2025 年度 修士論文

ランダムなしりとりのネットワーク解析

Network Analysis of Random Word Chain Games

大阪公立大学 大学院理学研究科
物理学専攻 非線形物理研究室

BHB24061 藤田悠朔

2026 年 2 月 23 日

概要

しりとりは，“りんご” → “ごりら” → “らっぱ” → · · · のように，前の単語の末尾文字から始まる単語をつなげていくゲームである．他の基本的なルールとしては (i) 一度使用した単語は以後使用できない (ii) 次の単語がなくなったら終了（負け）がある．本研究では，辞書（使用できる単語の集合）を定め，ゲームの勝敗や戦略を度外視したランダムなしりとりに着目する．つまり，次の単語は残されている選択可能な単語の中からランダムに選び，選択された単語は辞書から消す，という過程を終了するまで繰り返す．こうしてできる鎖長（単語列の長さ）がどのようになるのかに着目した．解析には辞書ネットワーク，すなわち辞書から構成される，文字を頂点，単語を辺とする多重有向グラフを用いた．これにより，ランダムなしりとりは辞書ネットワーク上の自己回避的なランダムウォークとみなすことができる．我々は「単一辞書」，あるいは各文字の次数のみを保存した辞書の集合である「シャッフル辞書群」でランダムなしりとりを実行したときの鎖長について，数値計算と理論の両面から統計的に解析した．

目次

記号一覧

記号	意味
θ, ϕ	文字を表す変数
“$\theta \cdots \phi$”	文字 θ で始まり文字 ϕ で終わる単語
Θ	辞書内の単語の先頭文字と末尾文字の集合
C	辞書内の単語の先頭文字と末尾文字の種類数
D	辞書の総単語数
$D^{(l)}$	l ステップ目の辞書の総単語数
L	鎖長（単語列の長さ）
$w^{(l)}$	l ステップ目の単語
$\theta^{(l)}$	l ステップ目の単語の末尾文字
$q_{\theta\phi}^{(l)}$	単一辞書で l ステップ目に単語 “$\theta \cdots \phi$” を選択する確率
$p(L)$	単一辞書の鎖長の分布
$p_\theta(L)$	単一辞書の文字 θ で終了する鎖長の分布
$p^T(L)$	試行回数 T で実測した単一辞書の鎖長の分布
Q_θ	単一辞書における文字 θ の終了確率
$\tilde{q}_{\theta\phi}^{(l)}$	シャッフル辞書群で l ステップ目に単語 “$\theta \cdots \phi$” を選択する確率
$\tilde{p}_\theta(L)$	シャッフル辞書群の文字 θ で終了する鎖長の分布
$\tilde{p}_\theta^N(L)$	冊数 N で実測したシャッフル辞書群の文字 θ で終了する鎖長の分布
$\tilde{Q}_\theta$	シャッフル辞書群における文字 θ の終了確率
$k_{\mathrm{in},\theta}$, $k_{\mathrm{out},\theta}$	初期状態の頂点 θ の入次数，出次数（単語辺型）
$\overrightarrow{k_{\mathrm{in}}}$, $\overrightarrow{k_{\mathrm{out}}}$	各頂点の入次数，出次数を並べたベクトル（単語辺型）
$k_{\mathrm{in},\theta}^{(l)}$, $k_{\mathrm{out},\theta}^{(l)}$	l ステップ目の頂点 θ の入次数，出次数（単語辺型）
$\kappa_{\mathrm{in},\theta\phi}$, $\kappa_{\mathrm{out},\theta\phi}$	初期状態の頂点 “$\theta \cdots \phi$” の入次数，出次数（単語頂点型）
$P_{\kappa_{\mathrm{in}}}$, $P_{\kappa_{\mathrm{out}}}$	入次数分布，出次数分布（単語頂点型）
A	初期状態のネットワークの隣接行列．$A_{\theta\phi}$ はその $\theta\phi$ 成分
$A^{(l)}$	l ステップ目のネットワークの隣接行列．$A_{\theta\phi}^{(l)}$ はその $\theta\phi$ 成分
$J^{(l)}$	l ステップ目の経路行列．$J_{\theta\phi}^{(l)}$ はその $\theta\phi$ 成分
J	鎖経路行列（これ以上伸びない経路に用いる）．$J_{\theta\phi}$ はその $\theta\phi$ 成分
$\deg(J^{(l)})$	経路行列 $J^{(l)}$ の縮退度（構成可能な末尾文字列の個数）

記号	意味
$HG(m_1; N, n_1, M)$	超幾何分布の確率質量関数
$NHG(m_2; N, n_2, m_1)$	負の超幾何分布の確率質量関数
$MHG(\{m_1, m_2, \cdots, m_C\}; N, \{n_1, n_2, \cdots, n_C\}, M)$	多変量超幾何分布の確率質量関数
$MIHG(\{m_2, m_3, \cdots, m_C\}; N, \{n_2, n_3, \cdots, n_C\}, m_1)$	多変量逆超幾何分布の確率質量関数
$d_2(f, g)$	確率質量関数 $f(x)$, $g(x)$ の L^2 ノルム
$d_{TV}(f, g)$	確率質量関数 $f(x)$, $g(x)$ の全変動距離
$a^{[n]}$	Pochhammer 記号. 整数 a, n $(n \geq 0)$ について, $a^{[n]} = a(a-1)\cdots(a-n+1)$, $a^{[0]} = 1$

第 1 章

序論

しりとりは，前の単語にその末尾文字から始まる単語を繋げていくゲームである．例えば，“りんご” → “ごりら” → “らっぱ” → · · · と続く．ここで，同じ単語を 2 回以上使うことはできない．プレイヤーは 1 人，または複数人で行う．単語が思いつかないか，語尾に「ん」のつく単語を言ったら終了，または負けとなる．

この極めて単純なルールは，世代を問わず誰でも楽しむことができる言葉遊びとして，しりとりを定着させた．さらに，本質的なルールは言語に依存しないため，日本のみならず世界各国で親しまれている．例えば，英語圏では word chain game という名前で知られている．しかし，ルールのシンプルさに反して多くの問題設定が考えられ，これまで様々な分野でしりとりが研究されてきた．特にしりとりの鎖長（単語列の長さ）に焦点を当てた研究がよく見られ，しりとりは最長で何単語続くのか [1] や，2 人しりとりにおける最も効果的な戦略は何か [2] といった先行研究が存在する．これらは戦略的にしりとりの鎖長を変化させる研究である．我々は戦略を持たないしりとりにおいて，鎖長がどうなるのかに着目した．

本研究では辞書（使用できる単語の集合）を定め，ゲームの勝敗や戦略を度外視したランダムなしりとりを考える．すなわち，次の単語は残されている選択可能な単語の中からランダムに選び，選択された単語は辞書から消す，という過程を終了するまで繰り返す．一連の過程はネットワーク上の通過した辺を回避するランダムウォークとして定式化し，解析することが可能である．本研究の目的は，ランダムなしりとりにおける鎖長の分布を数値的および理論的に解析し，その特徴を明らかにすることである．

本研究の問題設定はネットワーク科学にも関係している．ネットワーク上の自己回避ランダムウォーク (SAW) についてはいくつかの先行研究が存在し，ランダムネットワーク [3] やスケールフリーネットワーク [4] 上で SAW を実行したときの経路長分布が解析的に求められている．一方で，辞書から構成されるネットワークは先行研究のような典型的な次数分布を持たない．このようなネットワークにおける経路長分布がどのような特徴を持つのかは興味深いテーマである．

本研究で得られた主な結果は以下の 3 点である．(i) 辞書のネットワーク構造から，しりとりが終了する文字の条件を明らかにした（第 3 章）．(ii) 平均場近似を導入し，同じ条件を

持つ辞書の集合「シャッフル辞書群」に対する平均鎖長分布および分布統計量を解析的に求めた（第 4 章）．(iii) 単一の辞書に対し鎖長分布を効率的に求めるアルゴリズムを構築した（第 5 章）．

最後に本論文の構成について述べておく．第 2 章ではしりとりとそのモデルである SAW に関する先行研究を紹介する．第 3 章では本研究の詳しい問題設定と解析手法について記述する．第 4 章ではシャッフル辞書群における平均鎖長分布およびその分布統計量の解析的な導出を試みる．第 5 章では単一辞書における鎖長分布を求めるアルゴリズムを構築し，単一辞書の計算が難しい理由について考察する．第 6 章で本研究をまとめ，今後の課題を記す．また，付録 A, B に本論文で登場する辞書の詳細なデータを，付録 C, D, E に本論文で登場する数学的概念や定理の説明を，付録 F, G, H に本研究の発展的課題とその途中結果を記載している．なお，本論文の成果の一部は [5] に掲載されている．

第 2 章

先行研究

本章ではしりとり，およびそれをモデル化したネットワーク上の SAW に関する先行研究を紹介する．

2.1 しりとりについて

しりとりはこれまで数理科学から心理学，教育学と幅広い分野で研究されてきた．本節ではその中でも本研究との関連性が高いものを紹介する．ただし，単語の最後の文字を末尾文字と呼ぶことにする．

数理的研究の 1 つ目は，「しりとりは先手必勝か，後手必勝か」[6] である．2 人のしりとりは使用できる単語を定めることで組合せゲーム（二人確定完全情報ゲーム）となり，理論上の勝敗が一意に定まる．ただし，問題のサイズが大きい場合，ゲーム木の全探索によって勝敗を導き出すことは計算量の観点からほぼ不可能である．そこで，勝敗を保ったまま局面を簡略化する方法が提案された．これにより文字数が 3 以下の問題に帰着させることができれば，勝敗を直ちに判定できることが明らかとなった．

2 つ目は，「与えられた辞書の単語でしりとりをしたときに，最長でどれほど長く続くのか」[1] である．辞書をネットワークで表現すると，しりとりの単語列はネットワークの 2 頂点間を流れるフローとなり，この問題はフロー量の総和を最大化する問題に帰着される．これに対応する整数計画問題を直接解くのは難しいが，緩和問題を繰り返し解くことにより，約 19 万単語の辞書における最長しりとりを 1.36 秒で求めることが可能となった．より一般的な問題として，総文字数が最大となるしりとりの単語列を求める問題も分析されている [7].

3 つ目は，「2 人のしりとりにおける最も効果的な戦略は何か」[2] である．しりとりには特定の文字で終了する単語を繰り返し使う戦略が存在する．この研究ではどの文字で切り返すのが最も効果的かを，日本語の辞書を用いた統計調査によって導いている．それによれば，攻撃の仕掛けやすさや「ん」での終わりやすさから，末尾文字が「り」の単語を繰り返し使うのが最も効果的であると結論づけている．

他にも発達心理学の観点から，幼児が語彙を獲得する過程 [8] や，音韻意識を獲得する過

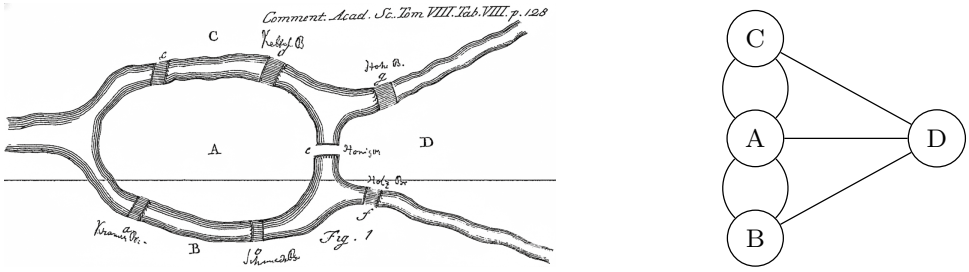
程 [9] が，しりとりを用いた調査により分析されている．また教育学の観点から，外国語の語彙習得にしりとりが効果的であることが，複数の実証実験 [10][11] によって報告されている．このようにしりとりと語彙習得の間には関連性が見られる．

最後に本研究との関連性について言及する．まず，数理的研究，特に鎖長を戦略的に変化させる研究において，戦略を考えないランダムなしりとりの鎖長は基礎的なデータとして有用である．また，心理学的，教育学的研究に関連して，語彙数（辞書の単語数）としりとりの鎖長の間の関係が判明すれば，しりとりをすると語彙力が測定される仕組みを構築できるかもしれない．本研究の問題設定は単なる数学的興味にとどまらず，しりとりの研究全体に多くの影響を与えることが期待される．

2.2 ネットワークについて

しりとりは使用できる単語の集合，辞書を定めることで，ネットワークを用いた解析が可能となる．本節ではネットワーク科学の起源と発展の歴史について整理する．

グラフとは，頂点とそれらを繋ぐ辺の集合のことである．そして，グラフが持つ性質を探求する数学の分野はグラフ理論と呼ばれる．グラフの概念を初めて導入したのは 18 世紀の Euler とされている．Euler は「図 2.1 (a) の 7 つの橋すべてを 1 度ずつ通ることはできるのか」という問題を，図 2.1 (b) のグラフ上の一筆書き問題に落とし込んで，それが不可能であることを証明した [12]．これに因んでか，グラフ上のすべての辺を一度ずつ通る経路，閉路はそれぞれ Euler 経路，Euler 閉路と呼ばれている．



(a) Euler による地図（[12] の図を改変）　(b) 陸地を頂点，橋を辺とするグラフ

図 2.1: Königsberg の橋問題

幅広い科学分野において，グラフはネットワークと呼ばれることが多い．そして，地理的なつながりに限らず，モノとモノのつながりをネットワークで捉え，解析する分野はネットワーク科学と呼ばれる．1998 年に Watts, Strogatz は，高いクラスター性と短い平均経路長を併せ持つ「スモールワールド性」が多くの現実のネットワークに見られることを明らかにした [13]．続く 1999 年に Barabási, Albert は，次数分布が冪乗則に従う「スケールフリー性」が多くの現実のネットワークに現れることを指摘した [14]．これら 2 つの研究を皮切りに，大規模なネットワークが多くの科学者によって解析されるようになった．今ではその応

用先は非常に広範囲であり，人間や企業などの社会的な関係から，情報分野，工学分野，生物学分野などの様々なシステムがネットワークによってモデル化されている [15]．単語間の関係もネットワークでモデル化され，文章中で隣接する単語を繋いだネットワーク [16] や，シソーラスで類義語どうしを繋いだネットワーク [17] から，自然言語が持つ特徴を分析する研究が行われている．しりとりに関しても，辞書から文字を頂点，単語を辺とする多重有向グラフ（辞書ネットワーク）を構成することで，数理的に解析することができるようになる．

2.3 SAW について

ランダムなしりとりは辞書ネットワーク上の通過した辺を回避するランダムウォークとしてモデル化される．本節ではこのランダムウォークに関する先行研究を紹介する．

自己回避ランダムウォーク (SAW) とは，同じ頂点を 2 度通らないランダムウォークのことである．この章で扱う SAW は，次の頂点をまだ訪れていない隣接頂点の中から選択し，どの隣接頂点もすでに訪れている状態になったら終了する．このような成長過程を持つ SAW は，kinetic growth walk [4] や genuine self-avoiding walk [18]，true self-avoiding walk [19]，myopic self-avoiding walk [19] とも呼ばれる．一方でランダムなしりとりに見られるような，同じ辺を 2 度通らないランダムウォークは自己回避トレイル (SAT) [20]，あるいは単にトレイル [21] と呼ばれる．ただし，あるネットワーク上の SAT はその線グラフ（頂点と辺を入れ替えたグラフ[*1]）上の SAW とみなすことができる（図 2.2）ので，この章ではよく研究されている SAW の方に焦点を当てる．

