we focus on the 387 players that played at least 200 rounds and choose $N_{fp} = 500$. The resulting first-passage time distribution is still very noisy. In order to reduce the noise we use the log-binning technique which yields

the distribution shown in Fig. 4.12. Inspection of that figure reveals that after some initial time regime a superdiffusive regime prevails, as indicated by a slope larger than 3/2, the characteristic value for a Gaussian process. As already mentioned, for any truncated heavy-tail distribution the long-time behavior should be normal diffusion, and we do observe the crossover from a superdiffusive to a normal diffusive behavior in Fig. 4.11 for the mean-squared displacement. For the first-passage time distribution obtained from the gambling logs the long-time normal diffusion decay with an exponent 3/2 is not readily observed, due to the shortness of the available time series.

4.5 Modeling online gambling through random walk models

In order to better understand this switch from a superdiffusive to a normal diffusive behavior in the net income random walk we discuss in the following three different random walk models. The aim of this investigation is not so much to find the best parameter sets to reproduce the empirical data, but instead to gain insights into the necessary ingredients to obtain from these models data with qualitative similar properties as those derived from the gambling logs.

In all three models we consider that at each round four players interact (this is mostly useful for the numerical simulations; the analytical results for the simpler models are valid for any number of gamblers interacting in a round). For each round the gamblers place a bet with a value taken from the continuous power-law distribution with exponential cutoff

$$P(b) = \frac{\lambda^{1-\alpha}}{\Gamma\left(1-\alpha, \lambda b_{min}\right)} b^{-\alpha} e^{-\lambda b}, \tag{4.13}$$

with $\lambda > 0$ and $b \geq b_{min}$ (we choose $b_{min} = 1$), whereas $\Gamma(\cdot, \cdot)$ is the incomplete Gamma function. This distribution (4.13), motivated by the data from the gambling logs that show a power-law behavior with an exponential cutoff, is the continuous version of the discrete distribution (4.4). For the results discussed in the following we fix the mean $\langle b \rangle = 100$. The two parameters α and λ are then not independent but related through that mean bet value as $\langle b \rangle = \frac{\Gamma(2-\alpha,\lambda)}{\lambda\Gamma(1-\alpha,\lambda)}$. We vary α between 1.2 and 1.6. We verified that qualitatively our results are unchanged if instead of using the distribution (4.13) we use a power-law distribution with a sharp truncation:

$$P(b) = C\, b^{-\alpha}, \tag{4.14}$$

with $b \in [b_{min}, b_{max}]$ and $C = \frac{\alpha-1}{b_{min}^{1-\alpha} - b_{max}^{1-\alpha}}$ where b_{max} and α are related when fixing the mean bet value.

Our first two models are focusing on a single gambler with infinite wealth. In model 1 [1]. we fix the winning chance of this gambler to be 1/4 (in a generalization to n interacting gamblers, the winning chance would be $1/n$). This model does not take into account that in the online game the winning chance is proportional to the bet value. We therefore consider a more realistic model 2 which implements this relationship between the bet value and the winning chance. Model 3, finally, is a more sophisticated version of model 2 where, similarly to the online game, a large pool of gamblers is available (the data shown below have been obtained for $N = 1,000,000$) and at each round $n = 4$ gamblers are selected randomly to play the round. We calculate quantities for all players, which are no longer independent, in contrast to models 1 and 2, and after each round we update the net income of all 4 players involved in that round. We also take into account in model 3 that the wealth of each player is finite: before the first round is played every gambler is assigned a wealth taken from the

[1] We thank an anonymous referee for suggesting this model.

power-law probability distribution

$$P(w) = \frac{1}{w_{min}} \left(\frac{w}{w_{min}} \right)^{-2}, \tag{4.15}$$

with $w_{min} = 1$. As in model 3 the gamblers have been provided with finite wealth, and the individual net income random walks all have an absorbing state of zero wealth. As soon as the wealth of a gambler is zero, this gambler is removed from the pool. While it is tempting to discuss our random walkers in the context of previous studies of random walk type motion with absorbing boundaries [157, 158, 159, 160], it is crucial to realize that our random walkers are not independent, but instead at each round the winner's step length is correlated to those of the losers.

Similarly to our analysis of the empirical data, we compute in the following for the different models the mean-squared displacement (MSD) of the net income as well as the distribution of the first-passage time at which the income of a gambler takes on a given target value.

We start by noting that for models 1 and 2 the mean-squared displacement as a function of time (i.e., the number of rounds played) can be computed exactly, see Appendix. For rounds involving each time n gamblers and a fixed mean bet value $\langle b \rangle$, the MSD is given for model 1 by

$$\mathrm{MSD}(t) = \left(\frac{2(n-1)}{n} \mu_2 + \frac{(n-1)(n-2)}{n} \langle b \rangle^2 \right) t \ , \tag{4.16}$$

with μ_2 being the second moment of the bet distribution, whereas for model 2 one obtains

$$\mathrm{MSD}(t) = (n-1) \langle b \rangle^2 t \ . \tag{4.17}$$

Fig. 4.13(a) and 4.13(b) display these curves for three different values of the parameter α found in the bet distribution (4.13), with $\langle b \rangle = 100$ and $n = 4$.

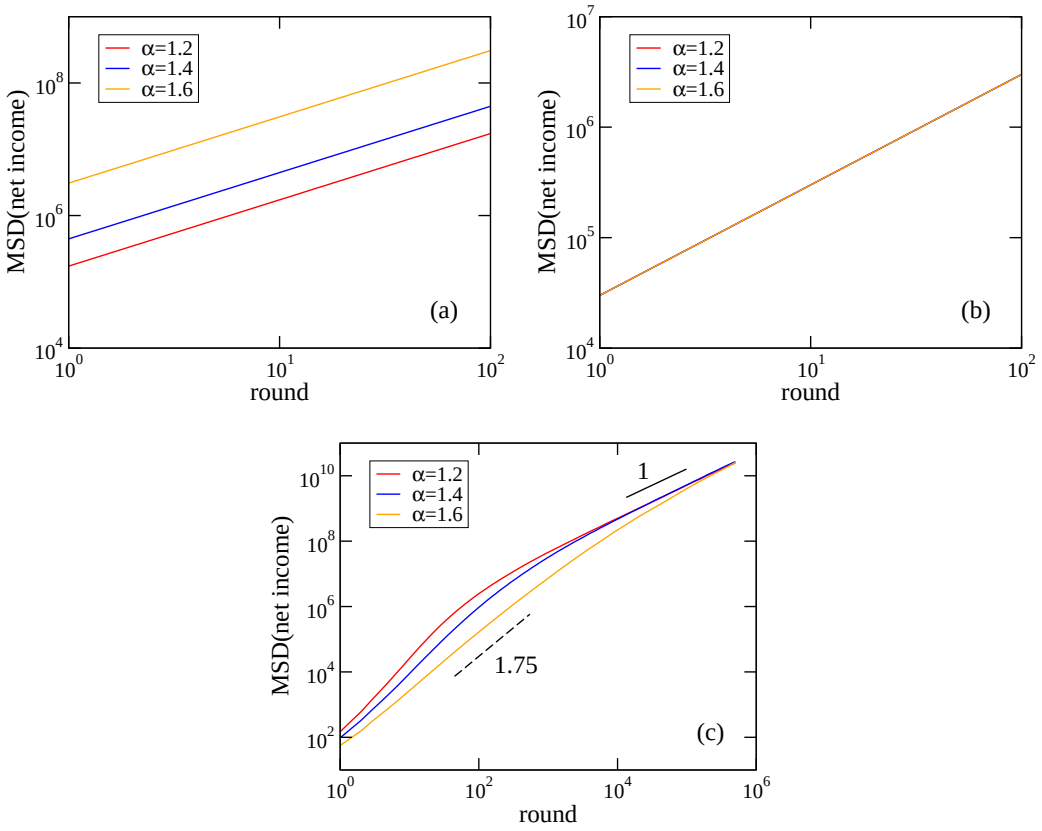

Figure 4.13: The mean-squared displacement for (a) model 1, (b) model 2, and (c) model 3. For models (1) and (2), the net income of each gambler performs an independent random walk where the step length is related to the bet distribution (4.13). In these two cases the mean-squared displacement increases linearly with time (i.e. the number of rounds played), in agreement with prior results. Model 3, on the other hand, reveals a crossover from superdiffusion to diffusion. The different curves are for different values of the parameter α in the continuous power-law distribution with an exponential cutoff (4.13).

Several comments are in order. First we note that although we consider a truncated power-law distribution, we obtain that the MSD increases linearly with time. This is in agreement with an early observation of a linearly increasing MSD encountered in simulations of truncated Lévy flights in two dimensions [161]. This linear time dependence is very general as the bet (i.e. step length) distribution only enters through the mean and the second moment. Especially for model 2 any distribution with the same mean yields the same MSD as expression (4.17) does not depend on the variance. While we are focusing on the two truncated

power-law distributions (4.13) and (4.14), even a distribution with finite mean and infinite second moment yields for model 2 a finite MSD growing linearly with time. This is different for model 1 as the second moment explicitly enters in expression (4.16). As a result of this dependence, the MSDs for different values of α, see Fig. 4.13a, are shifted vertically, due to the fact that changing α while keeping $\langle b \rangle$ constant changes the value of the second moment, see Appendix. We further note that these two models do not allow to obtain a behavior similar to that observed in Fig. 4.11 for the empirical data, namely a crossover from a superdiffusive behavior with an exponent larger than 1 to a normal diffusive behavior characterized by a linear increase of the MSD. This, however, is different for model 3 where we indeed observe a crossover from superdiffusion to normal diffusion, see Fig. 4.13c for data obtained for one million gamblers playing 50 million rounds, with each round involving four randomly selected gamblers. As we can not compute the MSD analytically for this model, we can only provide a heuristic argument for this observation. We note that in model 3 all gamblers have finite wealth taken from the distribution (4.15). One of the consequences of this is that we add an absorbing boundary (a gambler is removed once their wealth becomes zero), another one is that initially many players have a small wealth and therefore can only bet small amounts. Consequently, at early rounds the mean bet value of active players is smaller than the mean of the bet distribution (4.13), which makes the MSD to be smaller than what one obtains for models 1 and 2. As time increases, some gamblers are eliminated as their wealth hits the absorbing boundary. As a result, the wealth of the active gamblers increases until their mean bet values are getting close to the mean value of the bet distribution (4.13). At this point the MSD for model 3 shows a crossover from a superdiffusive behavior to a normal diffusive one.

Fig. 4.14 shows our results for the first-passage time distributions obtained from simulations of the three different models with the bet distribution (4.13). For models 1 and 2 we simulate

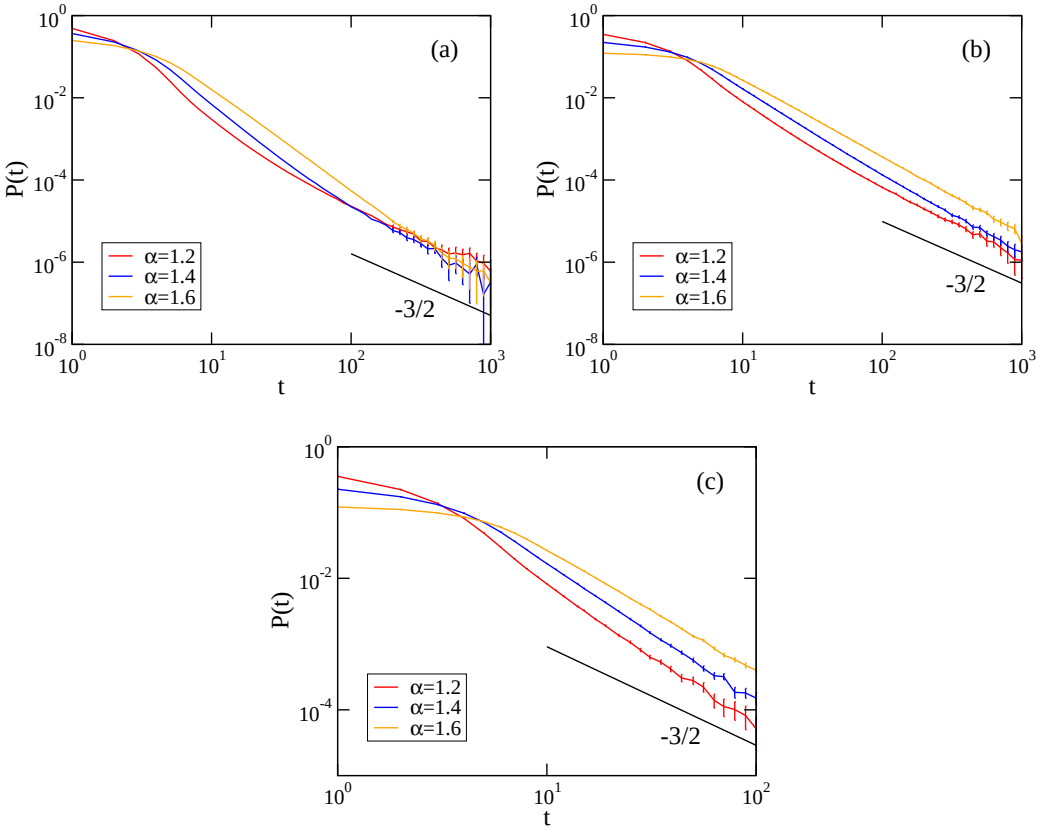

Figure 4.14: The first-passage time distribution for (a) model 1, (b) model 2, and (c) model 3. For all three models we observe a crossover from a superdiffusive behavior, revealed by a decay faster than $t^{-3/2}$, to a normal diffusive behavior proportional to $t^{-3/2}$. Error bars indicate 95% confidence intervals. The different curves are for different values of the parameter α in the bet distribution (4.13).

a gambler who plays 50 million rounds, which yields a time series of their net income of length 50 million. The data shown in Fig. 4.14(a) and 4.14(b) result from averaging over 300 independent runs for $\alpha = 1.2$, 250 independent runs for $\alpha = 1.4$, and 75 independent runs for $\alpha = 1.6$. These difference in the number of independent runs reflects an increase of computational costs when α increases, due to a decrease of the acceptance rates for generating random numbers. For model 3 we only made one run with 1 million players and 50 million rounds. For all three models we use $n = 4$, $\langle b \rangle = 100$, and $N_{fp} = 20$.

Interestingly, all three models show in the first-passage time distribution the expected

crossover from a superdiffusive behavior at early times, characterized by a decay with an effective exponent larger than 3/2, to a normal diffusive long-time behavior, where the distribution decays as $t^{-3/2}$. This crossover is rather sharp for model 1, whereas it is more gradual for the other two models. As already mentioned in Ref. [152], the first-passage time distribution does not suffer from the same restrictions than the MSD and is therefore the superior quantity for identifying the crossover between a superdiffusive and a normal diffusive regime.