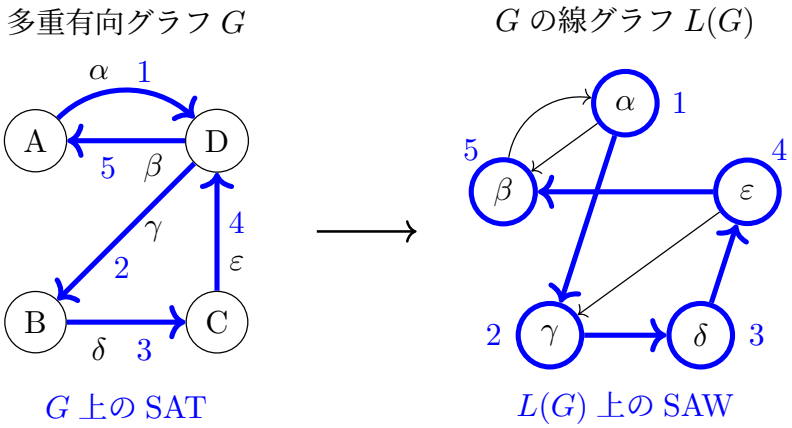


図 2.2: グラフ上の SAT はその線グラフ上の SAW に対応する．

SAW は高分子物理学に端を発している．高分子をセグメント（いくつかの原子からなる構成単位）の繋がりとみなし，排除体積効果（セグメント同士が重なり合わない効果）を考慮

[*1] 厳密には，ある有向グラフ G の線グラフ $L(G)$ は次のように構成される [22]．$L(G)$ の頂点を G のすべての辺とする．そして，G の任意の辺 e_1, e_2 について，e_1 の終点と e_2 の始点が一致しているとき，$L(G)$ 上で頂点 e_1 から頂点 e_2 への有向辺を張る．

に入れると，それは格子上の SAW としてモデル化される [23]．それゆえ，d 次元格子上の SAW について，長さ L の経路の総数 c_L や出発点からの平均距離 $\langle|\omega(L)|^2\rangle$ などの様々な性質が，理論と数値計算の両面から活発に研究されてきた [19]．

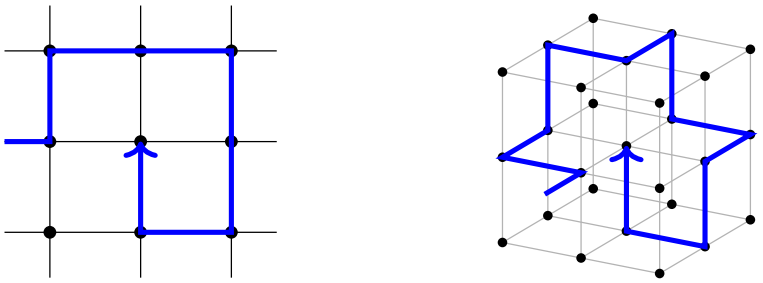
一方で，ネットワーク上の SAW の応用はあまり見られず，理論研究も格子上の研究ほど活発に行われていない．しかし，単純なランダムウォークよりも SAW の方がネットワーク上を効率的に探索できることが知られている [24]．ネットワーク上のランダムウォーク自体は現実の問題によく適用されており，ウェブのリンク構造から決まるウェブページの重要度の計測 [25] や，複雑ネットワークの次数分布や平均クラスター係数といったトポロジカルな構造の推測 [26] に利用されている．これらの探索を SAW が担うことで，ウェブページの重要度やネットワーク構造の計測が効率化されることが期待される．

2.4 SAW の経路長分布

ランダムなしりとりの鎖長分布を求める問題は，与えられたネットワーク上で SAW を実行したときに，その経路長（終了するまでのステップ数）L がどのように分布するのかを求める問題と考えられる．一般に SAW の経路長分布 $p(L)$ の厳密解を求めることは難しいが，ある種の格子，ネットワークにおいては数値解や近似解が求められている．本節ではそのいくつかを紹介し，各種格子，ネットワークにおける結果を表 2.1 にまとめる．

規則格子上の SAW は図 2.3 のような自己トラップが生じたときに終了する．P. C. Hemmer らは超立方格子上の SAW が確率 1 で終了することを証明した．つまり，いつまでも SAW が続くことはない．これは同じ頂点を k 回通過してよいとした場合でも同様に成り立ち，特に一次元格子上で $k = 2$ の SAW は平均 25 ステップ続くことが厳密に求められている [18]．S. Hemmer らは二次元正方格子上で 6 万回 SAW を実行するシミュレーションを行った．その結果，SAW はいずれも 490 ステップ以内に終了し，平均値は約 71 ステップ，最頻値は 33 ステップとなったことが報告されている．さらに，経路長分布は図 2.4 のようになり，これは表 2.1 に示した指数減衰に近い式で表されることが判明した [27]．I. Majid らは三次元立方格子上で SAW を実行した．三次元では図 2.3 (b) のような自己トラップが稀にしか起こらないため，平均約 5000 ステップと二次元の場合に比べて非常に長く続くことが報告されている [28]．

I. Tishby らは Erdős-Rényi ネットワーク上の SAW に関する解析を行った．Erdős-Rényi ネットワークの次数分布は Poisson 分布で表されるが，SAW により頂点が除去された後のネットワークもまた Poisson 分布に従うことが報告されている．また，SAW の経路長分布 $p(L)$ が解析的に求められ，ネットワークのサイズ N が十分大きい場合，表 2.1 に示したような Gompertz 分布に従うことが判明した．さらに $p(L)$ の振る舞いは，図 2.5 のように与えられたネットワークの平均次数 c によって変化し，c が小さいときは単調減少となり，ある値を超えるとピークを持ち始め，c が大きくなるほどピークは右に移動することが指摘されている [3]．



(a) 二次元格子　　(b) 三次元格子

図 2.3: 規則格子上の自己トラップの例（[18] の図を元に作成）

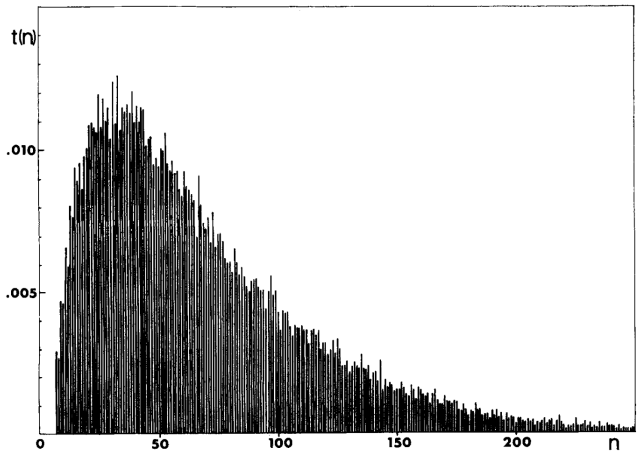


図 2.4: 二次元正方格子上の SAW の経路長分布．図は 6 万回 SAW を実行した結果であり，経路長が 250 を超えたものは表示していない．なお，経路長を n，SAW が長さ n で終了する確率を $t(n)$ で表している（[27] より転載）．

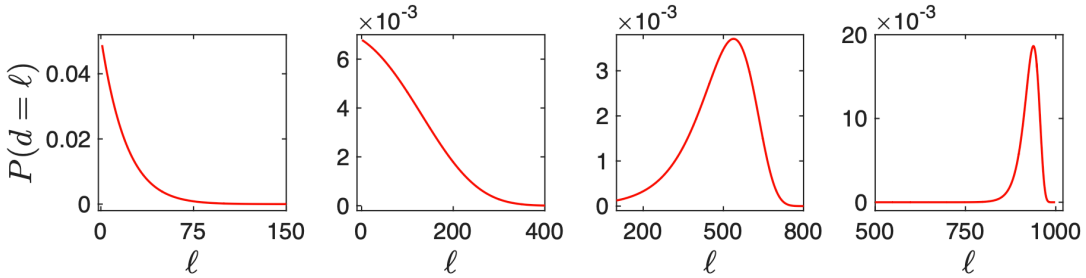
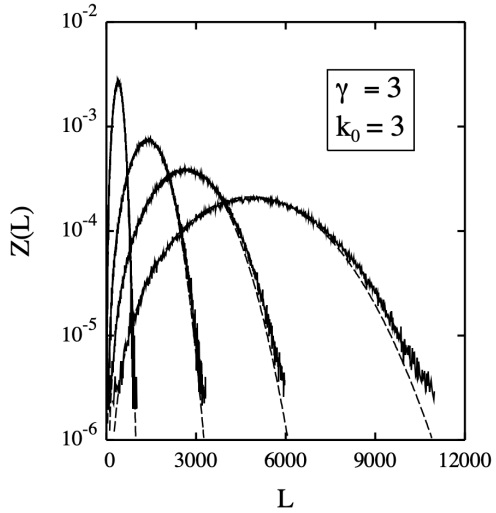
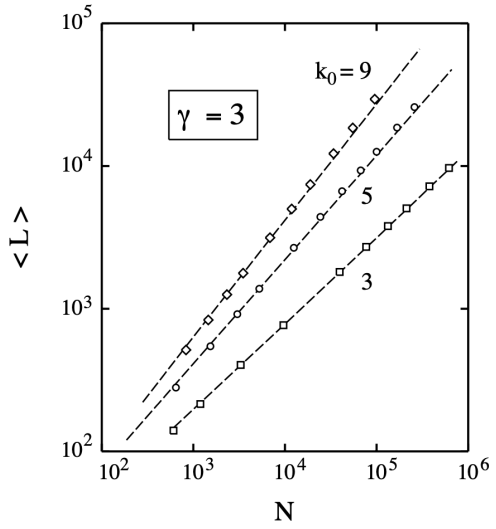


図 2.5: Erdős-Rényi ネットワーク上の SAW の経路長分布．図は解析的に得られた結果であり，初期ネットワークは頂点数が $N = 1000$，平均次数が左から $c = 3, 5, 10, 50$ としている．なお，経路長を l，SAW が長さ l で終了する確率を $P(d = l)$ で表している（[3] より転載）．

C. P. Herrero はスケールフリーネットワーク上の SAW に関する解析を行った．スケールフリーネットワークの次数分布は冪乗則に従う．その冪指数を γ とする．全頂点数 N が経路長 L に比べて十分大きいという近似のもと経路長分布 $p(L)$ を解析的に求めると，表 2.1 に示したような式で記述され，最小次数 k_0 に強く依存することが判明した．さらに平均経路長 $\langle L \rangle$ は N^{α} に比例し，$\gamma > 3$ のとき $\alpha = 1 - 1/k_0$ と書けることが明らかとなった．以上の解析結果は，図 2.6 のように数値シミュレーションともよく一致したことが報告されている [4]．C. P. Herrero は他にもスモールワールドネットワーク [29] や正則ランダムネットワーク（全ての頂点が同じ次数を持つランダムネットワーク）[30] 上の SAW に関する解析を行っており，結果の一部を表 2.1 に記載している．



(a) 経路長分布．ネットワークの冪指数は $\gamma = 3$，最小次数は $k_0 = 3$，サイズは左から $N = 3.3 \times 10^3, 2.6 \times 10^4, 7.7 \times 10^4, 2.1 \times 10^5$ のものを用いた．実線がシミュレーション結果，破線が理論値を表している．



(b) 平均経路長のサイズ N 依存性．ネットワークの冪指数は $\gamma = 3$，最小次数は左から $k_0 = 9, 5, 3$ のものを用いた．実線がシミュレーション結果，記号が理論値を表している．

図 2.6: スケールフリーネットワーク上の SAW の経路長分布と平均経路長のサイズ依存性．なお，この図では SAW が長さ L で終了する確率を $Z(L)$ で表している（[4] より転載）．

以上のネットワークに関する先行研究は，単一のネットワークの挙動ではなく，同一の性質を有するネットワークの集合に対する平均的な挙動に着目している．こうすることでパラメータの数が 2, 3 個に削減され，解析がしやすくなる．例えば，ランダムネットワークは頂点数と平均次数が与えられるだけなので，様々なネットワークが構成される．しかし，それらの平均を考えることで図 2.5 のような解析が可能となる．しりとりについても，単一の辞書ネットワークを扱う前に，ある性質を持つ辞書ネットワーク全体の平均をとった「シャッフル辞書群」で解析を行う．ただし，文字の種類数を C とすると，単一の辞書では隣接行列の各成分の C^2 個，シャッフル辞書群であっても各文字の入次数，出次数の $2C - 1$ 個のパラメータが必要となる．

ちなみに，単語頂点型辞書ネットワークの次数分布は Poisson 分布や冪乗則といった典型的な分布にはならない．その代わりに，同じ文字で始まる単語を表す頂点はすべて同じ入次数を持ち，同じ文字で終わる単語を表す頂点はすべて同じ出次数を持つ．つまり，単語頂点型の入次数や出次数は最大で C 種類しか存在せず，次数分布はとびとびの構造を持つことになる（付録 A を参照）．したがって，鎖長分布を計算することができれば，このようなネットワーク上の SAW の経路長分布が計算されたことになり，先行研究との比較が可能となる．

表 2.1: SAW の経路長分布

格子・ネットワーク	経路長分布	平均経路長	参照
二次元正方格子	$p(L) \sim (L-6)^{\frac{3}{5}} e^{-\frac{L}{40}}$（数値解）	$\langle L \rangle = 70.7 \pm 0.2$	[27]
ランダムネットワーク （頂点数 N，平均次数 c）	$p(L) \simeq \exp\left[-\frac{N}{c} e^{-c} (e^{\frac{c}{N} L} - 1) - \left(1 - \frac{L+1}{N}\right) c \right]$ *1	脚注に記載 *1*2	[3]
スモールワールドネットワーク （頂点数 N，平均次数 c，再接続確率 p）	$p(L) \simeq t(1-t)^L$ *1*4	脚注に記載 *1*3	[29]
スケールフリーネットワーク （頂点数 N，冪指数 γ，最小次数 k_0）	$p(L) \simeq \frac{k_0}{L} \left(\frac{L}{x_0}\right)^{k_0} \exp\left[-\left(\frac{L}{x_0}\right)^{k_0} \right]$ *1*5	$\langle L \rangle \simeq \frac{x_0}{k_0} \Gamma\left(\frac{1}{k_0}\right)$ *1*5*7	[4]
正則ランダムネットワーク （頂点数 N，次数 k）	$p(L) \simeq \left(\frac{(k-2)L}{kN}\right)^{k-1} \exp\left[-\left(\frac{L}{y_0}\right)^{k} \right]$ *1*6	$\langle L \rangle \simeq \frac{y_0}{k} \Gamma\left(\frac{1}{k}\right)$ *1*6	[30]

*1 $N \gg 1$，$N \gg L$ などの近似を使っている．

*2 $c < W(N)$ のとき，$\langle L \rangle \simeq 1 + e^c$，$c > W(N)$ のとき，$\langle L \rangle \simeq 1 + \left[N - \frac{N}{c} \left(\log \frac{N}{c} + \gamma \right) + \left(\frac{N}{c}\right)^2 e^{-c} - \frac{1}{4} \left(\frac{N}{c}\right)^3 e^{-2c} \right] \exp\left(\frac{N}{c} e^{-c}\right)$ である．ただし，$W(x)$ は Lambert W 関数，γ は Euler 定数を表している．

*3 $p \ll 1$ のとき，$\langle L \rangle \simeq \langle L \rangle_{2D} + \frac{q}{2} [\langle L \rangle_{2D} + 2 \langle L \rangle_{2D}^2 - \langle L^2 \rangle_{2D}]$，$p \to 1$ のとき，$\langle L \rangle \simeq 27 e^{2p} p^{-2} (1-p)^{-4}$ である．ただし，$\langle L^n \rangle_{2D}$ は二次元正方格子における経路長 n 次モーメント，μ は二次元正方格子の connective 定数を表し，$q = \frac{p}{2} \left(1 + \frac{4}{\mu}\right)$ である．

*4 t は各ステップで SAW が終了する条件付き確率を表す．特に $p \to 1$ のとき，$t \simeq \frac{1}{27} e^{-2p} p^2 (1-p)^4$ である．

*5 x_0 はネットワークから決まる定数で，$x_0 = \left(\frac{N^{k_0} \langle k \rangle}{N_{k_0} w^{k_0 - 1}} \right)^{\frac{1}{k_0}}$，$w = \frac{\langle k^2 \rangle - 2 \langle k \rangle}{\langle k \rangle^2}$，$N_{k_0}$ は次数 k_0 の頂点の数である．

*6 y_0 はネットワークから決まる定数で，$y_0 = k \left(\frac{N}{k-2}\right)^{1 - \frac{1}{k}}$ である．

*7 $N \to \infty$ の極限において，$2 < \gamma \leq 3$ のとき，$\langle L \rangle \sim \left(\frac{N}{\langle k^2 \rangle}\right)^{1 - \frac{1}{k_0}}$，$\gamma > 3$ のとき，$\langle L \rangle \sim N^{1 - \frac{1}{k_0}}$ となる．

第 3 章

問題設定

本章ではランダムなしりとりを定義し，本研究の問題設定を詳しく述べる．次に，解析手法として辞書ネットワークを導入し，ランダムなしりとりとの関係性を明らかにする．最後に，ランダムなしりとりを辞書ネットワークの観点から再定式化し，特定の手順を書き換えることで平均場近似を定義する．

3.1　ランダムなしりとりの鎖長分布

本研究では使用できる単語の集合，辞書を定めた上で，以下の設定でしりとりを行う．

1. 初期条件として文字をランダムに選択する．
2. 辞書の中から前の末尾文字から始まる単語をランダムに選択する．そして選択された単語を辞書から消す．これを繰り返す．
3. 単語を選択できなくなったら終了し，そのときの単語列の長さ L を鎖長とする．

このようなしりとりをランダムなしりとりと呼ぶ．なお，日本語のしりとりを考えると，末尾文字が拗音や長音，濁音，半濁音の場合にどう処理するのかで任意性が生じてしまう．そこで今回は英語のしりとりを考えることとする．またそれに伴って，特定の文字で終わる単語を選択したら終了というルールは考えない．我々の目的は，ランダムなしりとりを実行したときの鎖長分布を求めることである．

本研究で主に使用する辞書として，2025 年 5 月時点で国際連合に加盟している国の英語名 193 語 [31] を用いた（単語の一覧は付録 A に記載）．以下，この辞書を「国名辞書」と呼ぶ．国名辞書を用いて 100 万回ランダムなしりとりを実行したときの鎖長分布は図 3.1 のようになった．鎖長分布は横軸に鎖長 L，縦軸にその鎖長が実現する確率 $p(L)$ をとっている．100 万回実行した範囲では，鎖長の最小値は 2，最大値は 39，平均値は 17.9 ± 3.8 となった．理論上，最小値が “yemen”→“norway” の 2 であることは確実だが，最大値が 39 なのか，それよりも長い単語列が存在しているのかは定かではない．さらに，この非対称で歪な形状の分布はどのようにして得られるのかという疑問も提起される．

ちなみに，辞書を変えると鎖長分布が大きく変化する場合がある．例えば，文学作品 “Moby-Dick” [32] に登場する英語の名詞 19088 語からなる辞書，「Moby-Dick 辞書」で 10 万回ランダムなしりとりをした結果が図 3.2 である．国名辞書の鎖長分布と比べて分布の形状が大きく異なっていることがわかる．なお，これ以降の Moby-Dick 辞書による解析結果は付録 B に記載する．

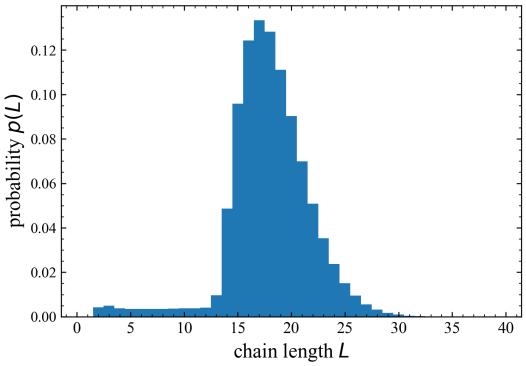


図 3.1: 国名辞書の鎖長分布（100 万回実測）

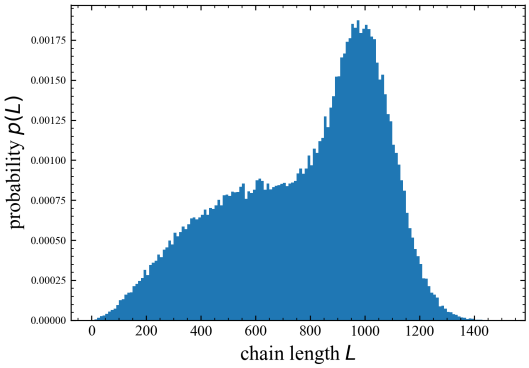


図 3.2: Moby-Dick 辞書の鎖長分布（10 万回実測）

国名辞書でしりとりを 100 万回試行することの妥当性について検討しておく．図 3.3 は各試行回数 $T_i = 10^{2+0.5i},\ i = 0, 1, 2, \cdots, 8$ に対して，試行回数 T_i の国名辞書の鎖長分布 $p^{T_i}(L)$ と，試行回数 T_{i-1} の分布 $p^{T_{i-1}}(L)$ との全変動距離 $d_{TV}(p^{T_i}, p^{T_{i-1}}),\ i = 1, 2, \cdots, 8$ を測定したものである．全変動距離 $d_{TV}(f, g)$ は，2 つの確率質量関数 $f(x), g(x)$ に対して次のように定義される距離である．

$$d_{TV}(f, g) \coloneqq \frac{1}{2} \sum_{x} |f(x) - g(x)|. \tag{3.1}$$

全変動距離の特徴は値の範囲が $0 \leq d_{TV}(f, g) \leq 1$ をとるところにある．試行回数が小さいうちは試行回数の変化による分布関数の変動が比較的大きいが，試行回数を増やすに従って変動は徐々に小さくなり，$T_i = 10^6$ 程度まで増やすと距離は 10^{-3} 程度になることが確認される．以上の理由から，国名辞書の試行回数は 100 万回を基準とした．

続いて，しりとりの終了文字（最後の単語の末尾文字）についての測定を行った．100 万回しりとりを実行したときに，終了文字として現れたのは a, y, o, q の 4 文字だけであった．さらに，全体の 9 割以上が a で終了する結果となった．ここで終了文字ごとの鎖長分布を定義しておこう．$p_\theta(L)$ は文字 θ で終了する鎖長の分布であり，

$$\sum_{\theta \in \Theta} p_\theta(L) = p(L) \tag{3.2}$$

が成り立つ．辞書を構成する単語の先頭文字と末尾文字の集合を Θ で表した．つまり，図 3.1 に示した鎖長分布 $p(L)$ は，異なる終了文字に対する鎖長分布の重ね合わせである．では，

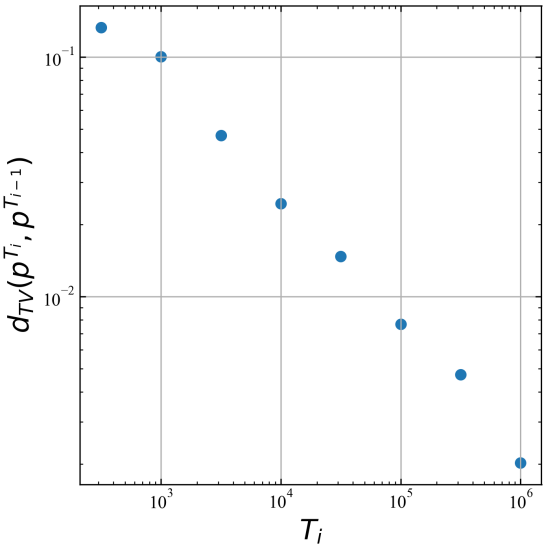


図 3.3: 国名辞書の試行回数 T_i, T_{i-1} の誤差 $d_{TV}(p^{T_i}, p^{T_{i-1}})$

終了文字ごとの鎖長分布はどのような形状をしているのだろうか．図 3.4 に国名辞書の終了文字 θ ごとの鎖長分布 $p_\theta(L)$ を示した．

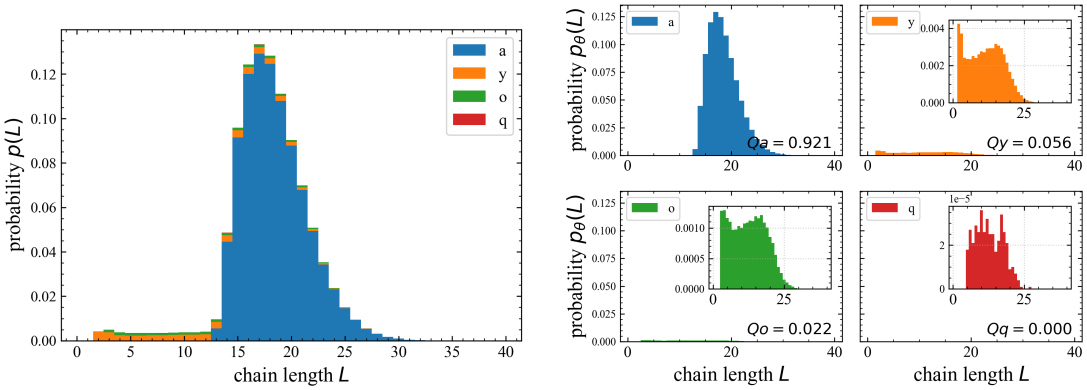
終了文字が a の分布 $p_{\mathrm{a}}(L)$ は，$L = 13$ で値を持ち始め，$L = 17$ にピークを持つ単峰性の分布となった．一方で，$p_{\mathrm{y}}(L)$ や $p_{\mathrm{o}}(L)$ は，L が 5 未満のところでいきなり最大値を持ち，その後 $L = 15$ 付近で再び極大値をとるような二峰性の分布を示した．このように終了文字ごとに鎖長分布の形状が異なっていることがわかる．したがって，図 3.1 の鎖長分布 $p(L)$ の形状を理解するためには，終了文字ごとの鎖長分布 $p_\theta(L)$ を求めればよい．本研究では終了文字別に分布の導出を試みる．関連する量として文字 θ の終了確率を定義しておこう．Q_θ はランダムなしりとりが文字 θ で終了する確率であり，

$$Q_\theta \coloneqq \sum_L p_\theta(L) \tag{3.3}$$

で定義される．このとき，

$$\sum_{\theta \in \Theta} Q_\theta = 1 \tag{3.4}$$

が成り立つ．表 3.1 に各文字の終了確率 Q_θ と，鎖長の最小値 $L_{\min}$，最大値 $L_{\max}$，平均値 L_{mean} をまとめた．例えば，q で終了する試行は 100 万回中わずか 37 回しか存在しない．つまり，しりとりの試行回数が 1 万回程度なら q で終了する単語列を発見できなかったかもしれない．逆に，試行回数をもっと増やせば a, y, o, q 以外の文字で終了するような単語列を見出す可能性がある．このように，試行回数が十分大きいかどうかは，鎖長分布の関数形に着目するか，すべての単語列を探索するかといった問題設定に依存すると考えられる．



(a) 終了文字ごとに色分けした分布　　(b) 終了文字ごとの分布

図 3.4: 国名辞書の終了文字ごとの鎖長分布（100 万回実測）

表 3.1: 国名辞書の終了文字ごとの終了確率 Q_θ および鎖長の最小値 $L_{\min}$，最大値 $L_{\max}$，平均値 L_{mean}（100 万回実測）

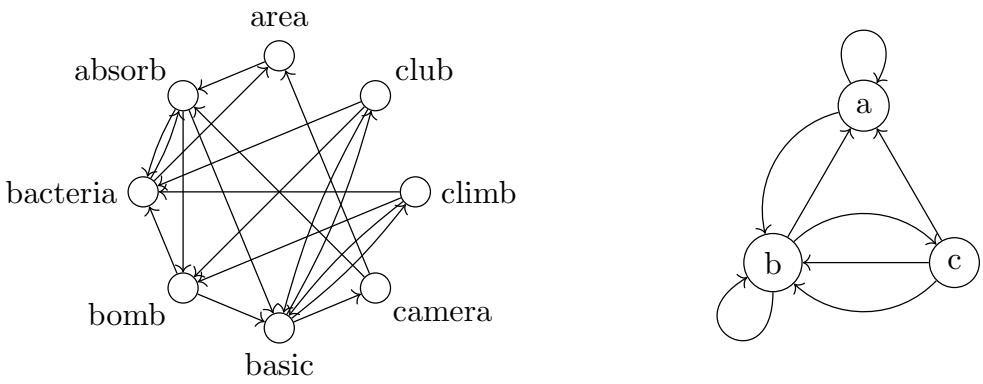
終了文字 θ	Q_θ	$L_{\min}$	$L_{\max}$	L_{mean}
a	0.921	13	39	18.5 ± 3.0
y	0.056	2	32	11.4 ± 6.1
o	0.022	3	32	12.5 ± 6.1
q	3.7×10^{-5}	5	27	12.7 ± 4.7
すべて	1	2	39	17.9 ± 3.8

3.2　ランダムなしりとりと辞書ネットワーク

辞書は 2 種類の辞書ネットワークを構成する．それぞれの構成方法を辞書 {area, absorb, bacteria, bomb, basic, camera, climb, club} ($D = 8, \Theta = \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$) を例にとって説明する．1 種類目は単語頂点型ネットワークである．これは単語を頂点として，しりとりで繋げられる関係を辺とする単純有向グラフである．文字 θ で終わる単語は文字 θ から始まるすべての単語への有向辺を持つ．例の場合，頂点 area からは頂点 absorb への有向辺があり，頂点 absorb からは 3 頂点 bacteria，bomb，basic への有向辺がある（図 3.5 (a)）．2 種類目は単語辺型ネットワークである．これは文字を頂点として，単語を辺とする多重有向グラフである．任意の文字 θ, ϕ に対して，θ で始まり ϕ で終わる単語が n 個あれば，θ から ϕ への有向辺を n 本張る．例の場合，area は頂点 a でループする辺，absorb は頂点 a から頂点 b への有向辺に対応する（図 3.5 (b)）[*1]．単語頂点型から単語辺型へは，頂点と辺を入れ替えて，

[*1] 単語頂点型では “area” や “bomb” などの先頭文字と末尾文字が一致する単語は自己ループを形成するが，後述する通過した頂点を回避するルールにより，このループは無視することができる．一方で，単語辺型の自

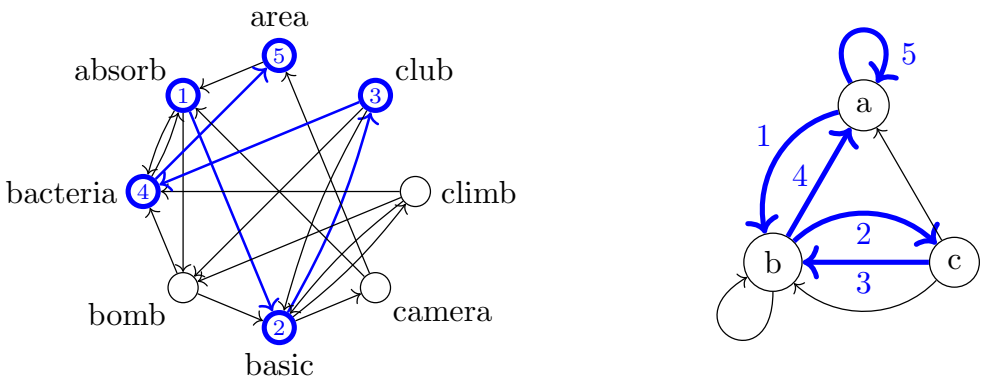
繋がっている文字が同じ頂点を同一視することで変換することができる[*2].