4.6 Summary

The quickly increasing video gaming industry has led to the development of other types of online entertainment, the prime example being online gambling. We considered in this work an online jackpot game as an example of virtual item gambling. Publicly available gambling logs permit a behavioral analysis at both the aggregate and individual levels. We analyzed the probability distribution functions and correlation coefficients in order to elucidate the relationships between some quantities derived from the gambling logs. Viewing the changes of the net income of a gambler as a random walk, the mean-squared displacement of the net income displays a crossover from a superdiffusive to a diffusive behavior. We discussed three different models, two of which are simple random walk models for a gambler with infinite wealth, whereas the third one considers many gamblers with finite wealth. All three models show a crossover from superdiffusive to normal diffusive behavior in the first-passage time distribution, but only the model with finite wealth displays a similar crossover in the mean-squared displacement.

Appendix

4.A Mean-squared displacements

In the following we briefly discuss the exact calculation of the mean-squared displacements for models 1 and 2. In both models we consider n players with infinite wealth who gamble with identically and independently distributed bet values b taken from a distribution $P(b)$. In the main text we consider a power-law distribution (4.13) as well as a power-law distribution with a sharp truncation (4.14).

Let $A, B, C, \cdots$ be the n players attending one round. We are going to focus on player A and compute the mean-squared displacement of their net income. For simplicity, we will also use $A, B, C, \cdots$ to represent the bet values of the corresponding players. We denote by $A_1, A_2, \cdots, A_t$ the bet values of player A in t rounds and call $\Omega_1, \Omega_2, \cdots, \Omega_t$ the sum of the bet values of the other players in the corresponding rounds. We use a_t to represent the net income of player A after t rounds and note that before the first round played the net income is zero, i.e., $a_0 = 0$. The mean-squared displacement is then given by

$$\mathrm{MSD}(t) = \left\langle (a_t - a_0)^2 \right\rangle = \left\langle a_t^2 \right\rangle \ . \tag{4.18}$$

When player A wins round t, then their net income increases by $\Omega_t = B_t + C_t + \cdots$, but the net income decreases by $-A_t$ in case of a loss. Models 1 and 2 now differ by the probability to win the round, with this probability being given by $1/n$ for model 1 and by $A_t/(B_t + C_t + \cdots)$ for model 2.

Let us first look at model 2. In that case the mean-squared displacement at round t is given

by

$$\mathrm{MSD}(t) = \int\limits_{A_1,\Omega_1,\cdots,A_t,\Omega_t} P\left(A_1,\Omega_1,\cdots,A_t,\Omega_t\right)\left(\frac{A_1}{A_1+\Omega_1}\cdots\frac{A_t}{A_t+\Omega_t}\left(\Omega_1+\cdots+\Omega_t\right)^2\right.$$
$$+\frac{\Omega_1}{A_1+\Omega_1}\cdots\frac{A_t}{A_t+\Omega_t}\left(-A_1+\cdots+\Omega_t\right)^2+\cdots+\frac{A_1}{A_1+\Omega_1}\cdots\frac{\Omega_t}{A_t+\Omega_t}\left(\Omega_1+\cdots-A_t\right)^2$$
$$\left.+\frac{\Omega_1}{A_1+\Omega_1}\cdots\frac{\Omega_t}{A_t+\Omega_t}\left(-A_1-\cdots-A_t\right)^2\right)\mathrm{d}A_1\mathrm{d}\Omega_1\cdots\mathrm{d}A_t\mathrm{d}\Omega_t.$$

After expanding the squared terms most terms cancel out, yielding after some simple algebraic manipulations

$$\begin{aligned}\mathrm{MSD}(t) &= \int\limits_{A_1,\Omega_1,\cdots,A_t,\Omega_t} P\left(A_1,\Omega_1,\cdots,A_t,\Omega_t\right)\left(A_1\Omega_1+\cdots+A_t\Omega_t\right)\mathrm{d}A_1\mathrm{d}\Omega_1\cdots\mathrm{d}A_t\mathrm{d}\Omega_t,\\ &= \int\limits_{A_1,\Omega_1} P(A_1,\Omega_1)A_1\Omega_1\mathrm{d}A_1\mathrm{d}\Omega_1+\cdots+\int\limits_{A_t,\Omega_t} P(A_t,\Omega_t)A_t\Omega_t\mathrm{d}A_t\mathrm{d}\Omega_t.\end{aligned}$$

All these t terms are identical and are given by

$$\begin{aligned}\int\limits_{A,\Omega} P(A,\Omega)A\Omega\mathrm{d}A\mathrm{d}\Omega &= \int\limits_{A,B,C,\cdots} P(A)P(B)P(C)\cdots\left(A(B+C+\cdots)\right)\mathrm{d}A\mathrm{d}B\mathrm{d}C\cdots,\\ &= \int\limits_{A} P(A)A\mathrm{d}A\left(\int\limits_{B} P(B)B\mathrm{d}B+\int\limits_{C} P(C)C\mathrm{d}C+\cdots\right),\\ &= \mu\left((n-1)\mu\right)=(n-1)\mu^2,\end{aligned}$$

where $\mu=\langle A\rangle=\int\limits_{A} P(A)A\mathrm{d}A$. This then yields the final result

$$\mathrm{MSD}(t)=(n-1)\langle A\rangle^2+\cdots+(n-1)\langle A\rangle^2=(n-1)\langle A\rangle^2 t.$$

We note that for model 2 the MSD grows linearly in time. As the MSD is independent of the second moment of the bet distribution, it is the same for any bet distribution with the

same mean, including distributions with finite mean and infinite second moment.

The calculation for model 1 closely follows that of model 2, but with the major change that for gambler A the probability of winning a round is $1/n$, whereas the probability of losing that round is $(n-1)/n$, with n being the number of gamblers involved in a round. This then yields the expression

$$\begin{aligned}\mathrm{MSD}(t) = \langle a_t^2 \rangle = \int\limits_{A_1,\Omega_1,\cdots,A_t,\Omega_t} P\left(A_1,\Omega_1,\cdots,A_t,\Omega_t\right) \Bigg(& \frac{1}{n}\cdots\frac{1}{n}\left(\Omega_1+\cdots+\Omega_t\right)^2 \\ & + \frac{n-1}{n}\cdots\frac{1}{n}\left(-A_1+\cdots+\Omega_t\right)^2 + \cdots + \frac{1}{n}\cdots\frac{n-1}{n}\left(\Omega_1+\cdots-A_t\right)^2 \\ & + \frac{n-1}{n}\cdots\frac{n-1}{n}\left(-A_1+\cdots-A_t\right)^2 \Bigg) \mathrm{d}A_1\mathrm{d}\Omega_1\cdots\mathrm{d}A_t\mathrm{d}\Omega_t \ ,\end{aligned}$$

and after some algebraic manipulations,

$$\begin{aligned}\mathrm{MSD}(t) &= \int\limits_{A_1,\Omega_1,\cdots,A_t,\Omega_t} P\left(A_1,\Omega_1,\cdots,A_t,\Omega_t\right) \left(\frac{n-1}{n}A_1^2 + \frac{1}{n}\Omega_1^2 + \cdots + \frac{n-1}{n}A_t^2 + \frac{1}{n}\Omega_t^2 \right) \\ & \qquad\qquad \mathrm{d}A_1\mathrm{d}\Omega_1\cdots\mathrm{d}A_t\mathrm{d}\Omega_t, \\ &= \int\limits_{A_1,\Omega_1} P(A_1,\Omega_1)\left(\frac{n-1}{n}A_1^2+\frac{1}{n}\Omega_1^2\right)\mathrm{d}A_1\mathrm{d}\Omega_1 + \cdots \\ & \qquad + \int\limits_{A_t,\Omega_t} P(A_t,\Omega_t)\left(\frac{n-1}{n}A_t^2+\frac{1}{n}\Omega_t^2\right)\mathrm{d}A_t\mathrm{d}\Omega_t.\end{aligned}$$

Again, these t terms are identical, with

$$\begin{aligned}
&\int\limits_{A,\Omega} P(A,\Omega)\left(\frac{n-1}{n}A^2+\frac{1}{n}\Omega^2\right)\mathrm{d}A\mathrm{d}\Omega \\
=&\int\limits_{A,B,C,\cdots} P(A,B,C,\cdots)\left(\frac{1}{n}(B+C+\cdots)^2+\frac{n-1}{n}A^2\right)\mathrm{d}A\mathrm{d}B\mathrm{d}C\cdots, \\
=&\frac{n-1}{n}\int\limits_{A} P(A)A^2\mathrm{d}A+\frac{1}{n}\int\limits_{B+C+\cdots} P(B)P(C)\cdots\left(B^2+C^2+2BC+\cdots\right)\mathrm{d}B\mathrm{d}C\cdots, \\
=&\frac{n-1}{n}\mu_2+\frac{1}{n}\left(\int\limits_{B} P(B)B^2\mathrm{d}B+\int\limits_{C} P(C)C^2\mathrm{d}C+\cdots\right)+\frac{1}{n}\left(\int\limits_{B,C} P(B)P(C)\ 2BC\ \mathrm{d}B\mathrm{d}C+\cdots\right), \\
=&\frac{n-1}{n}\mu_2+\frac{1}{n}(n-1)\mu_2+\frac{1}{n}\frac{(n-1)(n-2)}{2}2\mu^2, \\
=&\frac{2(n-1)}{n}\mu_2+\frac{(n-1)(n-2)}{n}\mu^2,
\end{aligned}$$

with the mean $\mu=\langle A\rangle=\int\limits_A P(A)A\mathrm{d}A$ and the second moment $\mu_2=\langle A^2\rangle=\int\limits_A P(A)A^2\mathrm{d}A$. It follows that the MSD for model 1 is given by

$$\mathrm{MSD}(t)=\left(\frac{2(N-1)}{N}\mu_2+\frac{(N-1)(N-2)}{N}\mu^2\right)t,$$

and is therefore still proportional to t, but now the pre-factor depends on both the mean and the second moment.

The table below provides the interested reader with the mean values and second moments for the two bet distributions considered in this work. $\Gamma(\cdot,\cdot)$ is the incomplete Gamma function.

Table 4.2: Mean and second moment for the different bet value distributions.

distribution model	mean μ	second moment μ_2
power law with exponential cutoff (4.13)	$\frac{1}{\lambda}\frac{\Gamma(2-\alpha,\lambda b_{min})}{\Gamma(1-\alpha,\lambda b_{min})}$	$\frac{1}{\lambda^2}\frac{\Gamma(3-\alpha,\lambda b_{min})}{\Gamma(1-\alpha,\lambda b_{min})}$
power law with sharp truncation (4.14)	$\frac{1-\alpha}{2-\alpha}\frac{b_{min}^{2-\alpha}-b_{max}^{2-\alpha}}{b_{min}^{1-\alpha}-b_{max}^{1-\alpha}}$	$\frac{1-\alpha}{3-\alpha}\frac{b_{min}^{3-\alpha}-b_{max}^{3-\alpha}}{b_{min}^{1-\alpha}-b_{max}^{1-\alpha}}$

Chapter 5

Wagering Distribution, Risk Attitude and Anomalous Diffusion in Online Gambling of Pure Chance

Under Dr. Michel Pleimling's supervision, I contributed all the content in this chapter.

5.1 Introduction

Gambling, as a form of games, appeared at a very early age in human history. Today, it has become a huge industry and has a huge social impact. According to a report by the American Gaming Association [162], commercial casinos in the United States alone made total revenue of over 40 billion US dollars in 2017. On the other hand, different studies reported that $0.12\% - 5.8\%$ of the adults and $0.2\% - 12.3\%$ of the adolescents across different countries in the world are experiencing problematic gambling [163, 164]. Studying the gamblers' behavior patterns not only contributes to the prevention of problematic gambling and adolescent gambling, but also helps to a better understand human decision-making processes. Researchers have put lots of attention on studying gambling-related activities. Economists have proposed many theories about how humans make decisions under different risk conditions. Several of them can also be applied to model gambling behaviors. For example, the prospect theory introduced by Kahneman and Tversky [165] and its variant cumulative prospect theory [166] have been adopted in modeling casino gambling [167]. In parallel to the theoretical approach, numerous studies focus on the empirical analysis of gambling behaviors, aiming at explaining the motivations behind problematic gambling behaviors. However, parametric models that quantitatively describe empirical gambling behaviors are still missing. Such models can contribute to evaluating gambling theories proposed by economists, as well as better understanding of the gamblers' behaviors. In this chapter, our goal is to provide such a parametric model for describing human wagering activities and risk attitude during gambling from empirical gambling logs. However, it is very difficult to obtain gambling logs from traditional casinos, and it is hard to collect large amounts of behavior data in a lab-controlled environment. Therefore in this chapter we will mainly focus on analyzing online gambling logs collected from online casinos.

Recent years have seen an increasing trend of online gambling due to its low barriers to entry,

high anonymity and instant payout. For researchers of gambling behaviors, online gambling games present two advantages: simple rules and the availability of large amounts of gambling logs. In addition to the usual forms of gambling games that can be found in traditional casinos, many online casinos also offer games that follow very simple rules, which makes analyzing the gambling behavior much easier as there are much fewer degrees of freedom required to be considered. On the other hand, many online casinos have made gambling logs publicly available on their websites, mainly for verification purposes, which provides researchers with abundant data to work on. Due to the high popularity of online gambling, in a dataset provided by an online casino there are often thousands or even hundreds of thousands of gamblers evolved. Such a large scale of data can hardly be obtained in a lab environment. Prior research has begun to make use of online gambling logs, for example, Meng's thesis [168] presented a pattern analysis of typical gamblers in Bitcoin gambling. It's worth arguing that although our work only focuses on the behaviors of online gamblers, there is no reason to think that our conclusions cannot be extended to traditional gamblers.

Naturally, we can treat the changing cumulative net income of a player during their gambling activities as a random walk process. We are particularly interested in the diffusive characteristics of the gambler's net income. This is another reason why we want to analyze the wager distribution and risk attitude of gamblers, since both distributions are closely related to the displacement distribution for the gambler's random walks. Within this chapter, we will mainly focus on the analysis at the population level. Physicists have long been studying diffusion processes in different systems, and recently, anomalous diffusive properties have been reported in many human activities, including human spatial movement [53, 169, 170], and information foraging [78] (see in Ch. 3). In Chapter 4, we have shown that in a parimutuel betting game (where players gamble against each other), a gambler's net income displays a crossover from superdiffusion to normal diffusion. We have reproduced this crossover in

simulations by introducing finite and overall conserved gamblers' wealth. However, this explanation cannot be used in other types of gambling games where there is no interaction among gamblers (e.g., fixed-odds betting games, which will be introduced below), as they violate the conservation of gamblers' overall wealth. In this chapter, we want to expand the scope of our study to more general gambling games, check the corresponding diffusive properties, and propose some explanations for the observed behaviors.

5.1.1 Game Types and Rules

To uncover the commonalities behind the behavior patterns of online gambling, we analyze the data from different online gambling systems. The first one is skin gambling, where the bettors are mostly video game players and where cosmetic skins from online video games are used as virtual currency for wagering [125, 171]. The other system is crypto-currency gambling, where the bettors are mostly crypto-currency users, and different types of crypto-currencies, for example Bitcoin, are used for wagering. As the overlap of these two communities, video game players and crypto-currency users, is relatively small for now, the common features of gambling patterns between these two gambling systems are possible to be common features among online gamblers.