(a) 単語頂点型の辞書ネットワーク　　(b) 単語辺型の辞書ネットワーク

図 3.5: 2 種類の辞書ネットワーク

しりとりは辞書ネットワーク上の辺を伝いながら頂点間を移動する．そして単語は一度しか使えないので，一度通った単語頂点または単語辺を再び通ることはできない．したがってランダムに単語を選ぶしりとりは，単語頂点型上の自己回避ランダムウォーク（SAW）または単語辺型上の自己回避トレイル（SAT）とみなすことができる．例の辞書において，“absorb” → “basic” → “club” → “bacteria” → “aria” と続くしりとりをした場合，SAW や SAT の動きは図 3.6 のように表される．そして，単語頂点型では area から進める頂点を，単語辺型では頂点 a から出ていく辺を使い果たすので，しりとりは鎖長 $L = 5$ で終了することがわかる．なお，図中の数字はステップ数であり，しりとりの何単語目かを表す．



(a) 単語頂点型上の SAW　　(b) 単語辺型上の SAT

図 3.6: 辞書ネットワーク上のランダムウォークの例

このように辞書をネットワークとみなす方法は 2 種類あるが，どちらの方が扱いやすいだろうか．単語頂点型では，辞書の単語数が増えるほど頂点の数が増加する．一方で，単語辺

己ループは鎖長分布を計算する上で意味があり，無視することはできない．

[*2] この変換を 2 部グラフの射影として捉えることもできる（付録 C を参照）.

型は辞書の単語数によらず頂点の数は高々文字の種類数にしかならない．後述する各頂点の次数や隣接行列を用いたネットワーク解析を行うには，ネットワークの頂点数が少ない方が扱いやすくなる．そのため，本研究に用いる辞書ネットワークには単語辺型を採用する．

単語辺型の辞書ネットワークでは各文字が入次数（頂点に入ってくる辺の数）と出次数（頂点から出ていく辺の数）を持つ．しりとり開始時の文字 θ の入次数を $k_{\mathrm{in},\theta}$，出次数を $k_{\mathrm{out},\theta}$ と書く．つまり，$k_{\mathrm{in},\theta}$ は θ で終わる単語の数，$k_{\mathrm{out},\theta}$ は θ から始まる単語の数を意味している．先ほどの例における各文字の $k_{\mathrm{in}}, k_{\mathrm{out}}$ の関係を表 3.2 と図 3.7 に示した．またステップごとに各文字の次数は変化するため，l ステップ目における文字 θ の入次数を $k^{(l)}_{\mathrm{in},\theta}$，出次数を $k^{(l)}_{\mathrm{out},\theta}$ と表す．このとき，$k^{(0)}_{\mathrm{in},\theta} = k_{\mathrm{in},\theta}$, $k^{(0)}_{\mathrm{out},\theta} = k_{\mathrm{out},\theta}$ である．

文字	k_{in}	k_{out}
a	3	2
b	4	3
c	1	3

表 3.2: $k_{\mathrm{in}} - k_{\mathrm{out}}$ 表

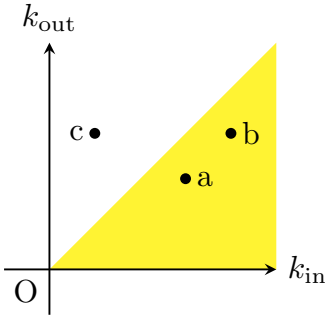


図 3.7: $k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面

しりとり開始時の辞書の総単語数を D, l ステップ目時点のしりとりで使える単語数を $D^{(l)}$ とする．このとき $D^{(0)} = D$ である．l ステップ目における入次数，出次数，しりとりで使える単語数の間には以下の関係式が成り立つ．

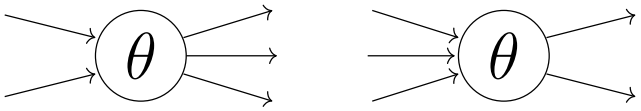
$$\sum_{\theta\in\Theta} k^{(l)}_{\mathrm{in},\theta} = \sum_{\theta\in\Theta} k^{(l)}_{\mathrm{out},\theta} = D^{(l)} = D - l. \tag{3.5}$$

しりとりの最後の単語の末尾文字を終了文字と呼ぶ．終了文字となるための条件を考えたい．文字 θ が図 3.8 (a) $k_{\mathrm{in},\theta} < k_{\mathrm{out},\theta}$ を満たす場合，頂点 θ に入って出ていくを繰り返した結果，入次数が先に 0 となり θ で終了することはない．一方で図 3.8 (b) $k_{\mathrm{in},\theta} > k_{\mathrm{out},\theta}$ を満たす場合，出次数が先に 0 となり θ で終了する可能性が生じる．また，$k_{\mathrm{in},\theta} = k_{\mathrm{out},\theta}$ の場合も，しりとりの最初の単語が θ から始まるときに，出次数が先に 0 となり θ で終了する可能性が生じる．よって，しりとりは次の式を満たす文字 θ で終了する．

$$k_{\mathrm{in},\theta} \geq k_{\mathrm{out},\theta}. \tag{3.6}$$

(3.6) 式を満たす文字を終了可能文字と呼び，満たさない文字を終了不能文字と呼ぶ．先ほどの例では (3.6) 式を満たすのは a と b であるため，しりとりは必ず a か b で終了する．

国名辞書の辞書ネットワークを図 3.9 に示す．しりとりに使われる文字の種類数を C で表すと，w と x は国名の最初と最後の文字に使われないので，国名辞書においては $C = 24$ である．よって，図 3.9 のネットワークは頂点数 $C = 24$，リンク数 $D = 193$ の多重有向グラ



(a) $k_{\mathrm{in},\theta} < k_{\mathrm{out},\theta}$　　(b) $k_{\mathrm{in},\theta} > k_{\mathrm{out},\theta}$

図 3.8: 終了可能性について

フとなっている．また，各文字の $k_{\mathrm{in}}, k_{\mathrm{out}}$ の関係は表 3.3，図 3.10 のようになった．この図表から終了可能文字，つまり (3.6) 式を満たす文字は黄色の領域にある 9 文字であることがわかる．100 万回の実測ではそのうち 4 文字しか終了文字として現れていない．なお，使われていない文字 w, x については第 5 章で触れる．

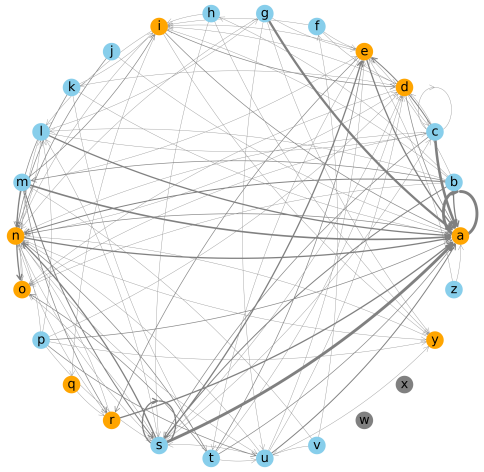


図 3.9: 国名辞書の辞書ネットワーク
辺の太さは有向辺の数に比例，頂点は青が終了可能文字，橙が終了不能文字を表す．孤立点 w, x は灰色で示している．

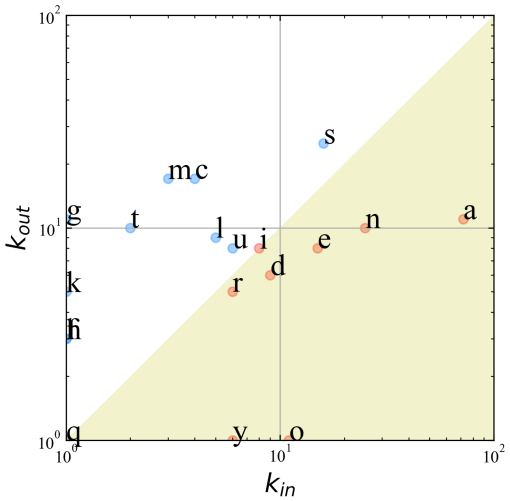


図 3.10: 国名辞書の $k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面
両対数のため，$k_{\mathrm{in}} = 0$ である b, j, p, v, z がプロットされていないことに注意．

表 3.3: 国名辞書の $k_{\mathrm{in}} - k_{\mathrm{out}}$ 表

文字	k_{in}	k_{out}	文字	k_{in}	k_{out}	文字	k_{in}	k_{out}	文字	k_{in}	k_{out}
a	72	11	g	1	11	m	3	17	s	16	25
b	0	17	h	1	3	n	25	10	t	2	10
c	4	17	i	8	8	o	11	1	u	6	8
d	9	6	j	0	3	p	0	9	v	0	3
e	15	8	k	1	5	q	1	1	y	6	1
f	1	3	l	5	9	r	6	5	z	0	2

3.3 単一辞書とシャッフル辞書群

多重有向グラフの隣接行列を定義する．ただし，有向辺 $\theta\phi$ とは頂点 θ から頂点 ϕ へ向かう有向辺のことである．

定義 3.3.1. N 頂点の多重有向グラフに対して，$\theta\phi$ 成分が有向辺 $\theta\phi$ の数であるような N 次正方行列 A を隣接行列と呼ぶ．

辞書ネットワークの隣接行列 A において，$A_{\theta\phi}$ は θ で始まり ϕ で終わる単語の数を表す．なお本論文では，隣接行列や後に登場する経路行列の要素をアルファベット順に並べるものとする．例えば，図 3.5 (b) の辞書ネットワークの隣接行列 A は次の 3×3 行列で表される．

$$A=\begin{pmatrix} A_{\mathrm{aa}} & A_{\mathrm{ab}} & A_{\mathrm{ac}} \\ A_{\mathrm{ba}} & A_{\mathrm{bb}} & A_{\mathrm{bc}} \\ A_{\mathrm{ca}} & A_{\mathrm{cb}} & A_{\mathrm{cc}} \end{pmatrix}=\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix}.$$

国名辞書の隣接行列 A も同様に求めることができ，次の 24×24 行列で表される．

$$A=\begin{pmatrix}
9&0&0&0&0&0&0&0&0&0&0&0&0&2&0&0&0&0&0&0&0&0&0&0\\
4&0&0&0&1&0&0&1&1&0&0&1&2&3&1&0&0&0&3&0&0&0&0&0\\
8&0&1&1&3&0&0&0&0&0&0&0&0&1&1&0&0&0&2&0&0&0&0&0\\
2&0&1&0&0&0&0&0&1&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0\\
4&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&2&0&1&0&0&0&0\\
0&0&0&1&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
7&0&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&1&0&1&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0\\
2&0&0&2&0&0&0&0&0&0&0&1&0&1&0&0&1&0&0&0&0&0&1&0\\
1&0&0&0&0&0&0&0&0&0&0&0&0&2&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&0&0&1&0&0&0&0&2&0&0&0&0&0&1&0&0&0&0\\
4&0&1&0&0&0&1&0&0&0&0&0&0&2&1&0&0&0&0&0&0&0&0&0\\
5&0&0&0&1&0&0&0&2&0&0&0&0&0&4&0&0&2&3&0&0&0&0&0\\
4&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&1&1&0&1&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
2&0&0&1&0&0&0&0&0&0&0&1&0&1&0&0&0&0&1&0&2&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
4&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
9&0&1&1&4&0&0&0&0&0&0&1&0&4&1&0&0&0&4&0&0&0&0&0\\
2&0&0&1&2&0&0&0&0&0&0&0&0&2&2&0&0&0&0&0&1&0&0&0\\
3&0&0&1&1&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&1&0\\
0&0&0&0&0&1&0&0&0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{pmatrix}.$$

ステップごとに隣接行列は変化するので，l ステップ目の隣接行列を $A^{(l)}$ と表す．ここで $A^{(0)}=A$ である．文字 θ の入次数 $k^{(l)}_{\mathrm{in},\theta}$，出次数 $k^{(l)}_{\mathrm{out},\theta}$ はそれぞれ隣接行列 $A^{(l)}$ の列和と行

和になっており，

$$k^{(l)}_{\mathrm{in},\theta} = \sum_{\phi\in\Theta} A^{(l)}_{\phi\theta} \tag{3.7}$$

$$k^{(l)}_{\mathrm{out},\theta} = \sum_{\phi\in\Theta} A^{(l)}_{\theta\phi} \tag{3.8}$$

と表される．例えば，例の辞書において $k_{\mathrm{in,a}} = A_{\mathrm{aa}} + A_{\mathrm{ba}} + A_{\mathrm{ca}} = 3$, $k_{\mathrm{out,a}} = A_{\mathrm{aa}} + A_{\mathrm{ab}} + A_{\mathrm{ac}} = 2$ が成り立つ．

単一辞書（1 つの辞書）が与えられたときのランダムなしりとりの一回の試行は，隣接行列を用いて以下のように記述できる．

1. 0 番目の単語の末尾文字をすべての文字の中から等確率に選択する．
2. l 番目の単語の末尾文字が θ のとき，$l+1$ 番目の単語の末尾文字に ϕ を選択する確率 $q^{(l)}_{\theta\phi}$ を以下のように定める．

$$q^{(l)}_{\theta\phi} = \frac{A^{(l)}_{\theta\phi}}{\sum_{\phi'\in\Theta} A^{(l)}_{\theta\phi'}} = \frac{A^{(l)}_{\theta\phi}}{k^{(l)}_{\mathrm{out},\theta}} \tag{3.9}$$

   単語選択後の隣接行列 $A^{(l+1)}$ は $\theta\phi$ 成分のみが 1 減少し，他の成分は変化しない．
3. いずれかの文字 θ で $k^{(l)}_{\mathrm{out},\theta} = 0$ となった後，末尾文字に θ を選択したらしりとりは終了し，そのときの単語列の長さ L を鎖長とする．

つまり，図 3.1, 3.4 の特徴的な分布は前のページの隣接行列で決まるのである．(3.9) 式は「末尾文字 ϕ を選択する確率」が「θ で始まる全単語の中から，θ で始まり ϕ で終わる単語を選択する確率」に等しいことを意味する．もし同じ単語を選んでもよいのなら，$A^{(l)}$ は定常な隣接行列となり，いわゆるマルコフ連鎖で解析できるようになる．しかし今の場合，$A^{(l)}$ は非定常な隣接行列であるため，$q^{(l)}_{\theta\phi}$ は過去のすべての選択に依存して決まることになる．したがって，鎖長 L かつ文字 θ で終了する確率 $p_\theta(L)$ を解析的に求めるのは容易ではないと考えられる．

以上の設定についての解析は第 5 章に回すとして，ここからは「末尾文字 ϕ を選択する確率」を「すべての単語の中から，ϕ で終わる単語を選択する確率」で近似することを考える．こうすることで，しりとりを非復元抽出型の壺モデル（詳しくは第 4 章に記述）として扱うことができ，鎖長分布を解析的に求められるようになる．この近似を平均場近似と呼び，平均場近似を施して得られる鎖長分布を $\tilde{p}_\theta(L)$ と表す．具体的な近似方法は，(3.9) 式を分母分子それぞれ θ について和をとったもの

$$\tilde{q}^{(l)}_{\theta\phi} = \frac{\sum_{\theta\in\Theta} A^{(l)}_{\theta\phi}}{\sum_{\theta\in\Theta}\sum_{\phi'\in\Theta} A^{(l)}_{\theta\phi'}} = \frac{k^{(l)}_{\mathrm{in},\phi}}{D^{(l)}} \tag{3.10}$$

で置き換える．$\tilde{q}^{(l)}_{\theta\phi}$ は隣接行列の各成分には依存せず，その列和すなわち $k^{(l)}_{\mathrm{in},\theta}$ にのみ依存する．したがって，$\tilde{p}_\theta(L)$ を求めるにあたっては，各文字の入次数と出次数のみが必要となる．

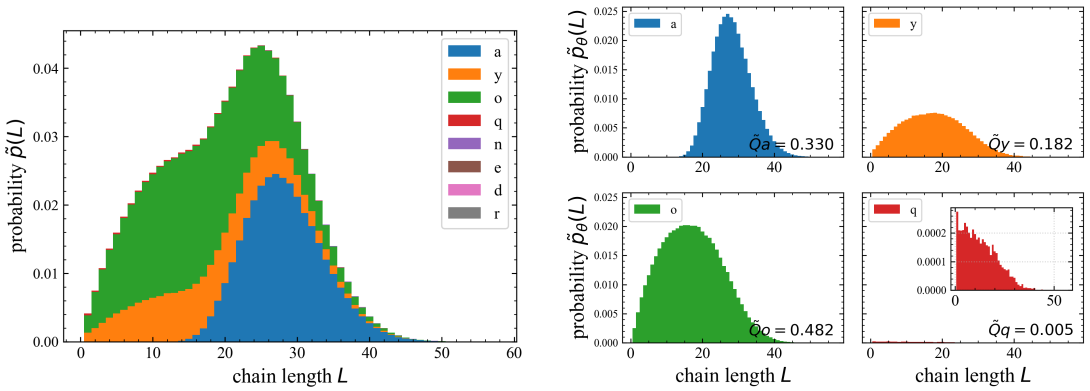
$\tilde{p}_\theta(L)$ の意味について説明する．与えられた辞書の各文字の入次数と出次数を保ったまま，頂点同士をランダムに繋ぎかえてできる辞書をシャッフル辞書と呼ぶ．シャッフル辞書を多数作成し，それぞれで $p_\theta(L)$ を求め，その平均をとった分布が $\tilde{p}_\theta(L)$ である．

平均場近似を施すと鎖長分布はどのようになるのだろうか．国名辞書のシャッフル辞書を 10000 冊作成し，各辞書で 100 回しりとりを実行したときの平均鎖長分布は図 3.11 のようになった．元の国名辞書の終了文字は a, y, o, q の 4 文字だったが，平均場近似を施した場合に出現した終了文字は a, y, o, q, n, e, d, r の 8 文字と大幅に増加した．さらに，単一辞書では二峰性分布であった y, o の分布が，シャッフル辞書群では単峰性分布に変化するなど，分布の形状にも大きな変化が見られる．シャッフル辞書群に対しても，(3.3) 式と同様に文字 θ で終了する確率 $\tilde{Q}_\theta$ を定義することができる．

$$\tilde{Q}_\theta \coloneqq \sum_L \tilde{p}_\theta(L), \tag{3.11}$$

$$\sum_{\theta\in\Theta} \tilde{Q}_\theta = 1. \tag{3.12}$$

各終了文字における $\tilde{Q}_\theta$, $L_{\min}$, $L_{\max}$, L_{mean} の値は表 3.4 のようになった．単一辞書ではほとんどが a で終了していたが，シャッフル辞書群では a より o で終了する回数の方が多くなっている．さらに，どの文字も鎖長のばらつきや最大値が単一辞書に比べて大きくなっていることがわかる．このように，各文字の鎖長分布や終了確率は元の国名辞書のものと大きく異なってしまう．しかし，平均場近似を施すことで鎖長分布を解析的に求めることができるようになるのである．



(a) 終了文字ごとに色分けした分布　　(b) 終了文字ごとの分布（a, y, o, q のみ表示）

図 3.11: 国名辞書のシャッフル辞書群の平均鎖長分布（1 万冊 × 各 100 回実測）

次章では (3.10) 式の平均場近似を採用した問題設定に関する解析結果を述べる．(3.9) 式を採用した場合の解析結果は第 5 章で述べる．

表 3.4: 国名辞書のシャッフル辞書群の終了文字ごとの終了確率 $\tilde{Q}_\theta$ および鎖長の最小値 $L_{\min}$，最大値 $L_{\max}$，平均値 L_{mean}（1 万冊 × 各 100 回実測）

終了文字 θ	$\tilde{Q}_\theta$	$L_{\min}$	$L_{\max}$	L_{mean}
o	0.482	1	53	17.2 ± 8.4
a	0.330	13	56	28.2 ± 5.4
y	0.182	1	55	17.6 ± 8.5
q	4.6×10^{-2}	1	48	12.7 ± 8.5
n	1.0×10^{-2}	18	57	36.2 ± 5.6
e	1.6×10^{-3}	19	49	36.5 ± 5.6
d	5.7×10^{-5}	18	49	31.9 ± 6.8
r	2.0×10^{-5}	16	43	27.7 ± 7.6
すべて	1	1	57	20.9 ± 9.1

第 4 章

シャッフル辞書群

平均場近似により，ランダムなしりとりは非復元抽出型の壺モデルとして扱うことができるようになる．本章では統計学的な準備を行った後，シャッフル辞書群の鎖長分布，およびその分布統計量を理論的に求める．なお今後は入次数と出次数をベクトルで扱う．つまり，文字の集合を $\Theta = \{\theta_1, \theta_2, \cdots, \theta_C\}$ としたとき，入次数ベクトルを $\overrightarrow{k_{\text{in}}} = (k_{\text{in},\theta_1}, k_{\text{in},\theta_2}, \cdots, k_{\text{in},\theta_C})$，出次数ベクトルを $\overrightarrow{k_{\text{out}}} = (k_{\text{out},\theta_1}, k_{\text{out},\theta_2}, \cdots, k_{\text{out},\theta_C})$ とする．

4.1 数学的準備

平均場近似を施すと，前の単語がどの文字で終わったかを考慮する必要がなくなる．つまり，l ステップ目に文字 θ で終わる単語を選ぶ確率は，前の末尾文字に関係なく $k^{(l)}_{\text{in},\theta}/D^{(l)}$ で表される．そして，θ で終わる単語を θ で始まる単語の数より多く選んだらしりとりは終了する．この過程は以下で説明する非復元抽出型の壺モデルに単純化することができる．非復元抽出とは取り出した球を壺に戻さないという意味である．

1. 壺の中に C 種類の文字が書かれた球を入れる．このとき，各文字 θ について，θ と書かれた球は $k_{\text{in},\theta}$ 個入れる．
2. 壺の中から非復元抽出で球を 1 つずつ取り出す．
3. いずれかの文字 θ で，θ と書かれた球を $k_{\text{out},\theta} + 1$ 個取り出したら終了する．

このモデルで l ステップ目に θ と書かれた球を取り出す確率は $k^{(l)}_{\text{in},\theta}/D^{(l)}$ と表され，シャッフル辞書群における単語の選択確率に一致することがわかる．第 1 章の辞書を例にとると，各文字の次数（表 3.2）から対応する壺は図 4.1 のようになり，a の球を 3 個取り出すか，b の球を 4 個取り出したら終了する．

本章で壺モデルを用いた解析を行うために，いくつかの確率分布を定義する．まずは要素の種類数が 2 の場合の非復元抽出について考える．

定義 4.1.1. 2 種類の要素 θ_1, θ_2 がそれぞれ n_1 個，$N - n_1$ 個ある合計 N 個の母集団から，

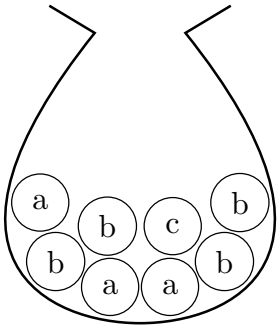


図 4.1: 非復元抽出型の壺モデル

M 個を非復元抽出で取り出したとする．このとき，抽出されたサンプルに含まれる θ_1 の個数 m_1 が従う確率分布を超幾何分布 (hypergeometric distribution) と呼ぶ.

超幾何分布の確率質量関数 $HG(m_1; N, n_1, M)$ は次で与えられる.

$$HG(m_1; N, n_1, M) = \frac{\binom{n_1}{m_1}\binom{N-n_1}{M-m_1}}{\binom{N}{M}}. \tag{4.1}$$

超幾何分布では試行回数を指定したが，代わりに一方の要素が抽出される回数を指定すると負の超幾何分布が得られる.

定義 4.1.2. 2 種類の要素 θ_1, θ_2 がそれぞれ $N - n_2$ 個，n_2 個ある合計 N 個の母集団から，非復元抽出を繰り返したとする．このとき，θ_1 が m_1 個取り出されるまでに抽出された θ_2 の個数 m_2 が従う確率分布を負の超幾何分布 (negative hypergeometric distribution) と呼ぶ.

負の超幾何分布の確率質量関数 $NHG(m_2; N, n_2, m_1)$ は，$m_1 + m_2 - 1$ 回目までに θ_1 が $m_1 - 1$ 個選択されて，かつ $m_1 + m_2$ 回目に θ_1 が選択される確率と考えられる．したがって次のように書ける.

$$\begin{aligned} NHG(m_2; N, n_2, m_1) &= HG(m_1 - 1; N, N - n_2, m_1 + m_2 - 1) \times \frac{(N - n_2) - (m_1 - 1)}{N - (m_1 + m_2 - 1)} \\ &= \frac{\binom{m_2+m_1-1}{m_2}\binom{N-m_2-m_1}{n_2-m_2}}{\binom{N}{n_2}}. \end{aligned} \tag{4.2}$$

続いて，要素の種類数が 2 以上の場合の非復元抽出を考える．まずは超幾何分布の多変量化を行う.

定義 4.1.3. C 種類 ($C \geq 2$) の要素 $\theta_1, \theta_2, \cdots, \theta_C$ がそれぞれ n_1 個，n_2 個，$\cdots$，n_C 個ある合計 N 個の母集団から，M 個を非復元抽出で取り出したとする．このとき，抽出されたサンプルに含まれる $\theta_1, \theta_2, \cdots, \theta_C$ の個数 $m_1, m_2, \cdots, m_C$ が従う確率分布を多変量超幾何分布 (multivariate hypergeometric distribution) と呼ぶ.

多変量超幾何分布の確率質量関数 $MHG(\{m_1, m_2, \cdots, m_C\}; N, \{n_1, n_2, \cdots, n_C\}, M)$ は次

で与えられる．

$$MHG(\{m_1, m_2, \cdots, m_C\}; N, \{n_1, n_2, \cdots, n_C\}, M) = \frac{\prod_{i=1}^{C} \binom{n_i}{m_i}}{\binom{N}{M}}. \tag{4.3}$$

$C = 2$ の場合，超幾何分布の確率質量関数に一致することがわかる．

次に負の超幾何分布を多変量化したい．(4.2) 式により超幾何分布から負の超幾何分布を記述することができたように，同じような関係式により多変量超幾何分布から新たな確率分布を定義する．

定義 4.1.4. C 種類 ($C \geq 2$) の要素 $\theta_1, \theta_2, \cdots, \theta_C$ がそれぞれ n_1 個，n_2 個，$\cdots$，n_C 個ある合計 N 個の母集団から，非復元抽出を繰り返したとする．このとき，θ_1 が m_1 個取り出されるまでに抽出された $\theta_2, \theta_3, \cdots, \theta_C$ の個数 $m_2, m_3, \cdots, m_C$ が従う確率分布を多変量逆超幾何分布 (multivariate inverse hypergeometric distribution)[*1] と呼ぶ．

多変量逆超幾何分布の確率質量関数 $MIHG(\{m_2, m_3, \cdots, m_C\}; N, \{n_2, n_3, \cdots, n_C\}, m_1)$ は，$M - 1$ 回目までに θ_1 が $m_1 - 1$ 個，θ_2 が m_2 個，$\cdots$，θ_C が m_C 個選択されて，かつ M 回目に θ_1 が選択される確率と考えられる．ここで，$M := \sum_{i=1}^{C} m_i$ とした．したがって次のように書ける．

$$\begin{aligned}
&MIHG(\{m_2, m_3, \cdots, m_C\}; N, \{n_2, n_3, \cdots, n_C\}, m_1) \\
&= MHG(\{m_1 - 1, m_2, \cdots, m_C\}; N, \{n_1, n_2, \cdots, n_C\}, M - 1) \times \frac{n_1 - (m_1 - 1)}{N - (M - 1)} \\
&= \frac{\binom{n_1}{m_1 - 1} \prod_{i=2}^{C} \binom{n_i}{m_i}}{\binom{N}{M-1}} \times \frac{n_1 - (m_1 - 1)}{N - (M - 1)}.
\end{aligned} \tag{4.4}$$