Not only do we consider different gambling systems, but also we discuss different types of gambling games. In this chapter, we discuss four types of solely probability-based gambling games, whose outcomes in theory will not benefit from the gamblers' skill or experience when the in-game random number generators are well designed. In general, there are two frameworks of betting in gambling: fixed-odds betting, where the odds is fixed and known before players wager in one round; and parimutuel betting, where the odds can still change after players place the bets until all players finish wagering. In fixed-odds betting, usually

players bet against the house/website, and there is no direct interaction among players; and in parimutuel betting, usually players bet against each other. The four types of games we discuss in this chapter will cover both betting frameworks.

When a player attends one round in any of those games, there are only two possible outcomes: either win or lose. When losing, the player will lose the wagers they placed during that round; whereas when winning, the prize winner receives equals their original wager multiplied by a coefficient. This coefficient is generally larger than 1, and in gambling terminology, it is called odds in decimal format [172, 173]. Here we will simply refer to it as odds. Note that the definition of odds in gambling is different than the definition of odds in statistics, and in this chapter, we follow the former one. When a player attends one round, their chance of winning is usually close to, but less than the inverse of the odds. The difference is caused by the players' statistical disadvantage in winning compared to the house due to the design of the game rules. In addition, the website usually charges the winner with a site cut (commission fee), which is a fixed percentage of the prize.

We further define the *payoff* o_p to be the net change of one player's wealth after they attend one round. Although the four types of games are based on different rules, the payoff all follow the same expression

$$o_p = \begin{cases} -b, & \text{with probability } p = 1 - \dfrac{1}{m} + f_{\mathrm{m}}, \\ (1-\eta)(m-1)b, & \text{with probability } q = 1 - p = \dfrac{1}{m} - f_{\mathrm{m}}, \end{cases} \tag{5.1}$$

where $b > 0$ is the wager the player places, $m > 1$ is the odds, $1 > \eta \geq 0$ corresponds to the site cut, and f_{m} is a non-negative value based on the odds representing the players' statistical disadvantage in winning we mentioned earlier. At least either η or f_{m} are non-zero.

From Eq. (5.1), we can obtain the expected payoff of attending one round

$$\begin{aligned} E\left(o_p|m,b\right) &= \Big(-(1-1/m+f_m) + (1-\eta)(m-1)(1/m-f_m)\Big)b, \\ &= -\Big((1-\eta)mf_m + (1-1/m+f_m)\eta\Big)b \equiv -\xi b, \end{aligned} \tag{5.2}$$

which is always negative since either η or f_{m} are non-zero. In gambling terminology, ξ is called the house edge, from which the websites make profits. The house edge represents the proportion the website will benefit on average when players wager. In the four types of games we discuss, the house edge ξ ranges from 1% to 8%. If there is no house edge $\xi = 0$, that means it is a fair game. In a fair game or when we ignore the house edge, the expected payoff would be 0. Below, we list the detailed rules of the different games:

Roulette We focus on a simplified version of roulette games that appears in online casinos, where a wheel with multiple slots painted with different colors will be spun, after which a winning slot will be selected. Each slot has the same probability to be chosen as the winning slot. Players will guess the color of the winning slot before the game starts. The players have a certain time for wagering, after which the game ends and a winning slot is selected by the website. Those players who successfully wagered on the correct color win, the others lose. As the chance of winning and odds for each color are directly provided by the website, roulette is a fixed-odds betting game.

Crash "Crash" describes a type of gambling games mainly hosted in online casinos. Before the game starts, the site will generate a crash point m_C, which is initially hidden to the players. With a lower boundary of 1, the crash point is distributed approximately in an inverse square law. The players need to place their wager in order to enter one round. After the game starts, on the player's user interface a number, called multiplier, will show up and gradually increase from 1 to the predetermined crash point m_C, after which the game ends.

During this process, if the player "cash-outs" at a certain multiplier m, before the game ends, they win the round; otherwise they lose. This multiplier m they cashed out at is the odds, which means when winning, the player will receive a prize equals his wager multiplied by m. When m_C is generated with a strict inverse-square-law distribution, the winning chance exactly equals the inverse of the player-selected odds m.

The player can also set up the cash-out multipliers automatically before the game starts, to avoid the possible time delay of manual cash-out. Since in a manual cash-out scenario, after the game starts, the multiplier will show up on the screen; at a given moment, the decision of the cash-out multiplier is based on the player's satisfaction with the current multiplier, and involves more complicated dynamics of decision-making processes. Meanwhile, in an auto cash-out scenario, the multiplier m is chosen before the game starts, which means the decision making is more "static." In this manuscript, due to the lack of high-resolution temporal information, when discussing the risk attitude in crash games, we will only focus on the auto cash-out scenario. In order to include more data, when discussing the wager distribution, both scenarios will be included in our analysis, since in both scenarios, the players need to wager first before the game starts. Crash is also a fixed-odds betting game where the odds are player-selected.

Satoshi Dice Satoshi Dice is one of the most popular games in crptocurrency gambling. In 2013, the transactions resulting from playing Satoshi Dice games accounted for about 60% of overall Bitcoin transactions [174]. When playing Satoshi Dice, the player needs to pick a number A within a range (L, U) provided by the website when they wager. The odds can be calculated with the expression $(U - L)/(A - L)$. Once the player finishes wagering, the website will pick another number B which is uniformly distributed on (U, L). If B is less than A, then the player wins the round, otherwise they lose.

Satoshi Dice is a fixed-odds betting game. In some online casinos, players cannot choose A

arbitrarily, but instead, they have to select A from a preset list provided by the gambling website. Since the odds m is determined from A, we are more interested in the case where the players can choose A arbitrarily, from which we can obtain a more detailed distribution of the odds m, which helps us to understand the players' risk attitude.

According to the rules of Satoshi Dice games, the maximum allowed bet is proportional to the inverse of A, which means the accepted range of wager is directly related to the odds. To simplify our modeling work, in wager distribution analysis, we will only focus on the bets that are wagered on the same odds.

Jackpot Unlike the games discussed above, Jackpot is a parimutuel betting game, where players gamble against each other. During the game, each player attending the same round will deposit their wager to a pool. The game-ending condition varies across different websites, it could be a certain pool size, a certain amount of players, or a preset time span. When the game ending condition is reached, each player's winning chance will be determined by the fraction of their wager in the wager pool, based on which one player will be chosen as the winner by the website. The winner will obtain the whole wager pool as the prize, after excluding the site cut. The odds can be calculated by the pool size divided by the player's wager, but it is unknown to the players at the moment they wager. In the previous chapter, we have already discussed the player's behavior in Jackpot games of skin gambling where in-game skins are directly used as wagers. In this chapter, we will extend the analysis to other cases.

5.2 Data Summary

For each type of game, we collect two datasets. In total, we analyze 8 datasets collected from 4 different online gambling websites, and the number of bet logs contained in each

dataset ranges from 0.3 million to 19.2 million. Due to the high variation of market prices of crypto-currencies and in-game skins, the wager and deposits are first converted into US cents based on their daily market prices.

CSGOFAST From the skin gambling website CSGOFAST [175], we collected four datasets on the Roulette, Crash and Jackpot games (*csgofast-Double*, *csgofast-X50*, *csgofast-Crash*, *csgofast-Jackpot*) it provides.

csgofast-Double (A) is a Roulette game in which players can bet on 3 different colors (Red, Black, Green), which respectively provide odds of (2, 2, 14). The data were collected in two different time periods, and the only difference between them is a change of the maximum allowed bet values. *csgofast-X50* (B) is also a Roulette game in which players can bet on 4 different colors (Red, Blue, Green, Gold), which respectively provide odds of (2, 3, 5, 50).

csgofast-Crash (C) is a Crash game. As we mentioned earlier, when analyzing the risk attitude of gamblers in Crash game, we are more interested in how players set up the odds (multiplier) with the automatically cash-out option. On CSGOFAST, under the automatically cash-out option, players can only setup odds ranging from 1.10 to 50. The interesting point about this dataset is that even if the player loses the round, if they used the automatically cash-out option, it still displays the player-selected odds (which is set before the game starts); meanwhile if they used the manually cash-out option, no odds is displayed. Therefore in early-crashed games ($m_C < 1.10$), all the displayed odds that is larger than 1.10 were placed with automatically cash-out option. These displayed odds will be used in odds distribution analysis. The data are also collected in two different periods, where the only difference is still a change of the maximum allowed bet value. Roulette and Crash games on CSGOFAST all use virtual skin tickets for wagering.

csgofast-Jackpot (H) is a Jackpot game, where in-game skins are directly placed as wagers. Each skin has a market value that ranges from 3 to 0.18 million US cents. A player can place at most 10 skins in one round.

CSGOSpeed From skin gambling website CSGOSpeed [176], we collected one dataset from its Jackpot game *csgospeed-Jackpot* (G), in which arbitrary amounts of virtual skin tickets can be used as wagers. The difference between datasets (H) and (G) focuses on whether the wagers are in-game skins or virtual skin tickets.

ethCrash ethCrash [177] is a cryto-currency gambling website providing a Crash game *ethCrash* (D). Players need to place wagers in Ethereum (ETH), one type of crypto-currency.

SatoshiDice SatoshiDice [178] is a cryto-currency gambling website which accepts Bitcoin Cash (BCH) as wagers. It provides a Satoshi Dice game *satoshidice* (E), where only 11 preset odds can be wagered on, ranging from 1.05 to 1013.74. Among the preset odds, we find that more than 30% of the bets are placed under the odds 1.98, and we will analyze those bets for wager distribution.

Coinroll Coinroll [179] is a cryto-currency gambling website which accepts Bitcoin (BTC) as wagers. It provides a Satoshi Dice game *Coinroll* (F), where players can either wager on the 8 preset odds listed by the website, or choose an odds of their own. When further analyzing the data, we find that a few players placed an unusual large amount of bets, where the top player placed more than 11 million bets. Although these large number of bets again prove the heavy-tailed distribution of the number of bets of individuals, we have doubts that these players are playing for the purpose of gambling. Indeed, prior studies have raised suspicion about the use of crypo-currency gambling websites as a way for money laundering [180]. We will therefore exclude from our analysis gamblers who placed more than half a million bets. For bets wagered on the preset odds, we find that more than 57% are placed under the odds

1.98, and we use these bets to analyze the wager distribution.

On the other hand, since player-selected odds show a better spectrum about the risk attitude of gamblers, we will focus on the odds distribution of the player-selected odds. However, we find that more than 78% of the bets on player-selected odds are placed by the top 40 gamblers, which only account for less than 1% of the players who try player-selected odds. Each of those top 40 gamblers placed at least half a million bets within years, therefore these bets have most likely been placed by programs, which is also a method offered by the website for wagering. As we have pointed out, all the games discussed in this chapter have negative expected payoffs, so these large amounts of bets are not likely to be placed for the purpose of gambling. As already mentioned, we will exclude the bets from these players from our odds distribution analysis.

Although crypto-currency has gained decent popularity in the financial and technological world, in this chapter we still measure the wager/wealth deposited in forms of crypto-currencies in US dollars, since the wagers in skin gambling are measured in US dollars. The historical daily price data of crypto-currencies (Bitcion, Ethereum, Bitcoin Cash) are obtained from CoinDesk [181] (for Bitcoin) and CoinMetrics [182] (for Ethereum and Bitcoin Cash).

5.2.1 Ethics for Data Analysis

The data collected and analyzed in this chapter are all publicly accessible on the internet, and we collect the data either with the consent of the website administrators or without violating the terms of service or acceptance usage listed on the hosting website. The data we use do not include any personally identifiable information (PII), and we further anonymize account-related information before storing them into our databases to preserve players' privacy. In

addition, our data collection and analysis procedures are performed solely passively, with absolutely no interaction with any human subject. To avoid abusing the hosting websites (i.e., the gambling websites), the request rates of data-collecting are limited to 1 request per second. Considering the legal concerns and potential negative effects of online gambling [126, 135, 138, 183, 184, 185, 186, 187], our analysis aims only to help better prevent adolescent gambling and problem gambling.

5.3 Parameter Estimation and Model Selection

In our analysis, the parameters of different distribution models are obtained by applying Maximum Likelihood Estimation (MLE). To select the best-fit distribution, we compare the models' Akaike weights derived from Akaike Information Criterion (AIC). Note that analyzing the fitting results, we constantly found that players show a tendency of using simple numbers when allowed to place wagers with arbitrary amounts of virtual currency. As a result, the curves of probability distribution functions appear to peak at simple numbers, and the corresponding cumulative distribution function shows a stepped behavior. This makes the fitting more difficult, especially for the determination of the start of the tail. To address this issue, we choose the start of the tail x_{min} such that we obtain a small Kolmogorov–Smirnov (K–S) distance between the empirical distribution and the fitting distribution, while maintaining a good absolute fit between the complementary cumulative distribution functions (CCDF) of the empirical distribution and the best-fitted distribution. Candidate models for model selection in this chapter include exponential distribution, power-law distribution, log-normal distribution, power-law distribution with sharp truncation, power-law distribution with exponential cutoff, and pairwise power-law distribution. More details about parameter fitting and model selection can be found in Ch. 3.

5.4 Wager Distribution

From the viewpoint of the interaction among players, the games discussed in this chapter can be grouped into two classes: in Roulette, Crash, and Satoshi Dice games, there is little or no interaction among players, whereas in Jackpot games, players need to gamble against each other.

At the same time, from the viewpoint of wager itself, the games can also be grouped into two classes: In games (A-G), the wagers can be an arbitrary amount of virtual currencies, such as virtual skin tickets or crypto-currency units, whereas in game (H), the wagers are placed in the form of in-game skins, which means the wager distribution further involves the distributions of the market price and availability of the skins.

Furthermore, from the viewpoint of the odds, considering the empirical datasets we have, when analyzing the wager distribution, there are three situations: i) For Roulette and Satoshi Dice games, the odds are fixed constants, and wagers placed with the same odds are analyzed to find the distribution. ii) For Crash games, the odds are selected by the players, and wagers placed with different odds are mixed together during distribution analysis. iii) For the Jackpot game, the odds are not fixed at the moment when the player wagers.

In Table 5.1 we categorize the 8 datasets based on the above information. At the same time, for each dataset we perform a distribution analysis of wagers at the aggregate level. Within the same dataset wagers placed under different maximum allowed bet values are discussed separately. We plot the complementary cumulative distribution function (CCDF) of the empirical data and the fitted distribution to check the goodness-of-fit, see Fig. 5.1. CCDF, sometimes also referred to as the survival function, is given by $\bar{F}(x) = P(X > x) = 1 - P(X \leq x)$.

It turns out that when players are allowed to place arbitrary wagers (games A-G in Ta-

Table 5.1: The best-fitted distribution and estimated parameters of wagers. For games (A, B, C, E, F, G) the best-fitted model is a log-normal distribution, and for game (D) the log-normal distribution is truncated at a maximum value. For game (H) the wager distribution follows a power law - exponential - power law pattern.

Game Name	Game Category	Wager Currency	Arbitrary Bet	Max Bet	Odds	Best-Fitted Model	Parameters
csgofast-Double (A)	Roulette	Virtual Skin Ticket	**Yes**	500,000 (A_1)	2 (Red)	**Log-normal**	$\mu = 3.689$, $\sigma = 1.952$ $x_{min} = 21$
					2 (Black)		$\mu = 3.807$, $\sigma = 1.922$ $x_{min} = 21$
					14 (Green)		$\mu = 3.972$, $\sigma = 1.647$ $x_{min} = 21$
				50,000 (A_2)	2 (Red)		$\mu = 2.936$, $\sigma = 2.108$ $x_{min} = 11$
					2 (Black)		$\mu = 3.175$, $\sigma = 2.118$ $x_{min} = 12$
					14 (Green)		$\mu = 2.633$, $\sigma = 2.113$ $x_{min} = 14$
csgofast-X50 (B)				50,000	2 (Blue)		$\mu = 2.734$, $\sigma = 1.930$ $x_{min} = 11$
					3 (Red)		$\mu = 2.450$, $\sigma = 2.030$ $x_{min} = 12$
					5 (Green)		$\mu = 2.814$, $\sigma = 1.999$ $x_{min} = 12$
					50 (Gold)		$\mu = 3.416$, $\sigma = 1.548$ $x_{min} = 11$
csgofast-Crash (C)	Crash			10,000 (C_1)	Player-Selected		$\mu = 1.647$, $\sigma = 2.226$ $x_{min} = 15$
				20,000 (C_2)			$\mu = 1.932$, $\sigma = 2.143$ $x_{min} = 11$
Ethcrash (D)		Crypto-currency		0.25 ETH			$\mu = -7.186$, $\sigma = 6.356$ $x_{min} = 1$
Satoshi dice (E)	Satoshi Dice			10 BCH	1.98		$\mu = 5.910$, $\sigma = 2.691$ $x_{min} = 34$
Coinroll (F)				3 BTC			$\mu = 1.930$, $\sigma = 2.638$ $x_{min} = 2$
csgospeed (G)	Jackpot	Virtual Skin Ticket		500,000	Not-fixed		$\mu = 5.167$, $\sigma = 1.301$ $x_{min} = 23$
csgofast-jackpot (H)		In-game Skin	**No**	15 items 180,000 per item		Power Law - Exponential - Power Law	$\alpha = 0.802$, $\delta = 2.457 \times 10^2$ $\beta = 7.080 \times 10^{-3}$, $\eta = 3.783$ $\lambda = 8.625 \times 10^{-5}$, $x_{min} = 250$

ble 5.1), the wager distributions can be in general best-fitted by log-normal distributions. In particular, in games (A, B, C, E, F, G), the wager distribution can be approximated by the

following expression

$$P(x) = \frac{\Phi\left(\frac{\ln(x+1)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{\ln(x_{\min})-\mu}{\sigma}\right)}, \tag{5.3}$$

with $x_{\min} \leq x$ and $\sigma > 0$. $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Meanwhile in game (D), the fitted log-normal distribution is truncated at an upper boundary $x_{\max}$, which might result from the small maximum allowed bet value and the huge variation of the market price of crypto-currencies.

During model selection, we notice that when we select different $x_{\min}$, occasionally power-law distribution with exponential cutoff is reported to be a better fit, but often it does not provide a decent absolute fit on the tail, and overall the log-normal distribution provides smaller K-S distances.

On the other hand, as we have pointed out in Ch. 4, when players are restricted to use in-game skins as wagers for gambling, the wager distribution can be best fitted by a shifted power law with exponential cutoff. Now, with a similar situation in game (H), where wagers can only be in-game skins, we find that the early part of the curve can be again fitted by a power law with exponential cutoff, as shown in Fig. 5.1(H). However, this time it does not maintain the exponential decay of its tail; instead, it changes back to a power-law decay. The overall distribution contains six parameters, given by the expression

$$P(x) = \begin{cases} \dfrac{1}{c_1 + c_2 c_3} \dfrac{(x-\delta)^{-\alpha}}{1 + e^{\lambda(x-\beta)}}, & \text{for } x \leq x_{\text{trans}}, \\ \dfrac{c_3}{c_1 + c_2 c_3}\, x^{-\eta}, & \text{for } x > x_{\text{trans}}, \end{cases} \tag{5.4}$$

where $c_1 = \sum_{x=x_{\min}}^{x_{\text{trans}}} \frac{(x-\delta)^{-\alpha}}{1+e^{\lambda(x-\beta)}}$, $c_2 = \zeta\left(\eta, x_{\text{trans}}\right)$, and $c_3 = x_{\text{trans}}^{\eta} \frac{(x_{\text{trans}}-\delta)^{-\alpha}}{1+e^{\lambda(x_{\text{trans}}-\beta)}}$.

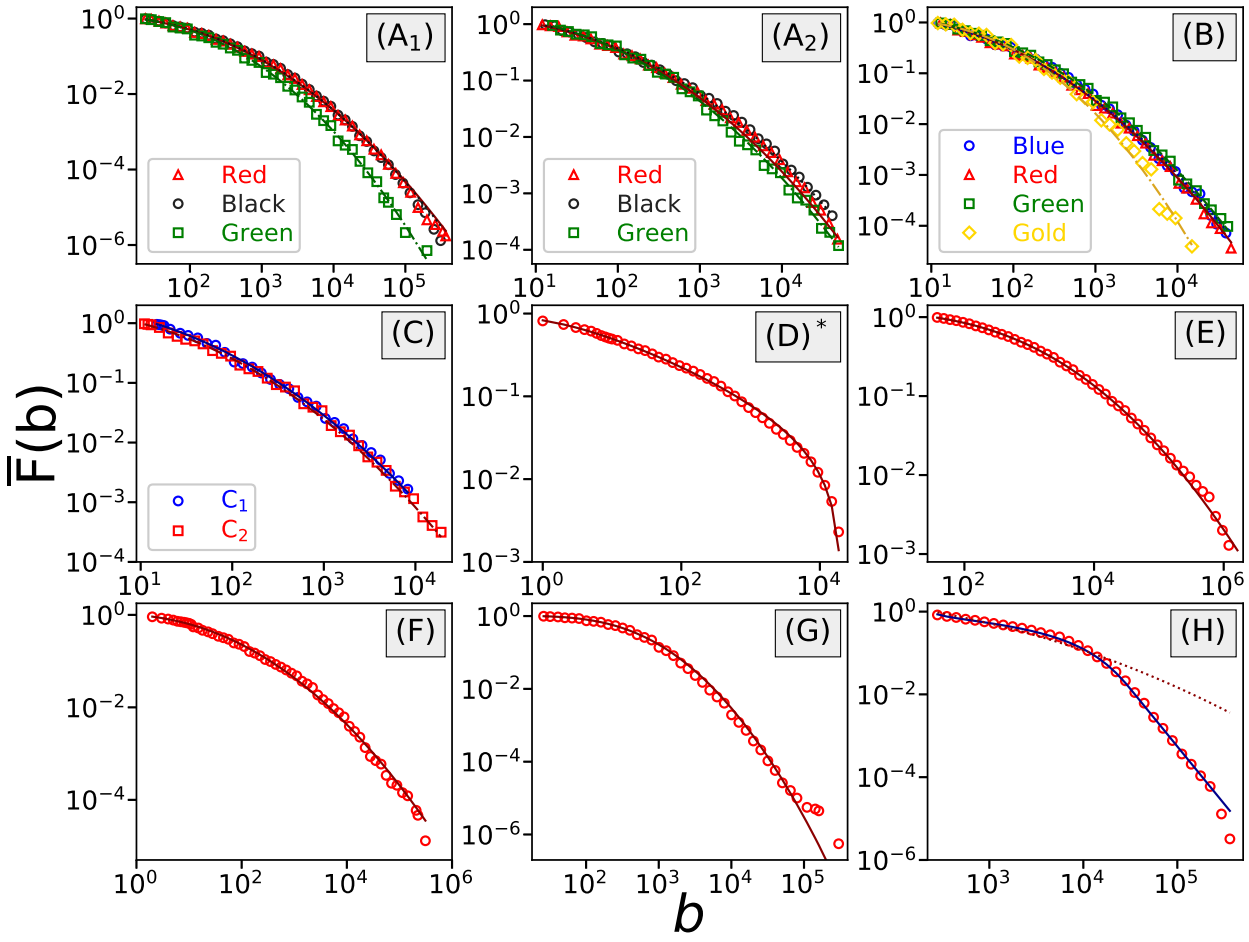

Figure 5.1: In games (A-G), where players are allowed to choose arbitrary bet values, the wager distribution can be best fitted by log-normal distributions (5.3). In game (D), the log-normal distribution is truncated at its maximum bet value, indicated by *. The fitting lines represent the log-normal fittings. Wagers placed under the different maximum allowed bet values are discussed separately, e.g., in game (A), (A_1) has a maximum bet value of $500,000$, and (A_2) has a maximum bet value of $50,000$. On the other hand, in game (H) where wagers can only be in-game skins, the wager distribution can be captured by a shape in a pairwise power law with an exponential transition, see Eq. (5.4). The red dotted line represents the log-normal fitting and the blue solid line represents the fitting of a pairwise power law with an exponential transition.

We believe that when players are restricted to use in-game skins as wagers, the decision to include one particular skin in their wager is further influenced by the price and availability of that skin. These factors make the wager distribution deviate from the log-normal distribution, which is observed in games (A-G). This is very clear when comparing the wager distributions of games (G) and (H) as both games are jackpot games of skin gambling, and the only difference is whether players are directly using skins as wagers or are using virtual skin tickets obtained from depositing skins. The power-law tail, which was not observed in the previous chapter, might result from the increment of the maximum allowed skin price (from \$400 to \$1800).

The above discussions, including the results for games (A-G) in Table 5.1, show that the wager distributions in pure probability-based gambling games, no matter whether the game follows parimutuel betting or fixed-odds (preset/player-selected) betting, stay log-normal as long as the players are allowed to place arbitrary amounts of wagers. This commonality of log-normal distribution no longer holds when this arbitrariness of wager value is violated, e.g., in the scenario where the player can only wager items (in-game skins).

Log-normal distribution has been reported in a wide range of economic, biological, and sociological systems [73], including income, species abundance, family size, etc. Economists have proposed different kinds of generative mechanisms for log-normal distributions (and power-law distributions as well). One particular interest for us is the multiplicative process [74, 188]. Starting from an initial value X_0, random variables in a multiplicative process follow an iterative formula

$$X_{i+1} = \exp(\nu_i) X_i \quad \text{or} \quad \ln X_{i+1} = \ln X_i + \nu_i. \tag{5.5}$$

If the v_i has finite mean and variance, and is independent and identically distributed, then

according to the central limit theorem, for large i, $\ln X_i$ will follow a normal distribution, which equivalently means X_i will follow a log-normal distribution.

If we want to check whether gamblers follow multiplicative processes when they wager, we can first check the correlation between consecutive bets (b_i, b_{i+1}). Due to the large variances of the wager distributions, Pearson's correlation coefficient may perform poorly. Instead, we adopt two rank-based correlation coefficients, Kendall's Tau τ_K and Spearman's Rho ρ_S. At the same time, we also check the mean and variance of the log-ratios $\ln(b_{i+1}/bi)$ between consecutive bets. These statistics can be found in Table 5.2. The results reveal that the values of consecutive bets exhibit a strong positive correlation, with all the correlation coefficients larger than 0.5. It shows that players' next bet values are largely dependent on their previous bet values. At the same time, the bet values are following gradual changes, rather than rapid changes. These conclusions can be confirmed by the small mean values and small variances of log-ratios between consecutive bets.

Table 5.2: Correlation analysis shows that there is a strong positive correlation between consecutive bets, along with the small mean values and variances of log-ratio between consecutive bets. Satoshi Dice (E) is excluded here as individual gamblers in the dataset are not distinguishable. csgofast-Jackpot (H) is excluded in the calculation of $P(b_i = b_{i+1})$ due to the low precision of bet values in this dataset.

Dataset	$\tau_K(b_i, b_{i+1})$	$\rho_s(b_i, b_{i+1})$	$\langle \log_{10}(b_{i+1}/b_i) \rangle$	$\mathrm{var}(\log_{10}(b_{i+1}/b_i))$	$P(b_i = b_{i+1})$
csgofast-Double (A)	0.596	0.737	0.010	0.183	0.342
csgofast-X50 (B)	0.692	0.803	0.007	0.102	0.512
csgofast-Crash (C)	0.858	0.909	0.004	0.038	0.802
ethCrash (D)	0.866	0.949	0.000	0.147	0.549
Coinroll (F)	0.826	0.925	0.000	0.282	0.497
csgospeed-Jackpot (G)	0.522	0.675	0.002	0.288	0.136
csgofast-Jackpot (H)	0.591	0.759	0.002	0.206	–

Further analysis of the distribution of ν shows an exponential decay on both of its tails, see Fig. 5.2. This means that ν approximately follows a Laplace distribution. However, compared to a Laplace distribution, the empirical log-ratio distribution shows a much higher probability at $\nu = 0$, whose value can be found in the last column of Table 5.2. We also