$C = 2$ の場合，負の超幾何分布の確率質量関数に一致することがわかる．

非復元抽出に関する 4 つの確率分布は以下のようにまとめられる．

表 4.1: 非復元抽出の確率分布

	試行回数固定	θ_1 の抽出回数固定
要素 2 種類	超幾何分布	負の超幾何分布
要素 C 種類	多変量超幾何分布	多変量逆超幾何分布

4.2 $C = 2$

準備が整ったので，文字の種類数 C が 2 のシャッフル辞書群における鎖長分布および分布統計量の求め方を説明する．$\Theta = \{x, z\}$ とし，$k_{\mathrm{in},x} > k_{\mathrm{out},x}$ であるとする．つまり，しりと

[*1] 意味合いとしては多変量負の超幾何分布 (multivariate negative hypergeometric distribution) の方がわかりやすいが，[33] によると多変量逆超幾何分布とは異なる定義でこの確率分布が存在している．

りは必ず x で終了する．

4.2.1 解析計算

鎖長分布 $\tilde{p}_x(L)$ の求め方について説明する．4.1 節の壺モデルを用いるのだが，初期条件として選択される文字，すなわち 1 単語目の最初の文字によって終了条件が変化する点に注意が必要である．最初に文字 z を選択した場合，x で終わる単語を $k_{\mathrm{out},x}+1$ 個選択したらしりとりは終了する．一方で最初に文字 x を選択した場合，x から始まる単語が 1 つ減るので，x で終わる単語を $k_{\mathrm{out},x}$ 個選択したらしりとりは終了する．これを踏まえて，$C=2$ の場合の壺モデルは以下のように書ける．ただし，l ステップ目の単語を $w^{(l)}$，$w^{(l)}$ の末尾文字を $\theta^{(l)}$ とする．

1. 壺の中に x と書かれた球を $k_{\mathrm{in},x}$ 個，z と書かれた球を $k_{\mathrm{in},z}$ 個入れる．
2. 壺の中から非復元抽出で球を 1 つずつ取り出す．
3. x と書かれた球を $\theta^{(0)}=x$ のとき $k_{\mathrm{out},x}$ 個，$\theta^{(0)}=z$ のとき $k_{\mathrm{out},x}+1$ 個取り出したら終了する．

つまり，しりとりが鎖長 L で終了する確率は，x と書かれた球を $k_{\mathrm{out},x}$ 個または $k_{\mathrm{out},x}+1$ 個取り出すまでに，z と書かれた球を $L-k_{\mathrm{out},x}$ 個または $L-(k_{\mathrm{out},x}+1)$ 個取り出す確率と考えられる．これは負の超幾何分布を用いて表される．初期条件として x,z が選択される確率はいずれも $1/2$ であるから，$\tilde{p}_x(L)$ は以下のように書くことができる[*2]．

$$\begin{aligned}\tilde{p}_x(L) &= \frac{1}{2}NHG(L-k_{\mathrm{out},x};D,D-k_{\mathrm{in},x},k_{\mathrm{out},x})\\ &\quad+\frac{1}{2}NHG(L-k_{\mathrm{out},x}-1;D,D-k_{\mathrm{in},x},k_{\mathrm{out},x}+1)\\ &=\frac{1}{2}\frac{\binom{L-1}{k_{\mathrm{out},x}-1}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}}}{\binom{D}{k_{\mathrm{in},x}}}+\frac{1}{2}\frac{\binom{L-1}{k_{\mathrm{out},x}}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}-1}}{\binom{D}{k_{\mathrm{in},x}}}.\end{aligned}\tag{4.5}$$

ただし，$k_{\mathrm{out},x}=0$ のときなどに現れる，上に整数 n，下に -1 がくる二項係数の値を [34] に倣って以下のように定義する．

$$\binom{n}{-1}=\begin{cases}0 & \text{if } n\geq 0,\\ 1 & \text{if } n=-1.\end{cases}\tag{4.6}$$

負の超幾何分布はしりとりの他にも様々な場面で登場する．例えば，心理検査の得点 [35] や鳥が記憶した餌箱に訪れるまでに探索する空箱の数 [36]，スロットマシンのボーナスゲームで獲得するスコア [37] に負の超幾何分布が現れることが報告されている．

[*2] 初期条件として辞書からランダムに単語を選択する場合，$\tilde{p}_x(L)$ は以下のように表される．

$$\tilde{p}_x(L)=\frac{k_{\mathrm{out},x}}{D}\frac{\binom{L-1}{k_{\mathrm{out},x}-1}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}}}{\binom{D}{k_{\mathrm{in},x}}}+\frac{D-k_{\mathrm{out},x}}{D}\frac{\binom{L-1}{k_{\mathrm{out},x}}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}-1}}{\binom{D}{k_{\mathrm{in},x}}}.$$

4.2.2 数値計算

4.2.1 節で得られた鎖長分布が正しいかどうかを，具体的な辞書「xz シャッフル辞書」を用いた数値計算により確認する．xz シャッフル辞書は $\Theta = \{x, z\}$，$D = 100$ のシャッフル辞書で，次数ベクトルは以下のように設定される．

$$\overrightarrow{k_{\mathrm{in}}} = (k_{\mathrm{in},x}, k_{\mathrm{in},z}) = (70, 30),$$
$$\overrightarrow{k_{\mathrm{out}}} = (k_{\mathrm{out},x}, k_{\mathrm{out},z}) = (20, 80).$$

この設定からしりとりは x で終了することがわかる．

xz シャッフル辞書を 10000 冊作成し，各辞書 100 回ずつしりとりを実行した．このときの鎖長分布の実測値と，(4.5) 式から計算される鎖長分布の理論値を比較した結果が図 4.2 である．実測値は理論値によく一致していることがわかる．

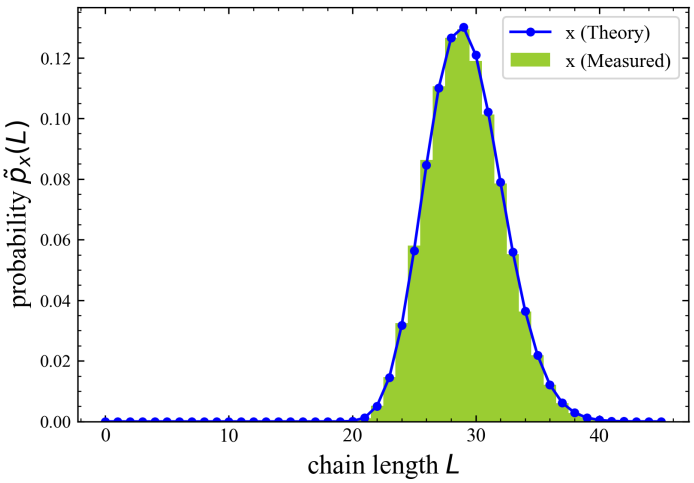


図 4.2: xz シャッフル辞書の鎖長分布（ヒストグラム：実測値，実線：理論値）

シャッフル辞書を 10000 冊用意した理由は次のとおりである．シャッフル辞書は $\overrightarrow{k_{\mathrm{in}}}$, $\overrightarrow{k_{\mathrm{out}}}$ が元の辞書と一致していればよいので，ランダムに選ばれたシャッフル辞書はそれぞれの特徴を持つ．しかし，平均場近似で予言しているのは，それらの平均的な性質である．図 4.3 は xz シャッフル辞書の冊数 N を変化させたときの，鎖長分布の実測値 $\tilde{p}_x^N(L)$ と理論値 $\tilde{p}_x(L)$ の L^2 ノルム $d_2(\tilde{p}_x^N, \tilde{p}_x)$ である．L^2 ノルム $d_2(f, g)$ は，2 つの確率質量関数 $f(x), g(x)$ に対して次のように定義される距離である．

$$d_2(f, g) := \sqrt{\sum_x (f(x) - g(x))^2}. \tag{4.7}$$

冊数が大きくなるほど距離は減少し続け，10000 冊付近では個別の辞書の特徴がほとんど現れなくなる．したがって，シャッフル辞書を 10000 冊用意すれば十分であると考えられる．

さらに，xz シャッフル辞書の冊数が大きい極限では，実測値が理論値に漸近することが予測される．

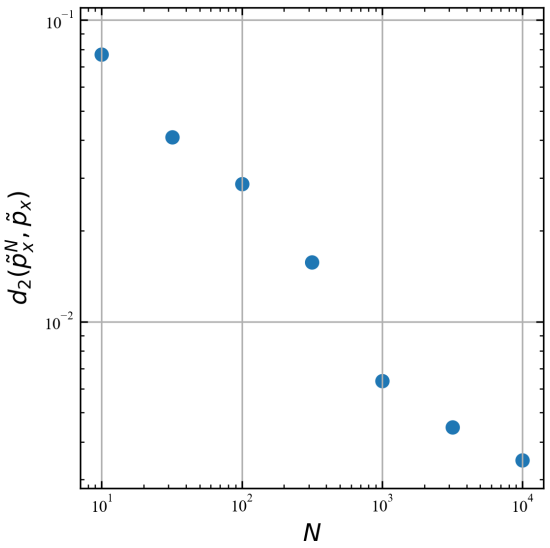


図 4.3: xz シャッフル辞書の冊数 N と理論値との距離 $d_2(\tilde{p}_x^N, \tilde{p}_x)$ の関係

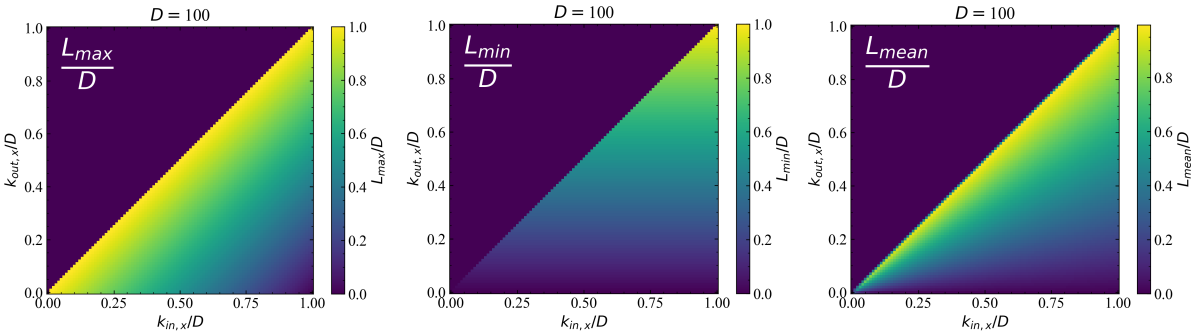
4.2.3 分布統計量

4.2.1 節で得られた鎖長分布をもとに，$C = 2$ シャッフル辞書群における鎖長の最大値，最小値，平均値を計算する．ただし，ここでは鎖長 L の代わりに，カバーレート L/D を用いる．カバーレートはしりとりの単語列が辞書内の単語を網羅している割合を表す．

まずは数値計算により予想を立てよう．図 4.4 はカバーレートの最大値 $L_{\max}/D$，最小値 $L_{\min}/D$，平均値 L_{mean}/D を各点 $(k_{\mathrm{in},x}/D, k_{\mathrm{out},x}/D)$ ごとに (4.5) 式から求めたヒートマップである．この図から，それぞれの等高線は (a) が $k_{\mathrm{out},x}/D = k_{\mathrm{in},x}/D$ に平行であり，(b) が $k_{\mathrm{out},x}/D =$ 一定であり，(c) が原点を通り傾き一定の直線に沿っているように見える．つまり，最小値は $k_{\mathrm{out},x}$ のみによって決まり，最大値と平均値はそれぞれ $k_{\mathrm{out},x}$ と $k_{\mathrm{in},x}$ の差と比によって決まることが予想される．本節ではこの予想が正しいことを解析的に導き出す．

初めにカバーレートの最大値 $L_{\max}/D$，最小値 $L_{\min}/D$ を求める．負の超幾何分布の確率質量関数 $NHG(m_2; N, n_2, m_1)$ は，確率変数 m_2 が $m_2 = 0, 1, \cdots, n_2$ のとき正の値をとるので，$\tilde{p}_x(L)$ が正の値をとるのは $L - k_{\mathrm{out},x} = 0, 1, \cdots, D - k_{\mathrm{in},x}$ のとき，または $L - k_{\mathrm{out},x} - 1 = 0, 1, \cdots, D - k_{\mathrm{in},x}$ のときである．よって鎖長 L のとりうる範囲は，

$$L = k_{\mathrm{out},x}, k_{\mathrm{out},x} + 1, \cdots, k_{\mathrm{out},x} + D - k_{\mathrm{in},x} + 1 \tag{4.8}$$



(a) $L_{\max}/D$ (b) $L_{\min}/D$ (c) L_{mean}/D

図 4.4: カバーレートの統計量のヒートマップ ($D = 100$)

である．したがって，カバーレートの最大値と最小値は,

$$\frac{L_{\max}}{D} = 1 + \frac{k_{\mathrm{out},x} - k_{\mathrm{in},x} + 1}{D}, \tag{4.9}$$

$$\frac{L_{\min}}{D} = \frac{k_{\mathrm{out},x}}{D} \tag{4.10}$$

と求められる．

続いて，カバーレートの平均値 L_{mean}/D を求める．その際に負の超幾何分布が (4.11) 式のように規格化されていることと，平均値が (4.12) 式で与えられることを用いる．これらの証明は付録 D に記載している．

$$\sum_{m_2=0}^{n_2} NHG(m_2; N, n_2, m_1) = 1, \tag{4.11}$$

$$E[m_2] \coloneqq \sum_{m_2=0}^{n_2} m_2 \cdot NHG(m_2; N, n_2, m_1) = \frac{m_1 n_2}{N - n_2 + 1}. \tag{4.12}$$

(4.5) 式を用いて鎖長の平均は以下のように計算される．

$$
\begin{aligned}
&L_{\mathrm{mean}} = \sum_{L} L \cdot \tilde{p}_x(L) \\
&= \frac{1}{2} \sum_{L-k_{\mathrm{out},x}=0}^{D-k_{\mathrm{in},x}} L \cdot NHG(L-k_{\mathrm{out},x}; D, D-k_{\mathrm{in},x}, k_{\mathrm{out},x}) \\
&\quad + \frac{1}{2} \sum_{L-k_{\mathrm{out},x}-1=0}^{D-k_{\mathrm{in},x}} L \cdot NHG(L-k_{\mathrm{out},x}-1; D, D-k_{\mathrm{in},x}, k_{\mathrm{out},x}+1) \\
&= \frac{1}{2} \sum_{L-k_{\mathrm{out},x}=0}^{D-k_{\mathrm{in},x}} (L-k_{\mathrm{out},x}) \cdot NHG(L-k_{\mathrm{out},x}; D, D-k_{\mathrm{in},x}, k_{\mathrm{out},x}) \\
&\quad + \frac{1}{2}\, k_{\mathrm{out},x} \sum_{L-k_{\mathrm{out},x}=0}^{D-k_{\mathrm{in},x}} NHG(L-k_{\mathrm{out},x}; D, D-k_{\mathrm{in},x}, k_{\mathrm{out},x}) \\
&\quad + \frac{1}{2} \sum_{L-k_{\mathrm{out},x}-1=0}^{D-k_{\mathrm{in},x}} (L-k_{\mathrm{out},x}-1) \cdot NHG(L-k_{\mathrm{out},x}-1; D, D-k_{\mathrm{in},x}, k_{\mathrm{out},x}+1) \\
&\quad + \frac{1}{2}(k_{\mathrm{out},x}+1) \sum_{L-k_{\mathrm{out},x}-1=0}^{D-k_{\mathrm{in},x}} NHG(L-k_{\mathrm{out},x}-1; D, D-k_{\mathrm{in},x}, k_{\mathrm{out},x}+1) \\
&= \frac{1}{2}\left\{ \frac{k_{\mathrm{out},x}(D-k_{\mathrm{in},x})}{k_{\mathrm{in},x}+1} + k_{\mathrm{out},x} \right\} + \frac{1}{2}\left\{ \frac{(k_{\mathrm{out},x}+1)(D-k_{\mathrm{in},x})}{k_{\mathrm{in},x}+1} + (k_{\mathrm{out},x}+1) \right\} \\
&= \frac{(D+1)(2k_{\mathrm{out},x}+1)}{2(k_{\mathrm{in},x}+1)}.
\end{aligned}
$$