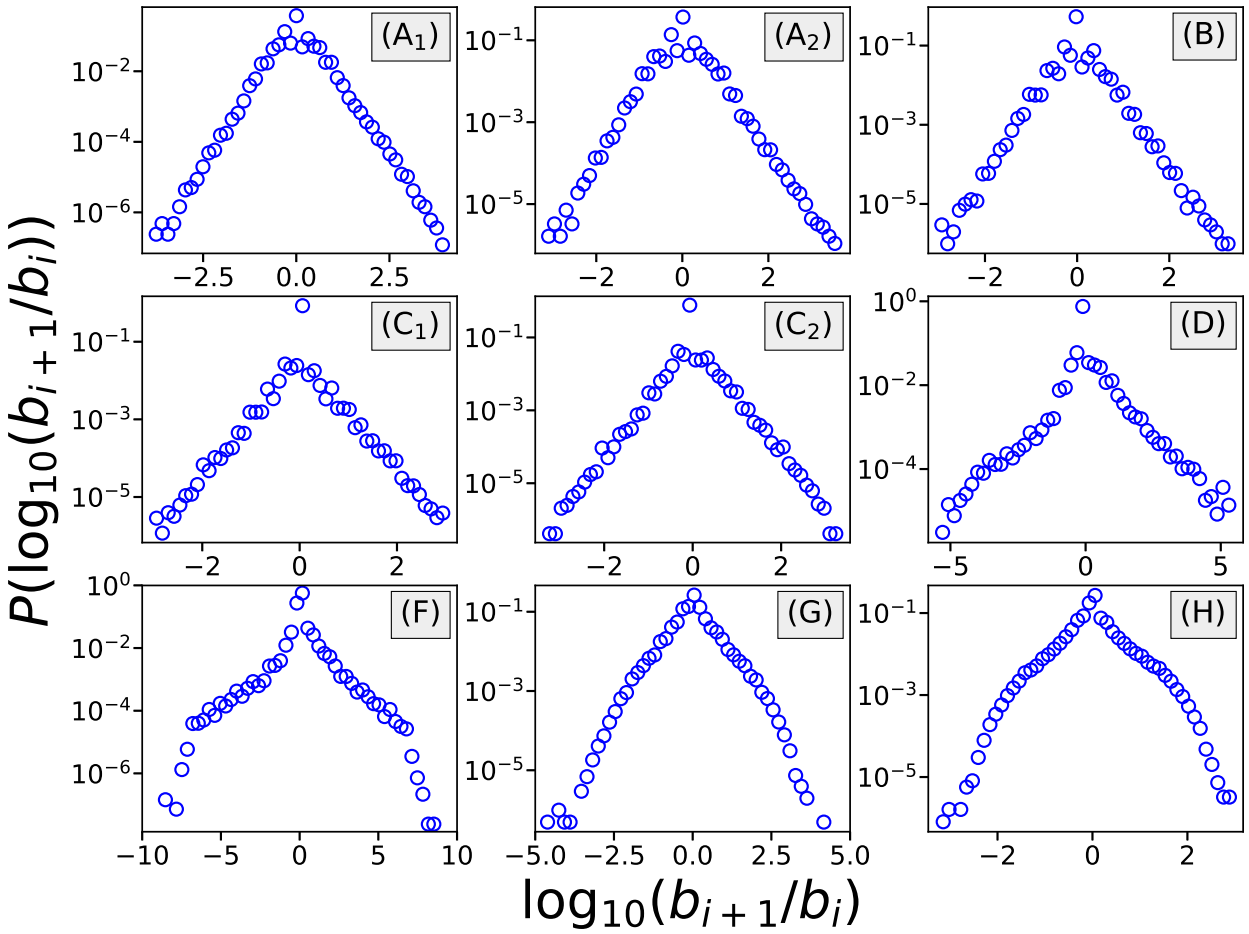

Figure 5.2: The distribution of the logarithmic of the ratio (log-ratio) between consecutive bet values. For games (A, B, C), the log-ratio can be described by a Laplace distribution. For games (D, F, G, H), the log-ratio presents bell-shaped distribution. In general, the distributions are symmetric with respect to the y-axis, except in games (D) and (F).

observe that ν presents higher probability densities around small integers/half-integers and their inverses. Due to the existence of these differences, we will skip the parameter fitting for the distribution of ν. The high probability of staying on the same wager is therefore one of the common strategies adopted by gamblers in fixed-wager betting.

Meanwhile, the high positive auto-correlations, along with the higher probability densities at small integers/half-integers and their inverses, provide evidence that gamblers also often follow a multiplicative process when wagering. The multiplication process can be explained

by the wide adoption of multiplicative betting systems. "Betting system" here refers to the strategy of wagering where the following bet value depends on the previous bet value and previous outcome [189]. Although betting systems will not provide a long-term benefit, as the expected payoff will always be 0 in a fair game, still they are widely adopted among gamblers. The most well-known multiplicative betting system is the Martingale (sometimes called as geometric progression) [189]. In Martingale betting, starting with an initial wager, the gambler will double their wager each time they lose one round, and return to the initial wager once they win. Martingale is a negative-progression betting system, meaning that the gambler will increase their wager when they lose, and/or decrease their wager when they win, which is opposite to a positive-progression betting system.

Apart from multiplicative betting, there are many other types of betting systems, such as additive betting and linear betting [189]. The reasons why multiplicative betting systems are dominant in our datasets are: 1) Martingale is a well-known betting system among gamblers; 2) Many online gambling websites provide a service for changing the bet value in a multiplicative way. For example, for the Crash games csgofast-Crash (C) and ethCrash (D), both websites provide a simple program for automatically wagering in a multiplicative way. For the Roulette games and Coinroll (F), the websites provide an interface with which the gambler can quickly double or half their wager. However, for Satoshi Dice (E) and csgospeed-Jackpot (G), no such function is provided, yet we still observe similar results, indicating that gamblers will follow a multiplicative betting themselves.

Fig. 5.2 provides us with the distribution of ν, however, it will not tell us whether the gamblers adopt the negative/positive-progression betting systems. Therefore we further analyze the effect on the bet values of winning/losing a round. How the gamblers adjust their wager after winning/losing rounds is shown in Table 5.3. We can see that although there is a high probability for sticking to the same bet values, the most likely outcome after

losing a round is that the gambler increases their wager. When winning one round, gamblers are more likely to decrease their wager. This means that negative-progression strategies are more common among gamblers than positive-progression strategies.

Table 5.3: Statistics about how gamblers change their bet values after winning/losing rounds. Apart from fixed-wagering betting, a comparison between the probabilities suggests gamblers prefer negative-progression betting rather than positive-progression betting. Satoshi Dice (E) and csgofast-Jackpot(H) are excluded from the analysis due to the reasons mentioned in the caption of Table 5.2.

Dataset	After Losing			After Winning		
	$P(b_{i+1} > b_i)$	$P(b_{i+1} = b_i)$	$P(b_{i+1} < b_i)$	$P(b_{i+1} > b_i)$	$P(b_{i+1} = b_i)$	$P(b_{i+1} < b_i)$
csgofast-Double (A)	0.432	0.319	0.249	0.228	0.383	0.388
csgofast-X50 (B)	0.293	0.500	0.207	0.167	0.541	0.292
csgofast-Crash (C)	0.201	0.685	0.114	0.076	0.854	0.069
ethCrash (D)	0.566	0.401	0.033	0.079	0.690	0.231
Coinroll (F)	0.560	0.377	0.061	0.121	0.606	0.274
csgospeed-Jackpot (G)	0.478	0.159	0.374	0.415	0.104	0.480

5.5 Risk Attitude

We now turn to the following question: When a player is allowed to choose the odds themselves in a near-fair game, how would they balance the risk and potential return? Higher odds means a lower chance of winning and higher potential return, for example, setting odds of 10 means that the winning chance is only 1/10, but the potential winning payoff equals 9 times the original wager. In our analysis, we can examine such behaviors based on the gambling logs from Crash and Satoshi Dice games. For the Crash game, only CSGOFAST.COM provides the player-selected odds even when players lose that round, therefore we can analyze the distribution of player-selected odds (from those rounds where every player with player-selected odds lost). For the Satoshi Dice game, only Coinroll accepts player-selected odds. We will focus on the data collected on these two websites. For the Crash game on CSGOFAST.COM, the odds can only be set as multiples of 0.01, whereas for the Satoshi

Dice game on Coinroll, the odds can be set to $0.99 \cdot 65536/i$ where i is a positive integer less than 64000. To simplify our modeling work, we will convert the odds on Coinroll to be multiples of 0.01 (same as the Crash game).

It turns out that in both cases the odds can be modeled with a truncated shifted power-law distribution, which has the expression

$$P(m) = \begin{cases} \dfrac{(\ m-\delta)^{-\alpha}}{\zeta(\alpha,\ m_{\min}-\delta)}, & \text{for } m_{\min} \le m < m_{\max}, \\ \dfrac{\zeta(\alpha,\ m_{\max}-\delta)}{\zeta(\alpha,\ m_{\min}-\delta)}, & \text{for } m = m_{\max}, \end{cases} \tag{5.6}$$

where $\zeta(\cdot,\cdot)$ is the incomplete Zeta function, and $m_{\max}$ is the upper truncation. Note that there is a jump at $m_{\max}$, meaning that the players are more likely to place bets on the maximum allowed odds than on a slightly smaller odds. The estimated parameters can be found in Table 5.4. From the comparison between the CCDFs of empirical data and fitting curves, as shown in Fig. 5.3, we can see that the truncated shifted power law can capture the overall decaying trends of odds distribution. Some noticeable deviations, including the stepped behavior, result from the gamblers' preference of simple numbers.

Table 5.4: The odds distribution can be best fitted by a truncated shifted power-law distribution, and the exponents are both smaller than 2.

Game Name	Game Category	Wager Currency	Best-Fitted Model	Parameters
csgofast-Jackpot	Crash	Virtual Skin Ticket	Truncated Shifted Power Law	$\alpha = \mathbf{1.881}$, $\delta = 0.849$ $m_{\min} = 1.15$
Coinroll	Satoshi Dice	Bitcoin		$\alpha = \mathbf{1.423}$, $\delta = 2.217$ $m_{\min} = 2.58$

A distribution that is close to a power law indicates that a gambler's free choice on odds shows scaling characteristics (within the allowed range) in near-fair games. It also means that when gamblers are free to determine the risks of their games, although in most times they will stick

to low risks, showing a risk-aversion attitude, they still present a non-negligible probability of accepting high risks in exchange for high potential returns. The scaling properties of risk attitude might not be unique to gamblers, but also may help to explain some of the risk-seeking behaviors in stock markets or financial trading.

Now we re-examine the distributions from the point of view of estimating the crash point m_C (Satoshi Dice games can be explained with the same mechanism). The true distribution of m_C generated by the websites follow a power-law decay with an exponent of 2 (with some small deviation due to the house edge). Meanwhile, a closer look at the fitted exponents listed in Table 5.4 gives us two empirical exponents of 1.423 and 1.881, both of which are smaller than 2. The smaller exponents reveal that gamblers believe that they have a larger chance to win a high-odds game than they actually do. Or equivalently, it means the gamblers over-weight the winning chance of low-probability games. At the same time, the "shifted" characteristics here lead to more bets on small odds, which also indicates that the gamblers over-estimate the winning chance of high-probability games. As a result, they under-weight the winning chances of mild-probability games. These are clear empirical evidence of probability weighting among gamblers, which is believed to be one of the fundamental mechanisms in economics [167].

5.6 Wealth Distribution

In the previous study of skin gambling [125], we pointed out that the wealth distribution of skin gamblers shows a pairwise power-law tail. This time, by considering the players' deposits to the gambling site as the wealth data, we find that the pairwise power-law tails are also observed for bitcoin gambling. We find that on the gambling website Coinroll, starting from 5660 cents, the players' wealth distribution follows a pairwise power-law distribution, with

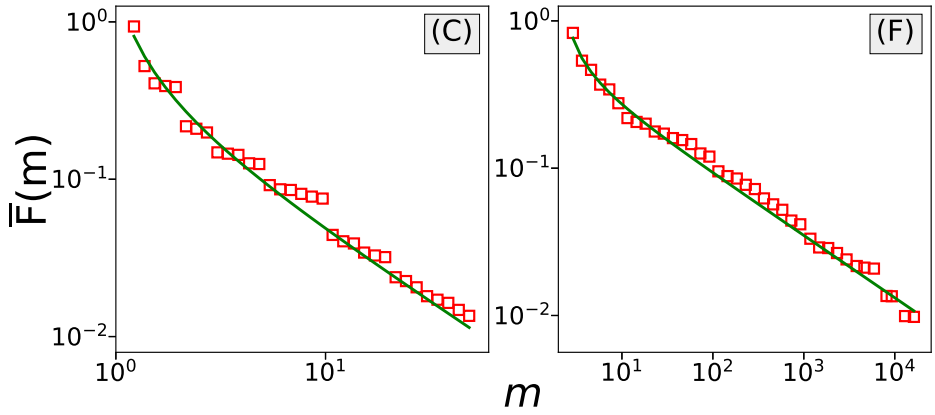

Figure 5.3: Odds distributions can be well-fitted by truncated shifted power-law distributions.

the power of the first regime to be 1.585, and the power of the second regime to be 3.258, see Fig. 5.4. The crossover happens at 1.221×10^5 cents. As both wealth distributions of skin gambling and bitcoin gambling can be approximated by a pairwise power distribution, we believe that it is a good option for modeling the tails of gambler wealth distribution in different scenarios.

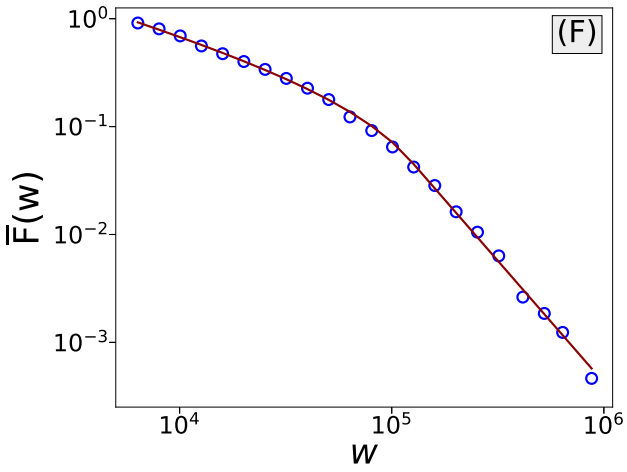

Figure 5.4: The tail of the wealth distribution of Bitcoin gamblers follows a pairwise power-law distribution.

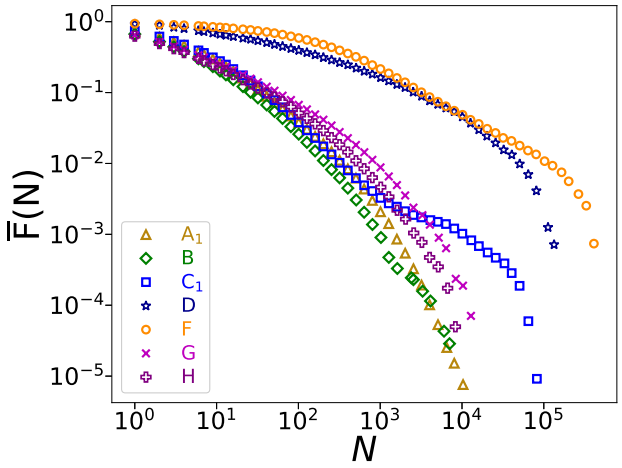

Figure 5.5: In all datasets, the distributions of the number of bets placed by individual players present heavy-tailed properties. Closer inspection also shows that the cryto-currency gamblers, i.e., in games (D) and (F), tend to place more bets (heavier tails) when comparing to gamblers in skin gambling.

5.7 Removing the Effects of Inequality of the Number of Bets

In the above sections, we have analyzed the distributions of several quantities at the population level. However, as shown in Fig. 5.5, there is a huge inequality of the number of placed bets among gamblers. We therefore wonder whether those distributions we obtain result from the inequality of number of bets among individuals. To remove the effects of this inequality, we randomly sample in each dataset the same number of bets from heavy gamblers. We re-analyze the wager distribution and odds distribution with the sample data to see if we obtain the same distribution as before. In each dataset we randomly sample 500 bets from each of those gamblers who placed at least 500 bets above $b_{\min}$ given in Table 5.1. Some datasets are excluded here as either they don't have enough data or we cannot identify individual gamblers. When re-analyzing the odds distribution, to ensure we have enough data, we respectively sample 100 and 2000 bets from each of those gamblers in games (C) and

(F) who have at least 100 and 2000 valid player-selected odds above $m_{\min}$ given in Table 5.4. According to the results in Fig. 5.6, the wager distributions after removing the inequality can still be approximated by log-normal distributions, but some deviation can be observed. Similarly, the odds distributions after removing the inequality, as shown in Fig. 5.7, again follow truncated shifted power-law distributions. These results demonstrate that the shape of the distributions we obtained in the above sections is not a result of the inequality of the number of bets.