4 つ目の等号で (4.11), (4.12) 式を用いた．したがってカバーレートの平均値は,

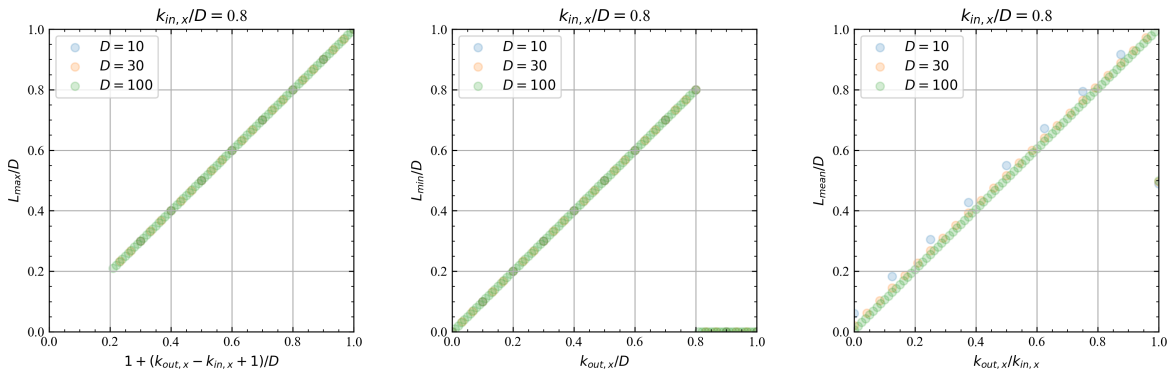
$$
\frac{L_{\mathrm{mean}}}{D} = \frac{D+1}{D} \cdot \frac{2k_{\mathrm{out},x}+1}{2k_{\mathrm{in},x}+2} \tag{4.13}
$$

と求められる．よって $D \gg 1$, $k_{\mathrm{in},x} \gg 1$, $k_{\mathrm{out},x} \gg 1$ を仮定すると,

$$
\frac{L_{\mathrm{mean}}}{D} \simeq \frac{k_{\mathrm{out},x}}{k_{\mathrm{in},x}} \tag{4.14}
$$

が得られる．(4.9), (4.10), (4.14) 式はヒートマップによる予想と一致する．

以上の結果が正しいことを数値的に確認する．$k_{\mathrm{in},x}/D = 0.8$ として (4.5) 式を計算し，$1 + (k_{\mathrm{out},x} - k_{\mathrm{in},x} + 1)/D$ に対する L_{max}/D，$k_{\mathrm{out},x}/D$ に対する L_{min}/D，$k_{\mathrm{out},x}/k_{\mathrm{in},x}$ に対する L_{mean}/D をプロットした結果が図 4.5 である．辞書のサイズに依らず (4.9), (4.10) 式が成り立つことが確認され，また辞書のサイズが大きい場合に (4.14) 式が成り立つことが確認される．



(a) $L_{\max}/D$ の次数依存性　(b) $L_{\min}/D$ の次数依存性　(c) L_{mean}/D の次数依存性

図 4.5: カバーレートの統計量と次数の関係 ($k_{\mathrm{in},x}/D = 0.8$)

4.3 $C \geq 3$

続いて，$C \geq 3$ のシャッフル辞書群における鎖長分布の求め方を説明し，分布統計量に関する考察を述べる．ここでは特に $C = 3$ に着目するが，$C > 3$ の場合も同様に求められる．$\Theta = \{x, y, z\}$ とし，$k_{\mathrm{in},x} > k_{\mathrm{out},x}$, $k_{\mathrm{in},y} > k_{\mathrm{out},y}$ であるとする．つまり，しりとりは必ず x か y で終了する．

4.3.1 解析計算

終了文字ごとの鎖長分布は $\tilde{p}_x(L)$ と $\tilde{p}_y(L)$ の 2 つが得られる．まずは $\tilde{p}_x(L)$ の求め方について説明する．4.2.1 節と同様に初期条件として選択される文字で場合分けを行うと，壺モデルは以下のように書ける．

1. 壺の中に x と書かれた球を $k_{\mathrm{in},x}$ 個，y と書かれた球を $k_{\mathrm{in},y}$ 個，z と書かれた球を $k_{\mathrm{in},z}$ 個入れる．
2. 壺の中から非復元抽出で球を 1 つずつ取り出す．
3. x と書かれた球を $\theta^{(0)} = x$ のとき $k_{\mathrm{out},x}$ 個，それ以外のとき $k_{\mathrm{out},x} + 1$ 個取り出したら終了する．

つまり，しりとりが鎖長 L（$L > 0$）かつ文字 x で終了する確率は，x と書かれた球を $k_{\mathrm{out},x}$ 個または $k_{\mathrm{out},x} + 1$ 個取り出した時点で，合計で L 個の球が取り出されている確率と考えられる．ここで，y, z と書かれた球をそれぞれ k_y, k_z 個取り出すとすると，$k_{\mathrm{out},x} + k_y + k_z$ または $k_{\mathrm{out},x} + 1 + k_y + k_z$ は L に等しくなる．そして k_y の範囲は，$\theta^{(0)} = y$ のとき $0 \leq k_y \leq k_{\mathrm{out},y} - 1$，それ以外のとき $0 \leq k_y \leq k_{\mathrm{out},y}$ をとり，k_z の範囲は，$\theta^{(0)} = z$ のとき $0 \leq k_z \leq k_{\mathrm{out},z} - 1$，それ以外のとき $0 \leq k_z \leq k_{\mathrm{out},z}$ をとる．よって $\tilde{p}_x(L)$ は $L > 0$ のと

き，多変量逆超幾何分布を用いて以下のように表すことができる[*3]．

$$
\begin{aligned}
\tilde{p}_x(L) &= \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}}\sum_{k_z=0}^{k_{\mathrm{out},z}} MIHG(\{k_y,k_z\};D,\{k_{\mathrm{in},y},k_{\mathrm{in},z}\},k_{\mathrm{out},x})\delta_{k_{\mathrm{out},x}+k_y+k_z,L} \\
&\quad + \frac{1}{3}\sum_{k_y=1}^{k_{\mathrm{out},y}}\sum_{k_z=0}^{k_{\mathrm{out},z}} MIHG(\{k_y-1,k_z\};D,\{k_{\mathrm{in},y},k_{\mathrm{in},z}\},k_{\mathrm{out},x}+1)\delta_{k_{\mathrm{out},x}+k_y+k_z,L} \\
&\quad + \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}}\sum_{k_z=1}^{k_{\mathrm{out},z}} MIHG(\{k_y,k_z-1\};D,\{k_{\mathrm{in},y},k_{\mathrm{in},z}\},k_{\mathrm{out},x}+1)\delta_{k_{\mathrm{out},x}+k_y+k_z,L} \\
&= \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}} MIHG(\{k_y,L-k_{\mathrm{out},x}-k_y\};D,\{k_{\mathrm{in},y},k_{\mathrm{in},z}\},k_{\mathrm{out},x}) \\
&\quad + \frac{1}{3}\sum_{k_y=1}^{k_{\mathrm{out},y}} MIHG(\{k_y-1,L-k_{\mathrm{out},x}-k_y\};D,\{k_{\mathrm{in},y},k_{\mathrm{in},z}\},k_{\mathrm{out},x}+1) \\
&\quad + \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}} MIHG(\{k_y,L-k_{\mathrm{out},x}-k_y-1\};D,\{k_{\mathrm{in},y},k_{\mathrm{in},z}\},k_{\mathrm{out},x}+1) \\
&= \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}} \frac{\binom{k_{\mathrm{in},x}}{k_{\mathrm{out},x}-1}\binom{k_{\mathrm{in},y}}{k_y}\binom{k_{\mathrm{in},z}}{L-k_{\mathrm{out},x}-k_y}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},x}-k_{\mathrm{out},x}+1}{D-L+1} \\
&\quad + \frac{1}{3}\sum_{k_y=1}^{k_{\mathrm{out},y}} \frac{\binom{k_{\mathrm{in},x}}{k_{\mathrm{out},x}}\binom{k_{\mathrm{in},y}}{k_y-1}\binom{k_{\mathrm{in},z}}{L-k_{\mathrm{out},x}-k_y}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},x}-k_{\mathrm{out},x}}{D-L+1} \\
&\quad + \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}} \frac{\binom{k_{\mathrm{in},x}}{k_{\mathrm{out},x}}\binom{k_{\mathrm{in},y}}{k_y}\binom{k_{\mathrm{in},z}}{L-k_{\mathrm{out},x}-k_y-1}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},x}-k_{\mathrm{out},x}}{D-L+1}.
\end{aligned}
\tag{4.15}
$$

$C=2$ の場合と違って k_y についての和をとっているため，$\tilde{p}_x(L)$ を求めるには 1 重ループを処理する必要がある．$L=0$ となる確率は，$k_{\mathrm{out},x}=0$ であるときに，最初に文字 x が選択される確率であるから，

$$
\tilde{p}_x(L=0) = \frac{1}{3}\delta_{k_{\mathrm{out},x},0} \tag{4.16}
$$

と表される．

$\tilde{p}_y(L)$ も $\tilde{p}_x(L)$ と同様の手順により以下のように計算できる．

[*3] 初期条件として辞書からランダムに単語を選択する場合の $\tilde{p}_x(L)$, $\tilde{p}_y(L)$ は，(4.15), (4.17) 式の各項で Σ の前についている 1/3 がそれぞれ $k_{\mathrm{out},x}/D$, $k_{\mathrm{out},y}/D$, $k_{\mathrm{out},z}/D$ に変化する．

$L > 0$ のとき：

$$
\begin{aligned}
\tilde{p}_y(L) = &\frac{1}{3}\sum_{k_x=0}^{k_{\mathrm{out},x}}\sum_{k_z=0}^{k_{\mathrm{out},z}} MIHG(\{k_x,k_z\};D,\{k_{\mathrm{in},x},k_{\mathrm{in},z}\},k_{\mathrm{out},y})\delta_{k_{\mathrm{out},y}+k_x+k_z,L} \\
&+\frac{1}{3}\sum_{k_x=1}^{k_{\mathrm{out},x}}\sum_{k_z=0}^{k_{\mathrm{out},z}} MIHG(\{k_x-1,k_z\};D,\{k_{\mathrm{in},x},k_{\mathrm{in},z}\},k_{\mathrm{out},y}+1)\delta_{k_{\mathrm{out},y}+k_x+k_z,L} \\
&+\frac{1}{3}\sum_{k_x=0}^{k_{\mathrm{out},x}}\sum_{k_z=1}^{k_{\mathrm{out},z}} MIHG(\{k_x,k_z-1\};D,\{k_{\mathrm{in},x},k_{\mathrm{in},z}\},k_{\mathrm{out},y}+1)\delta_{k_{\mathrm{out},y}+k_x+k_z,L} \\
=&\frac{1}{3}\sum_{k_x=0}^{k_{\mathrm{out},x}}\frac{\binom{k_{\mathrm{in},y}}{k_{\mathrm{out},y}-1}\binom{k_{\mathrm{in},x}}{k_x}\binom{k_{\mathrm{in},z}}{L-k_x-k_{\mathrm{out},y}}}{\binom{D}{L-1}}\times\frac{k_{\mathrm{in},y}-k_{\mathrm{out},y}+1}{D-L+1} \\
&+\frac{1}{3}\sum_{k_x=1}^{k_{\mathrm{out},x}}\frac{\binom{k_{\mathrm{in},y}}{k_{\mathrm{out},y}}\binom{k_{\mathrm{in},x}}{k_x-1}\binom{k_{\mathrm{in},z}}{L-k_x-k_{\mathrm{out},y}}}{\binom{D}{L-1}}\times\frac{k_{\mathrm{in},y}-k_{\mathrm{out},y}}{D-L+1} \\
&+\frac{1}{3}\sum_{k_x=0}^{k_{\mathrm{out},x}}\frac{\binom{k_{\mathrm{in},y}}{k_{\mathrm{out},y}}\binom{k_{\mathrm{in},x}}{k_x}\binom{k_{\mathrm{in},z}}{L-k_x-k_{\mathrm{out},y}-1}}{\binom{D}{L-1}}\times\frac{k_{\mathrm{in},y}-k_{\mathrm{out},y}}{D-L+1}.
\end{aligned}
\tag{4.17}
$$

$L = 0$ のとき：

$$
\tilde{p}_y(L=0) = \frac{1}{3}\delta_{k_{\mathrm{out},y},0}. \tag{4.18}
$$

一般の C $(C \geq 3)$ における鎖長分布も多変量逆超幾何分布により記述することができる．文字の集合を $\Theta = \{\theta_1, \theta_2, \cdots, \theta_C\}$ として，$\theta_1, \theta_2, \cdots, \theta_{C-1}$ を終了可能文字としたときの鎖長分布 $\tilde{p}_{\theta_1}(L)$ は次のように表される．なお，他の終了可能文字の分布 $\tilde{p}_{\theta_i}(L)$ $(i = 2, 3, \cdots, C-1)$ は $\tilde{p}_{\theta_1}(L)$ に現れる θ_1 と θ_i を入れ替えることにより得られる．