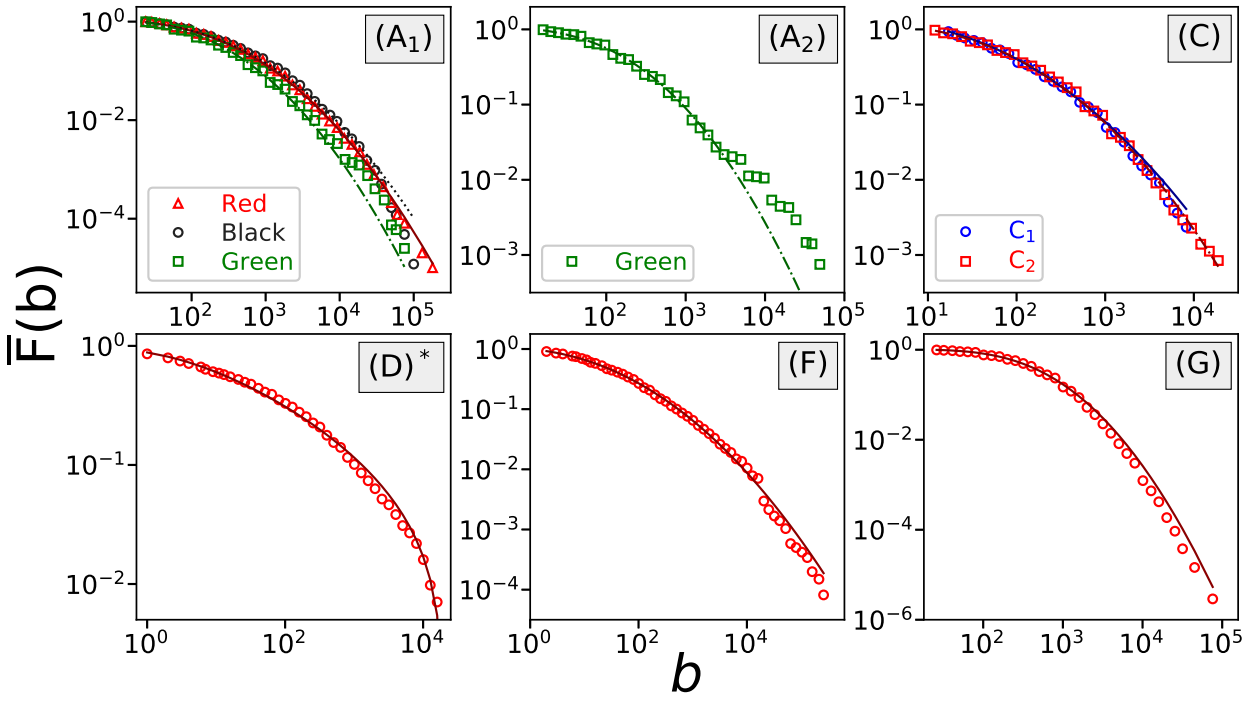

Figure 5.6: The wagers obtained from random sampling of top gamblers' bets still present log-normal distributions, although there are some observable deviations.

Now our question becomes whether the conclusion of the distribution at the population level can be extended to the individual level. Here due to the limitation of data, we will only discuss the wager distribution. Analyzing the individual distribution of top gamblers, we find that although heavy-tailed properties can be widely observed at the individual level, only a small proportion of top gamblers presents log-normal distributed wagers. In Fig. 5.8, we show

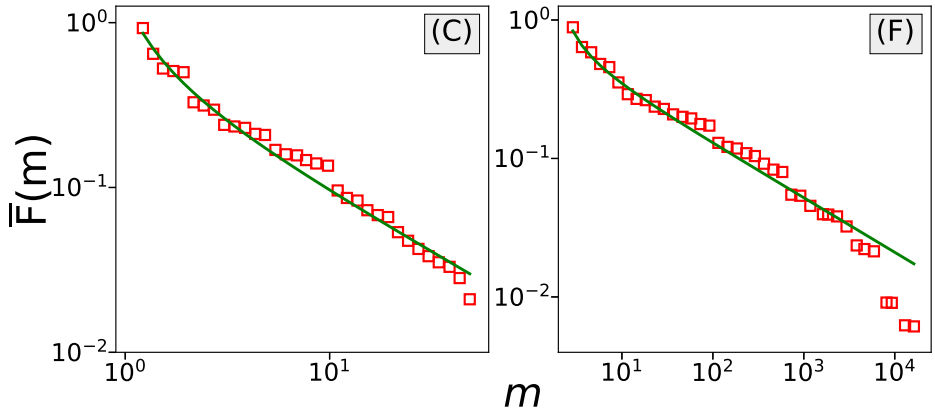

Figure 5.7: The player-selected odds obtained from the random samples of top gamblers' selected odds still present truncated shifted power-law distributions, with observable deviation at the far tails of the distributions. The exponents are respectively 1.712 and 1.394 for games (C) and (F).

the wager distribution of several typical gamblers whose wager distributions respectively follow a log-normal distribution, a power-law distribution, a power-law distribution with exponential cutoff, a pair-wise power-law distribution, an irregular heavy-tailed distribution, and one that only has a few values. The diversity of the wager distribution at the individual level suggests a diversity of individual betting strategies. Also, it indicates that a gambler may not stick to only one betting strategy. It follows that the log-normal wager distribution observed at the population level is very likely to be an aggregate result.

5.8 Diffusive process

For an individual player's gambling sequence we define "time" t as the number of bets one player has placed so far, and define as net income the sum of the payoffs of those bets. In all the games we analyze, there are only two possible outcomes: a win or a loss. The player's net income will change each time they place a bet in a round, with the step length to be the payoff from that bet. We can treat the change of one player's net income as a random walk

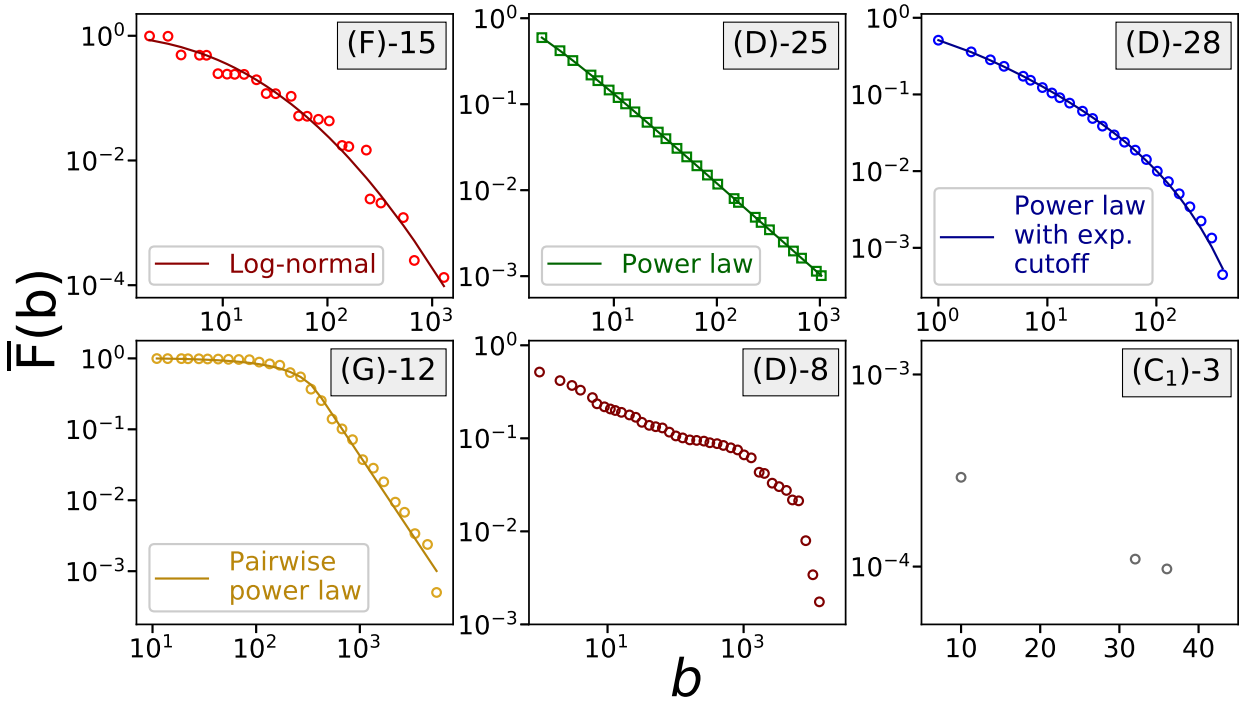

Figure 5.8: Although heavy-tailed properties can be commonly observed in wager distributions, the log-normal distribution is not universal at the individual level. In the figure, we show 6 typical gamblers whose wager distributions respectively follow a log-normal distribution: (F)-15; a power-law distribution: (D)-25; a power-law distribution with exponential cutoff: (D)-28; a pair-wise power-law distribution: (G)-12; an irregular heavy-tailed distribution: (D)-8; and one that only has a few values: (C_1)-3. The figure label indicates the index of the gambler based on the dataset and on the number of bets they placed, e.g., (F)-15 means the gambler placed 15th most bets in dataset (F).

in a one-dimensional space. The time t will increase by 1 when the player places a new bet, therefore the process is a discrete-time random walk. Now, let us focus on the analysis of the diffusive process of the gamblers' net incomes, starting with the analysis of the growth of the mean-squared displacement (MSD), defined as

$$\mathrm{MSD}(t) = \left\langle (\Delta w(t))^2 \right\rangle = \left\langle \left(\sum_{i=1}^{t} o_p(i) \right)^2 \right\rangle, \tag{5.7}$$

where $\Delta w(t)$ is the cumulative change of a player's wealth after they attend t rounds, and $o_p(i)$ represents the payoff from the i_{th} round the player attended. $\langle \cdot \rangle$ represents an ensemble

average over the population, i.e., over all the players who placed bets. For a normal diffusive process, $MSD(t) \sim t$, otherwise it means the random walk follows an anomalous diffusive behavior. More specifically, when the MSD growth is faster (respectively, slower) than linear, we call it superdiffusion (respectively, subdiffusion).

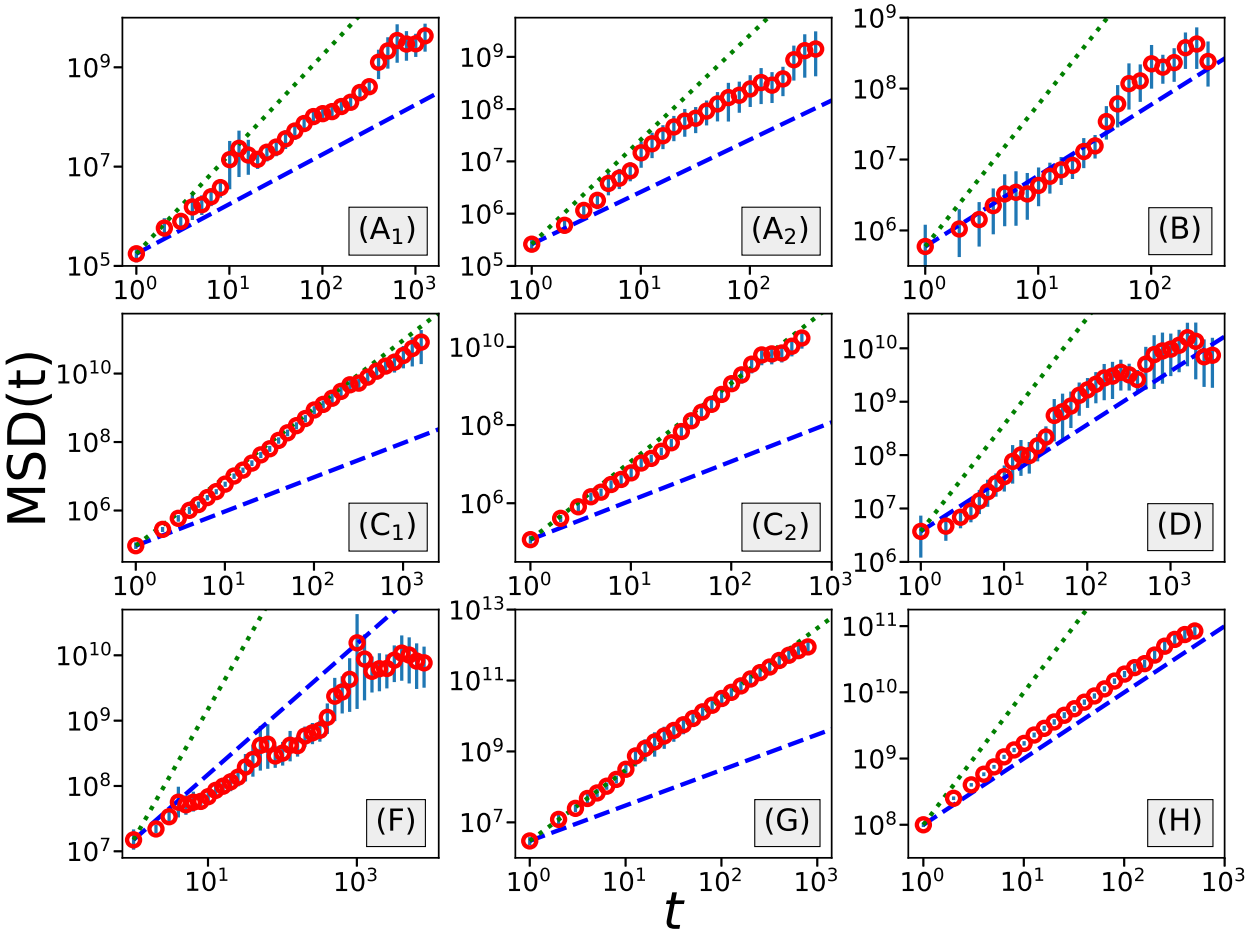

Figure 5.9: The growth of mean-squared displacement in different datasets presents different diffusive behaviors. In the figures, the error bars represent 95% confidence intervals, blue dashed lines follow linear functions (slope = 1), and green dotted lines follow quadratic functions (slope = 2).

In Fig. 5.9, we present the growth of the mean-squared displacement against time for each of the datasets. To reduce the coarseness, MSD curves are smoothed with log-binning technique. The error bars in Fig. 5.9 represent 90% confidence intervals computed with boot-

strapping using 2000 independent re-sampling runs. It is interesting to see that for different datasets we observe different diffusive behaviors. For games csgofast-Crash (C) and csgospeed (G) we observe that the MSD grows approximately as a square law with time, suggesting superdiffusive behavior. Meanwhile, for games csgofast-Double (A), ethCrash (D), and csgofast-Jackpot (H), the MSD first presents a superdiffusive regime, followed by a crossover to a normal diffusive regime. For games csgofast-X50 (B) and Coinroll (F), although the MSD roughly present a linear and a sublinear growth, a careful inspection suggests that both curves consist of several convex-shaped regimes, indicating a more complex behavior.

In Ch. 4, we argued that the crossover from a superdiffusive regime to a normal diffusive regime in a parimutuel game is due to the limitation of individuals' wealth and the conservation of total wealth. A similar crossover is observed in the same game (H), a parimutuel betting game using skins as wagers, where the same explanation can be applied. However, the crossover is not seen in the other parimutuel betting game (G). On the other hand, this crossover is found in a roulette game and in a Crash game, where there is no interaction among gamblers. The limitation of an individual's wealth can still be a partial explanation, but the conservation of total wealth no longer holds. A different explanation needs to be proposed to model this crossover.

In general, the diverse diffusive behaviors found in different datasets indicate that human gambling behavior is more complex than random betting and simple betting systems. Further studies of gambling behaviors are required in order to fully understand the differences.

5.9 Modeling

In this section, we want to discuss how we can obtain from different gambling models the different diffusive processes observed in the previous section. We will not attempt to reproduce the parameters we obtained from the gambling logs, but rather try to explore the possible reasons for the anomalous diffusion we reported.

For a gambling process, if the gambler's behavior is independent among different rounds, i.e., the wager and odds are respectively independent and identically distributed (IID), with no influence from the previous outcomes, and if the wager b has finite variance and the odds m has finite mean, then MSD's growth will be a linear function of time t.

$$\mathrm{MSD}(t) = (\langle m \rangle - 1) \left\langle b^2 \right\rangle t, \tag{5.8}$$

where $\langle m \rangle$ is the mean value of odds distribution and $\langle b^2 \rangle$ is the second moment of the wager distribution. But normal diffusion is only found in few datasets, the remaining datasets presenting anomalous diffusion which conflicts with the IID assumption.

Having shown the popularity of betting systems among gamblers, we would like to check how different betting systems affect diffusive behaviors. First, we simulate gamblers that follow Martingale strategies in a Crash game. We assume that the selection of odds follows a power-law distribution with an exponent α, with a minimum odds of 1 and a maximum odds of 50, where the maximum odds is set to ensure a finite variance of the odds distribution. Starting from a minimum bet of 1, we multiply wagers by a ratio γ each time the gamblers lose one round and return to the minimum bet each time they win. Once the wager reaches a preset maximum bet value 10000, we reset the gambler with a minimum bet. MSDs obtained from 10 billion individual simulations are shown in Fig. 5.10. Different curves correspond to different exponents in odds distribution. We can see that the MSD initially

presents an exponential-like growth, before the growths reduce to a linear function. It is easy to explain the exponential growth since many gamblers lose the rounds and therefore increase their wager by the factor γ, which leads to an increase in the average bet value. The superdiffusion here suggests that Martingale strategy increases gamblers' risks of huge losses. Considering the wide adoption of Martingale among gamblers, this could be a reason for the superdiffusion as well as the crossover to normal diffusion we found in several datasets. Comparison of the MSD curves of different α suggests that a more aggressive risk attitude leads to a higher risk of huge losses (as well as higher potential winnings).

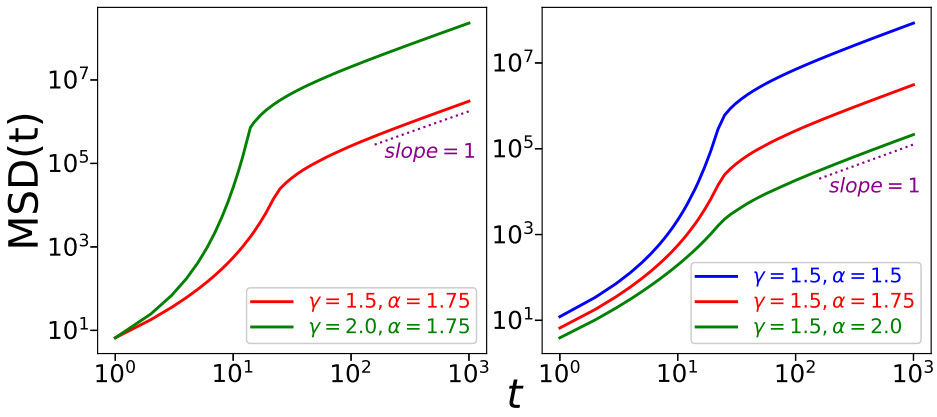

Figure 5.10: A betting system similar to Martingale will lead to a crossover from superdiffusion to normal diffusion according to the growth of mean-squared displacement. Comparison between curves of different parameters shows that higher γ and lower α both will lead to a higher chance of huge losses/winnings.

However, as has been pointed out by Meng [168], gamblers show a huge diversity of betting strategies, and even individual gamblers constantly change their betting strategy. Differences in the fractions of gamblers playing in different betting strategies could be a reason why we see the different diffusive behaviors in different datasets.

5.10 Discussion

In this chapter, we have shown that log-normal distributions can be widely used to describe the wager distributions of online gamblers at the aggregate level. Also the risk attitude of online gamblers shows scaling properties, which indicates that although most gamblers are risk-averse, they sometime will take large risks in exchange for high potential gains.

Our analysis of the growth of the mean-squared displacement of net income shows different diffusive behaviors in different datasets, indicating that the gamblers' behaviors are more complex than simple betting systems. We propose some explanation for the anomalous diffusion we found, but more work is required to fully understand the gamblers' behaviors.

Chapter 6

Conclusions

I have presented three projects in this dissertation studying the behavior patterns in certain human online activities: the first one focuses on information foraging on search engines, and the other two discuss online gambling behaviors. All of these projects provide evidence for the wide existence of heavy-tailed properties in human online behaviors, such as when describing jump lengths, waiting times, gambling wagers, risk attitudes, etc. Detailed distributions for each quantity are also provided. Using the analogy of viewing the behavior dynamics as random walk processes, I reported that those activities all present anomalous diffusion. With further discussions about correlations and entropy, I showed that the quantities are not independent, and the systems are not at stable states. These investigations can help to infer insights about the commonalities among human online dynamics.

For the analysis of information foraging patterns, we found a switch from Lévy-flight-like superdiffusion to Brownian-motion-like normal diffusion when the search engine ranking algorithm was improved. Further analysis shows that when the resource is sparse, the foraging processes are a combination of local searches and relocation phases that are power-law distributed. Our investigation therefore highlights the presence of intermittent search processes in online searches, where phases of local explorations are separated by power-law distributed relocation jumps. These results in general provide a better understanding of how humans perform searching tasks, and search engine providers can benefit from our findings to improve their services, especially when designing the layouts of result pages. On the other hand, the

switch between the diffusive patterns can be useful when designing an adaptive searching algorithm which takes into account of the resource distribution so that it can maintain its efficiency no matter the resource is abundant or sparse.

For the analysis of gambling behaviors, we provided a quantitative description of the wagering patterns of online gamblers. In general, when gamblers are allowed to place arbitrary amounts of wagers, the bet values at the aggregate level approximately follow a log-normal distribution; and when gamblers can only wager items, the wager distribution is affected by the market price and availability of those items. The analysis of individual wager distributions also reported heavy-tailed characteristics, however the diversity of the wager distribution at the individual level suggests a diversity of individual betting strategies. On the other hand, by analyzing the growth of mean-squared displacement, we showed that different games present different diffusive behaviors. In particular, in an online lottery game (Jackpot) that uses in-game skins as wagers, the mean-squared displacement presents a crossover from a superdiffusive to a normal diffusive regime, which is reproduced using simulations and explained analytically. These results can be used to propose better models of describing how and why humans gamble, which may help to prevent problem gambling and adolescent gambling behaviors. In addition, due to the similarities between the gambling and economic behaviors, the scaling characteristics and probability re-weighting in risk attitude we found among online gamblers may help to interpret similar behaviors in economic systems.

In my dissertation, although I only studied certain types of online activities, our findings in those projects reveal very complex humans behaviors. In Chapter 1, I have addressed the importance of studying human behavior patterns to sociophysics research. Clearly, to better understand human behaviors we need to put more attention into the research of human dynamics. And I hope that my work can provide inspirations for others that are interested in this field.

Bibliography

[1] C. Castellano, S. Fortunato, and V. Loreto, Reviews of Modern Physics **81**, 591 (2009).

[2] F. Schweitzer, Physics Today **71**, 40 (2018).

[3] S. Galam, *Sociophysics: A Physicist's Modeling of Psycho-Political Phenomena* (Springer-Verlag New York, 2012).

[4] D. Stauffer, Journal of Statistical Physics **151**, 9 (2013).

[5] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, 2009).

[6] D. Simchi-Levi, Manufacturing & Service Operations Management **16**, 2 (2014).

[7] W. Shih and S. Chai, Academy of Management Proceedings **2016**, 14843 (2016).

[8] J. Fan, F. Han, and H. Liu, National Science Review **1**, 293 (2014).

[9] S. Leonelli, Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences **43**, 1 (2012).

[10] J. C. Molloy, PLOS Biology **9**, 1 (2011).

[11] M. Gurstein, First Monday **16** (2011), 10.5210/fm.v16i2.3316.

[12] H. A. Piwowar and T. J. Vision, PeerJ **1**, e175 (2013).

[13] L. Quetelet, *Sur l'homme et le développement de ses facultés ou Essai de physique sociale*, Sur l'homme et le développement de ses facultés ou Essai de physique sociale No. v. 1 (Bachelier, 1835).

[14] H. C. Carey, *Principles of Social Science*, Vol. 1 (JB Lippincott & Company, 1858).

[15] H. C. Carey, *Principles of Social Science*, Vol. 2 (JB Lippincott & Company, 1863).

[16] H. C. Carey, *Principles of Social Science*, Vol. 3 (JB Lippincott & Company, 1867).

[17] V. Pareto, *Cours d'économie Politique: Professé à l'Université de Lausanne*, Cours d'économie politique: professé à l'Université de Lausanne No. v. 2 (F. Rouge, 1897).

[18] J. Q. Stewart, Sociometry **11**, 31 (1948).

[19] W. J. Reilly, *The Law of Retail Gravitation* (WJ Reilly, 1931).

[20] J. Q. Stewart, Impact of Science on Society **3**, 110 (1952).

[21] J. Q. Stewart, Science **106**, 179 (1947).

[22] J. Q. Stewart, Scientific American **178**, 20 (1948).

[23] J. Q. Stewart, American Journal of Physics **18**, 239 (1950).

[24] S. M. Ulam, *A Collection of Mathematical Problems*, Vol. 8 (Interscience Publishers, New York, 1960).

[25] R. Albert and A.-L. Barabási, Reviews of Modern Physics **74**, 47 (2002).

[26] M. Newman, SIAM Review **45**, 167 (2003).

[27] M. Eigen and P. Schuster, *The Hypercycle: A Principle of Natural Self-Organization* (Springer-Verlag Berlin Heidelberg, New York, 1979).

[28] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. Newman, *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* (Oxford University Press, Inc., Oxford, UK, 1992).

[29] J. G. Oliveira and A.-L. Barabási, Nature **437**, 1251 (2005).

[30] T. Zhou, H. A.-T. Kiet, B. J. Kim, B.-H. Wang, and P. Holme, Europhysics Letters **82**, 28002 (2008).

[31] J. A. Hołyst, K. Kacperski, and F. Schweitzer, in *Annual Reviews Of Computational PhysicsIX* (World Scientific, 2001) pp. 253–273.

[32] R. Hegselmann, U. Krause, *et al.*, Journal of artificial societies and social simulation **5** (2002).

[33] V. Loreto, A. Baronchelli, A. Mukherjee, A. Puglisi, and F. Tria, Journal of Statistical Mechanics: Theory and Experiment **2011**, P04006 (2011).

[34] V. Loreto and L. Steels, Nature Physics **3**, 758 (2007).

[35] B. S. Kerner, *The Physics of Traffic: Empirical Freeway Pattern Features, Engineering Applications, and Theory* (Springer-Verlag Berlin Heidelberg, New York, 2004).

[36] S. Maerivoet and B. D. Moor, Physics Reports **419**, 1 (2005).

[37] A.-L. Barabási, Nature **435**, 207 (2005).

[38] A. Vázquez, Physical Review Letters **95**, 248701 (2005).

[39] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, Proceedings of the National Academy of Sciences **106**, 12640 (2009).

[40] J. Leskovec and E. Horvitz, in *Proceedings of the 17th International Conference on World Wide Web*, WWW '08 (ACM, New York, 2008) pp. 915–924.

[41] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási, Physical Review E **73**, 066132 (2006).

[42] S. K. Baek, T. Y. Kim, and B. J. Kim, Physica A **387**, 3660 (2008).

[43] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási, Physical Review E **73**, 036127 (2006).

[44] A. Vázquez, Physica A **373**, 747 (2007).

[45] C. A. Hidalgo R, Physica A **369**, 877 (2006).

[46] J. G. Oliveira and A. Vázquez, Physica A **388**, 187 (2009).

[47] Y. Wu, C. Zhou, J. Xiao, J. Kurths, and H. J. Schellnhuber, Proceedings of the National Academy of Sciences **107**, 18803 (2010).

[48] R. D. Malmgren, D. B. Stouffer, A. S. L. O. Campanharo, and L. A. N. Amaral, Science **325**, 1696 (2009).

[49] D. B. Stouffer, R. D. Malmgren, and L. A. Amaral, arXiv preprint physics/0605027 (2006).

[50] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral, Proceedings of the National Academy of Sciences (2008), 10.1073/pnas.0800332105.

[51] D. Brockmann, L. Hufnagel, and T. Geisel, Nature **439**, 462 (2006).

[52] M. C. González, C. A. Hidalgo, and A.-L. Barabási, Nature **453**, 779 (2008).

[53] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, IEEE/ACM Transactions on Networking **19**, 630 (2011).

[54] X.-W. Wang, X.-P. Han, and B.-H. Wang, PLOS ONE **9**, 1 (2014).

[55] D. A. Raichlen, B. M. Wood, A. D. Gordon, A. Z. P. Mabulla, F. W. Marlowe, and H. Pontzer, Proceedings of the National Academy of Sciences **111**, 728 (2014).

[56] S. Bertrand, J. M. Burgos, F. Gerlotto, and J. Atiquipa, ICES Journal of Marine Science **62**, 477 (2005).

[57] N. E. Humphries, H. Weimerskirch, N. Queiroz, E. J. Southall, and D. W. Sims, Proceedings of the National Academy of Sciences **109**, 7169 (2012).

[58] C. Song, T. Koren, P. Wang, and A.-L. Barabási, Nature Physics **6**, 818 (2010).

[59] X.-P. Han, Q. Hao, B.-H. Wang, and T. Zhou, Physical Review E **83**, 036117 (2011).

[60] X.-P. Han, X.-W. Wang, X.-Y. Yan, and B.-H. Wang, PLOS ONE **10**, 1 (2015).

[61] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, Scientific Reports **5**, 9136 (2015).

[62] X.-Y. Yan, X.-P. Han, B.-H. Wang, and T. Zhou, Scientific Reports **3**, 2678 (2013).

[63] X. Liang, X. Zheng, W. Lv, T. Zhu, and K. Xu, Physica A **391**, 2135 (2012).

[64] A. Bazzani, B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini, Journal of Statistical Mechanics: Theory and Experiment **2010**, P05001 (2010).

[65] A. Baronchelli and F. Radicchi, Chaos, Solitons & Fractals **56**, 101 (2013).

[66] T. Rhodes and M. T. Turvey, Physica A **385**, 255 (2007).

[67] F. Radicchi, A. Baronchelli, and L. A. N. Amaral, PLOS ONE **7**, 1 (2012).

[68] F. Radicchi and A. Baronchelli, Physical Review E **85**, 061121 (2012).

[69] A. Clauset, C. Shalizi, and M. Newman, SIAM Review **51**, 661 (2009).

[70] S. Foss, D. Korshunov, S. Zachary, *et al.*, *An Introduction to Heavy-Tailed and Subexponential Distributions*, Vol. 6 (Springer, New York, 2011).

[71] M. Levy and S. Solomon, Physica A **242**, 90 (1997).

[72] E. L. Crow and K. Shimizu, *Lognormal Distributions* (Marcel Dekker, New York, 1987).

[73] E. Limpert, W. A. Stahel, and M. Abbt, BioScience **51**, 341 (2001).

[74] M. Mitzenmacher, Internet Mathematics **1**, 226 (2003).

[75] D. W. Sims, D. Righton, and J. W. Pitchford, Journal of Animal Ecology **76**, 222 (2007).

[76] E. Bonnet, O. Bour, N. E. Odling, P. Davy, I. Main, P. Cowie, and B. Berkowitz, Reviews of geophysics **39**, 347 (2001).

[77] E. P. White, B. J. Enquist, and J. L. Green, Ecology **89**, 905 (2008).

[78] X. Wang and M. Pleimling, Physical Review E **95**, 032145 (2017).

[79] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose, Science **280**, 95 (1998).

[80] R. Fagin, A. R. Karlin, J. Kleinberg, P. Raghavan, S. Rajagopalan, R. Rubinfeld, M. Sudan, and A. Tomkins, The Annals of Applied Probability **11**, 810 (2001).

[81] A. Chmiel, K. Kowalska, and J. A. Hołyst, Physical Review E **80**, 066122 (2009).

[82] F. Chierichetti, R. Kumar, P. Raghavan, and T. Sarlos, in *Proceedings of the 21st International Conference on World Wide Web* (ACM, New York, 2012) pp. 609–618.

[83] Z.-D. Zhao, Z.-G. Huang, L. Huang, H. Liu, and Y.-C. Lai, Physical Review E **90**, 050802 (2014).

[84] Z.-D. Zhao, S.-M. Cai, and Y. Lu, Chaos **25**, 063106 (2015).

[85] M. Szell, R. Sinatra, G. Petri, S. Thurner, and V. Latora, Scientific Reports **2**, 457 (2012).

[86] R. Sinatra and M. Szell, Entropy **16**, 543 (2014).

[87] Z.-D. Zhao, Z. Yang, Z. Zhang, T. Zhou, Z.-G. Huang, and Y.-C. Lai, Scientific Reports **3**, 3472 (2013).

[88] T. Joachims, in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2002) pp. 133–142.

[89] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan, in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (ACM, New York, 2004) pp. 118–126.

[90] N. Craswell and M. Szummer, in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York, 2007) pp. 239–246.

[91] F. Radlinski, M. Kurup, and T. Joachims, in *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (ACM, New York, 2008) pp. 43–52.

[92] G. M. Viswanathan, M. G. Da Luz, E. P. Raposo, and H. E. Stanley, *The Physics of Foraging: An Introduction to Random Searches and Biological Encounters* (Cambridge University Press, Cambridge, UK, 2011).

[93] A. M. Edwards, R. A. Phillips, N. W. Watkins, M. P. Freeman, E. J. Murphy, V. Afanasyev, S. V. Buldyrev, M. G. da Luz, E. P. Raposo, H. E. Stanley, *et al.*, Nature **449**, 1044 (2007).

[94] D. W. Sims, E. J. Southall, N. E. Humphries, G. C. Hays, C. J. Bradshaw, J. W. Pitchford, A. James, M. Z. Ahmed, A. S. Brierley, M. A. Hindell, *et al.*, Nature **451**, 1098 (2008).

[95] N. E. Humphries, N. Queiroz, J. R. Dyer, N. G. Pade, M. K. Musyl, K. M. Schaefer, D. W. Fuller, J. M. Brunnschweiler, T. K. Doyle, J. D. Houghton, *et al.*, Nature **465**, 1066 (2010).

[96] M. de Jager, F. J. Weissing, P. M. J. Herman, B. A. Nolet, and J. van de Koppel, Science **332**, 1551 (2011).

[97] P. Schultheiss, K. Cheng, and A. M. Reynolds, Learning and Motivation **50**, 59 (2015), bee and Insect learning and Cognition.

[98] M. A. Lomholt, K. Tal, R. Metzler, and K. Joseph, Proceedings of the National Academy of Sciences **105**, 11055 (2008).

[99] O. Bénichou, C. Loverdo, M. Moreau, and R. Voituriez, Reviews of Modern Physics **83**, 81 (2011).

[100] C. T. Brown, L. S. Liebovitch, and R. Glendon, Human Ecology **35**, 129 (2007).

[101] E. Korobkova, T. Emonet, J. M. Vilar, T. S. Shimizu, and P. Cluzel, Nature **428**, 574 (2004).

[102] G. Ariel, A. Rabani, S. Benisty, J. D. Partridge, R. M. Harshey, and A. Be'Er, Nature Communications **6**, 8396 (2015).

[103] T. H. Harris, E. J. Banigan, D. A. Christian, C. Konradt, E. D. T. Wojno, K. Norose, E. H. Wilson, B. John, W. Weninger, A. D. Luster, *et al.*, Nature **486**, 545 (2012).

[104] A. M. Edwards, Ecology **92**, 1247 (2011).

[105] A. Kölzsch, A. Alzate, F. Bartumeus, M. de Jager, E. J. Weerman, G. M. Hengeveld, M. Naguib, B. A. Nolet, and J. van de Koppel, Proc. R. Soc. B **282**, 20150424 (2015).

[106] G. H. Pyke, Methods in Ecology and Evolution **6**, 1 (2015).

[107] A. Reynolds, Physics of Life Reviews **14**, 59 (2015).

[108] M. E. Newman, Contemporary Physics **46**, 323 (2005).

[109] S. Chakraborty, Journal of Statistical Distributions and Applications **2**, 6 (2015).

[110] H. A. SIMON, Biometrika **42**, 425 (1955).

[111] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*, 2nd ed. (Springer, New York, 2002).

[112] L. Peng and Y. Qi, The Annals of Statistics **34**, 1964 (2006).

[113] J. Kyselý, Theoretical and Applied Climatology **101**, 345 (2010).

[114] O. Bénichou, M. Coppey, M. Moreau, P.-H. Suet, and R. Voituriez, Physical Review Letters **94**, 198101 (2005).

[115] O. Bénichou, C. Loverdo, M. Moreau, and R. Voituriez, Physical Review E **74**, 020102 (2006).

[116] G. Oshanin, K. Lindenberg, H. S. Wio, and S. Burlatsky, Journal of Physics A: Mathematical and Theoretical **42**, 434008 (2009).

[117] M. Kendall and J. Gibbons, *Rank Correlation Methods*, 5th ed., A Charles Griffin Book (Edward Arnold, London, 1990).

[118] J. M. G. Taylor, Biometrics **43**, 409 (1987).

[119] Sogou Labs, "Search engine click-through log (sogouq)," http://www.sogou.com/labs/dl/q-e.html (2012), accessed on June 5, 2015.

[120] Yahoo Labs, "Anonymized yahoo search logs with relevance judgments version 1.0," https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=50 (2013), accessed on June 5, 2015.

[121] Yahoo Labs, http://labs.yahoo.com/ (), accessed on June 5, 2015.

[122] Yahoo Labs, "The yahoo webscope program," http://webscope.sandbox.yahoo.com/ (), accessed on June 5, 2015.

[123] H. Bauke, The European Physical Journal B **58**, 167 (2007).

[124] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, ACM Trans. Math. Softw. **23**, 550 (1997).

[125] X. Wang and M. Pleimling, Physical Review E **98**, 012126 (2018).

[126] I. Redondo, International Journal of Mental Health and Addiction **13**, 584 (2015).

[127] M. D. Owens, Gaming Law Review and Economics **20**, 567 (2016).

[128] S. M. Gainsbury, Y. Liu, A. M. Russell, and T. Teichert, Computers in Human Behavior **55**, 717 (2016).

[129] A. L. Goldstein, N. Vilhena-Churchill, S. H. Stewart, P. N. S. Hoaken, and G. L. Flett, Journal of Behavioral Addictions **5**, 68 (2016).

[130] M. Chóliz, Journal of Gambling Studies **32**, 749 (2016).

[131] J. Konietzny, International Gambling Studies **17**, 144 (2017).

[132] K. S. Montes and J. N. Weatherly, Journal of Gambling Studies **33**, 85 (2017).

[133] R. Bitar, C. Nordt, M. Grosshans, M. Herdener, E. Seifritz, and J. Mutschler, European Addiction Research **23**, 106 (2017).

[134] N. Hing, A. M. Russell, and M. Browne, Frontiers in Psychology **8**, 779 (2017).

[135] A. González-Roz, J. R. Fernández-Hermida, S. Weidberg, V. Martínez-Loredo, and R. Secades-Villa, Journal of Gambling Studies **33**, 371 (2017).

[136] M. Auer and M. D. Griffiths, Journal of gambling studies **33**, 795—806 (2017).

[137] R. Edgren, S. Castrén, H. Alho, and A. H. Salonen, Computers in Human Behavior **72**, 46 (2017).

[138] D. Martinelli, Gaming Law Review **21**, 557 (2017).

[139] J. T. Holden and S. C. Ehrlich, Gaming Law Review **21**, 566 (2017).

[140] R. Sylvester and P. Rennie, Gaming Law Review **21**, 625 (2017).

[141] M. Griffiths, Casino & Gaming International **28**, 59 (2017).

[142] C. Grove, "Understanding skin gambling," (2016), accessed on January 8, 2019.

[143] K. Park and E. Domany, Europhysics Letters **53**, 419 (2001).

[144] T. Ichinomiya, Physica A **368**, 207 (2006).

[145] S. Pigolotti, S. Bernhardsson, J. Juul, G. Galster, and P. Vivo, Physical Review Letters **108**, 088701 (2012).

[146] J. Juul, A. Kianercy, S. Bernhardsson, and S. Pigolotti, Physical Review E **88**, 022806 (2013).

[147] Y. Zhao, Q. Chen, and Y. Wang, International Journal of Modern Physics C **25**, 1440002 (2014).

[148] CSGO Casino, `http://www.csgo-casino.com/`, accessed on February 15, 2018.

[149] CS:GO BACKPACK, `http://www.csgobackpack.net/`, accessed on February 15, 2018.

[150] V. M. Yakovenko and J. B. Rosser, Reviews of Modern Physics **81**, 1703 (2009).

[151] R. N. Mantegna and H. E. Stanley, Physical Review Letters **73**, 2946 (1994).

[152] J.-i. Inoue and N. Sazuka, Physical Review E **76**, 021111 (2007).

[153] K. Ito and S. Miyazaki, Progress of Theoretical Physics **110**, 875 (2003).

[154] Y. Maruyama and J. Murakami, Physical Review B **67**, 085406 (2003).

[155] D. Stauffer, C. Schulze, and D. W. Heermann, Journal of Biological Physics **33**, 305 (2008).

[156] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, 2001).

[157] S. Buldyrev, M. Gitterman, S. Havlin, A. Y. Kazakov, M. da Luz, E. Raposo, H. Stanley, and G. Viswanathan, Physica A **302**, 148 (2001).

[158] Y. Kantor and M. Kardar, Physical Review E **76**, 061121 (2007).

[159] D. S. Novikov, E. Fieremans, J. H. Jensen, and J. A. Helpern, Nature physics **7**, 508 (2011).

[160] B. Dybiec, E. Gudowska-Nowak, E. Barkai, and A. A. Dubkov, Physical Review E **95**, 052102 (2017).

[161] M. Ghaemi, Z. Zabihinpour, and Y. Asgari, Physica A **388**, 1509 (2009).

[162] A. G. Association, (2018), accessed October 1, 2018.

[163] F. Calado and M. D. Griffiths, Journal of Behavioral Addictions **5**, 592 (2016).

[164] F. Calado, J. Alexandre, and M. D. Griffiths, Journal of Gambling Studies **33**, 397 (2017).

[165] D. Kahneman and A. Tversky, Econometrica **47**, 263 (1979).

[166] A. Tversky and D. Kahneman, Journal of Risk and Uncertainty **5**, 297 (1992).

[167] N. Barberis, Management Science **58**, 35 (2012).

[168] J. Meng, *Understanding Gambling Behavior and Risk Attitudes Using Massive Online Casino Data*, Bachelor's thesis, Dartmouth College, Hanover, New Hampshire, USA (2018).

[169] D. Brockmann, The European Physical Journal Special Topics **157**, 173 (2008).

[170] S. Kim, C.-H. Lee, and D. Y. Eun, IEEE Transactions on Mobile Computing **9**, 288 (2010).

[171] J. T. Holden, UCLA Entertainment Law Review **25**, 41 (2018).

[172] R. Buhagiar, D. Cortis, and P. W. Newall, Journal of Behavioral and Experimental Finance **18**, 85 (2018).

[173] P. Rodríguez, B. R. Humphreys, and R. Simmons, *The Economics of Sports Betting* (Edward Elgar Publishing, 2017).

[174] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, in *Proceedings of the 2013 Conference on Internet Measurement Conference* (ACM, New York, 2013) pp. 127–140.

[175] CSGOFAST, `https://csgofast.com/`, accessed on April 20, 2018.

[176] CSGOSpeed, `https://csgospeed.com/home`, accessed on April 20, 2018.

[177] ethCrash, `https://www.ethcrash.io/play`, accessed on August 1, 2018.

[178] SatoshiDICE, `https://www.satoshidice.com/`, accessed on August 1, 2018.

[179] Coinroll, `https://coinroll.com/home`, accessed on August 1, 2018.

[180] I. Fiedler, "Online gambling as a game changer to money laundering?" (2013), accessed on January 8, 2019.

[181] Coindesk, "Bitcoin (usd) price," `https://www.coindesk.com/price/bitcoin/`, accessed on August 1, 2018.

[182] CoinMetrics, "Data downloads," `https://coinmetrics.io/data-downloads/`, accessed on August 1, 2018.

[183] S. I. Millar, Gaming Law Review **22**, 174 (2018).

[184] S. Kairouz, C. Paradis, and L. Nadeau, Cyberpsychology, Behavior, and Social Networking **15**, 175 (2012).

[185] S. M. Gainsbury, Current Addiction Reports **2**, 185 (2015).

[186] J. Banks, *Gambling, Crime and Society* (Palgrave Macmillan, London, 2017).

[187] J. Macey and J. Hamari, New Media & Society **21**, 20 (2019).

[188] R. Gibrat, Bulletin de Statistique General, France **19**, 419 (1930).

[189] R. A. Epstein, *The Theory of Gambling and Statistical Logic* (Academic Press, 2012).