$L > 0$ のとき：

$$
\begin{aligned}
&\tilde{p}_{\theta_1}(L) \\
&= \frac{1}{C} \sum_{k_{\theta_2}=0}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=0}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=0}^{k_{\mathrm{out},\theta_C}} MIHG(\{k_{\theta_2}, k_{\theta_3}, \cdots, k_{\theta_C}\}; D, \{k_{\mathrm{in},\theta_2}, k_{\mathrm{in},\theta_3}, \cdots, k_{\mathrm{in},\theta_C}\}, k_{\mathrm{out},\theta_1}) \\
&\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L} \\
&\quad + \frac{1}{C} \sum_{k_{\theta_2}=1}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=0}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=0}^{k_{\mathrm{out},\theta_C}} MIHG(\{k_{\theta_2}-1, k_{\theta_3}, \cdots, k_{\theta_C}\}; D, \{k_{\mathrm{in},\theta_2}, k_{\mathrm{in},\theta_3}, \cdots, k_{\mathrm{in},\theta_C}\}, k_{\mathrm{out},\theta_1}+1) \\
&\quad\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L} \\
&\quad + \frac{1}{C} \sum_{k_{\theta_2}=0}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=1}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=0}^{k_{\mathrm{out},\theta_C}} MIHG(\{k_{\theta_2}, k_{\theta_3}-1, \cdots, k_{\theta_C}\}; D, \{k_{\mathrm{in},\theta_2}, k_{\mathrm{in},\theta_3}, \cdots, k_{\mathrm{in},\theta_C}\}, k_{\mathrm{out},\theta_1}+1) \\
&\quad\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L} \\
&\quad + \cdots \\
&\quad + \frac{1}{C} \sum_{k_{\theta_2}=0}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=0}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=1}^{k_{\mathrm{out},\theta_C}} MIHG(\{k_{\theta_2}, k_{\theta_3}, \cdots, k_{\theta_C}-1\}; D, \{k_{\mathrm{in},\theta_2}, k_{\mathrm{in},\theta_3}, \cdots, k_{\mathrm{in},\theta_C}\}, k_{\mathrm{out},\theta_1}+1) \\
&\quad\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L} \\
&= \frac{1}{C} \sum_{k_{\theta_2}=0}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=0}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=0}^{k_{\mathrm{out},\theta_C}} \frac{\binom{k_{\mathrm{in},\theta_1}}{k_{\mathrm{out},\theta_1}-1}\binom{k_{\mathrm{in},\theta_2}}{k_{\theta_2}}\binom{k_{\mathrm{in},\theta_3}}{k_{\theta_3}}\cdots\binom{k_{\mathrm{in},\theta_C}}{k_{\theta_C}}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},\theta_1}-k_{\mathrm{out},\theta_1}+1}{D-L+1} \\
&\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L} \\
&\quad + \frac{1}{C} \sum_{k_{\theta_2}=1}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=0}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=0}^{k_{\mathrm{out},\theta_C}} \frac{\binom{k_{\mathrm{in},\theta_1}}{k_{\mathrm{out},\theta_1}}\binom{k_{\mathrm{in},\theta_2}}{k_{\theta_2}-1}\binom{k_{\mathrm{in},\theta_3}}{k_{\theta_3}}\cdots\binom{k_{\mathrm{in},\theta_C}}{k_{\theta_C}}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},\theta_1}-k_{\mathrm{out},\theta_1}}{D-L+1} \\
&\quad\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L} \\
&\quad + \frac{1}{C} \sum_{k_{\theta_2}=0}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=1}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=0}^{k_{\mathrm{out},\theta_C}} \frac{\binom{k_{\mathrm{in},\theta_1}}{k_{\mathrm{out},\theta_1}}\binom{k_{\mathrm{in},\theta_2}}{k_{\theta_2}}\binom{k_{\mathrm{in},\theta_3}}{k_{\theta_3}-1}\cdots\binom{k_{\mathrm{in},\theta_C}}{k_{\theta_C}}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},\theta_1}-k_{\mathrm{out},\theta_1}}{D-L+1} \\
&\quad\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L} \\
&\quad + \cdots \\
&\quad + \frac{1}{C} \sum_{k_{\theta_2}=0}^{k_{\mathrm{out},\theta_2}} \sum_{k_{\theta_3}=0}^{k_{\mathrm{out},\theta_3}} \cdots \sum_{k_{\theta_C}=1}^{k_{\mathrm{out},\theta_C}} \frac{\binom{k_{\mathrm{in},\theta_1}}{k_{\mathrm{out},\theta_1}}\binom{k_{\mathrm{in},\theta_2}}{k_{\theta_2}}\binom{k_{\mathrm{in},\theta_3}}{k_{\theta_3}}\cdots\binom{k_{\mathrm{in},\theta_C}}{k_{\theta_C}-1}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},\theta_1}-k_{\mathrm{out},\theta_1}}{D-L+1} \\
&\quad\quad \times \delta_{k_{\mathrm{out},\theta_1}+k_{\theta_2}+k_{\theta_3}+\cdots+k_{\theta_C},L}.
\end{aligned}
\tag{4.19}
$$

$L = 0$ のとき：

$$
\tilde{p}_{\theta_1}(L=0) = \frac{1}{C}\delta_{k_{\mathrm{out},\theta_1},0}. \tag{4.20}
$$

$\tilde{p}_{\theta_1}(L)$ を計算するには，文字の種類数 C に対して $C-2$ 重ループを計算する必要があることがわかる．

4.3.2 数値計算

4.3.1 節で得られた鎖長分布が正しいかどうかを，具体的な辞書「xyz シャッフル辞書」を用いた数値計算により確認する．xyz シャッフル辞書は $\Theta = \{x, y, z\}$，$D = 100$ のシャッフル辞書で，次数ベクトルは以下のように設定される．

$$\overrightarrow{k_{\mathrm{in}}} = (k_{\mathrm{in},x}, k_{\mathrm{in},y}, k_{\mathrm{in},z}) = (50, 30, 20),$$
$$\overrightarrow{k_{\mathrm{out}}} = (k_{\mathrm{out},x}, k_{\mathrm{out},y}, k_{\mathrm{out},z}) = (25, 10, 65).$$

この設定からしりとりは x か y で終了することがわかる．

xyz シャッフル辞書を 10000 冊作成し，各辞書 100 回ずつしりとりを実行した．このときの鎖長分布の実測値と，(4.15), (4.17) 式から計算される鎖長分布の理論値を比較した結果が図 4.6 である．実測値は理論値によく一致していることがわかる．

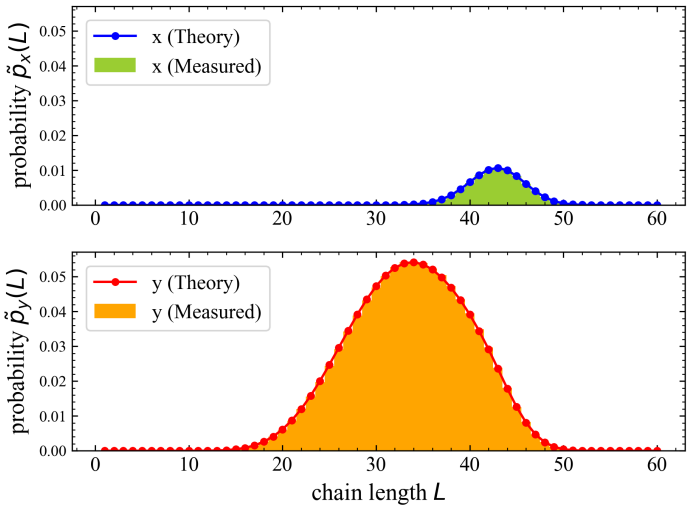


図 4.6: xyz シャッフル辞書の鎖長分布（ヒストグラム：実測値，実線：理論値）

xyz シャッフル辞書の冊数 N を変化させたときの，鎖長分布の実測値 $\tilde{p}_\theta^N(L)$ と理論値 $\tilde{p}_\theta(L)$ の L^2 ノルム $d_2(\tilde{p}_\theta^N, \tilde{p}_\theta)$，$\theta = x, y$ を測定した．結果は図 4.7 のようになった．冊数が大きくなると距離は減少し続けることがわかる．したがって，xyz シャッフル辞書の冊数が大きい極限では，実測値が理論値に漸近すると考えられる．

4.3.3 分布統計量

$C = 2$ と同様にカバーレートの統計量を求めたい．図 4.8 は D と $(k_{\mathrm{in},y}, k_{\mathrm{out},y})$ を固定したときの $(k_{\mathrm{in},x}/D, k_{\mathrm{out},x}/D)$ ごとの $L_{\max}/D$, $L_{\min}/D$, L_{mean}/D のヒートマップである．これらを解析的に求めることはできていないものの，$C = 2$ の場合と同様に，最小値は $k_{\mathrm{out},x}$ のみの関数であり，最大値と平均値がそれぞれ $k_{\mathrm{out},x}$, $k_{\mathrm{in},x}$ の差と比の関数であることが示

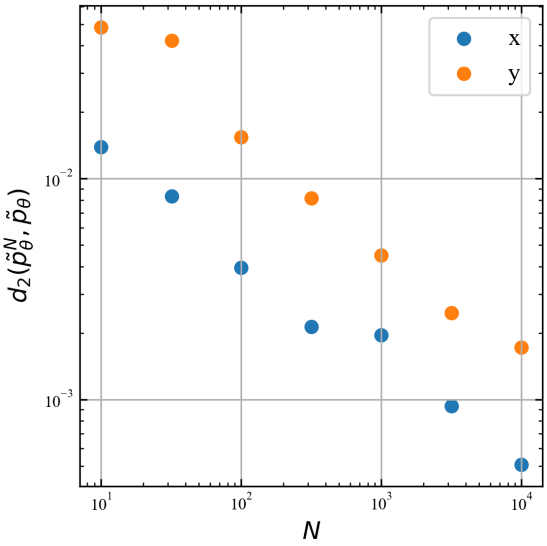
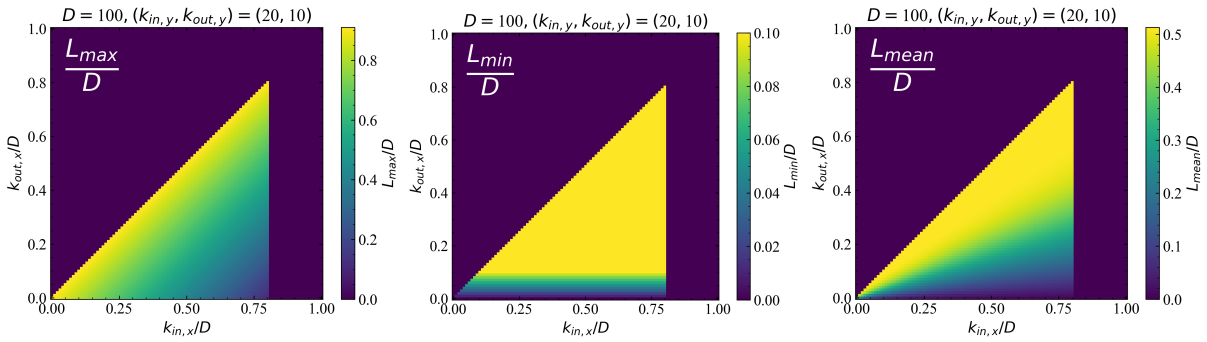


図 4.7: xyz シャッフル辞書の冊数 N と理論値との距離 $d_2(\tilde{p}_\theta^N, \tilde{p}_\theta)$ の関係

唆される．

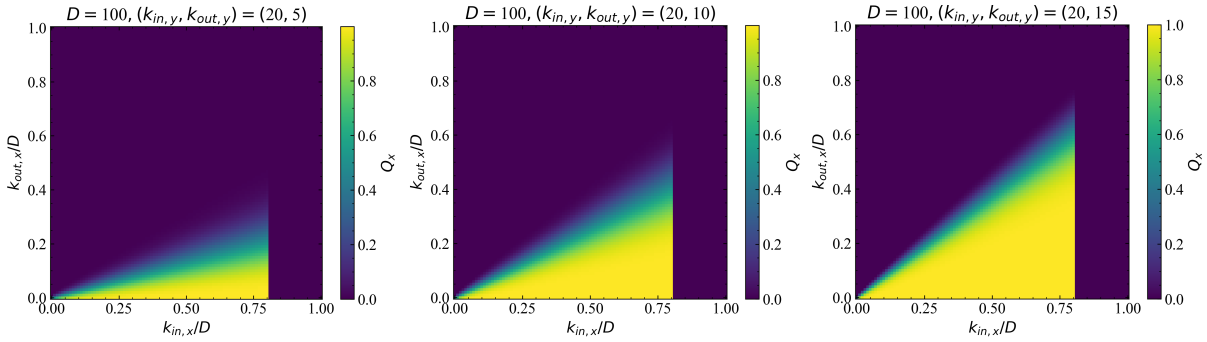


(a) $L_{\max}/D$　　(b) $L_{\min}/D$　　(c) L_{mean}/D

図 4.8: カバーレートの統計量のヒートマップ $(D = 100, (k_{\mathrm{in},y}, k_{\mathrm{out},y}) = (20, 10))$

$C \geq 3$ の場合，鎖長分布の統計量は $L_{\max}/D$, $L_{\min}/D$, L_{mean}/D に加えて，(3.11) で定義したしりとりが文字 $\theta \in \Theta$ で終了する確率 $\tilde{Q}_\theta$ を考えることができる．$C = 2$ のときは $\tilde{Q}_\theta = 1$ となり意味をもたなかったが，$C \geq 3$ では $\tilde{Q}_\theta$ が文字 θ での終了しやすさを意味するようになる．

図 4.9 は D を固定し $(k_{\mathrm{in},y}, k_{\mathrm{out},y})$ を変化させたときの $(k_{\mathrm{in},x}/D, k_{\mathrm{out},x}/D)$ ごとの $\tilde{Q}_x$ のヒートマップである．こちらも解析的な計算はできていないが，$\tilde{Q}_x$ は $k_{\mathrm{out},x}/k_{\mathrm{in},x}$ の関数であり，$k_{\mathrm{out},x}/k_{\mathrm{in},x} \approx k_{\mathrm{out},y}/k_{\mathrm{in},y}$ で急激に変化することが示唆される．



(a) $(k_{\mathrm{in},y}, k_{\mathrm{out},y}) = (20, 5)$　(b) $(k_{\mathrm{in},y}, k_{\mathrm{out},y}) = (20, 10)$　(c) $(k_{\mathrm{in},y}, k_{\mathrm{out},y}) = (20, 15)$

図 4.9: $\tilde{Q}_x$ のヒートマップ $(D = 100)$

4.4　実際の辞書への適用

本節では国名辞書のシャッフル辞書群における鎖長分布を解析的に計算する．計算にあたっては文字数の圧縮を行う．文字数の圧縮とは，辞書ネットワークの複数の頂点を同一視することである．シャッフル辞書群では，複数の文字の入次数，出次数を足し合わせて 1 つの文字とみなす操作にあたる．圧縮を行う理由は，文字数が増えるほど計算量が増加するからである．$C = 2$ の場合，$\tilde{p}_x(L)$ は (4.5) 式で表され，1 回の処理で計算することができる．一方で $C = 3$ の場合，$\tilde{p}_x(L)$ は (4.15) 式で表され，k_y についての和，すなわち 1 重ループを処理する必要がある．$C = 4, 5, \cdots$ と増加すると，(4.19) 式のように鎖長分布に総和記号が 1 つずつ追加され，それに伴って $C - 2$ 重ループの処理が発生する．ゆえに，文字数の大きな辞書においては圧縮が不可欠となる．

圧縮は鎖長分布にどう影響するのだろうか．まず，すべての終了不能文字を 1 文字に圧縮しても鎖長分布は変化しない．これは壺モデルで終了条件のない球を同一視しても問題がないことから明らかである．終了可能文字を含めて圧縮した場合は鎖長分布が変化するが，その影響の大きさはよく分かっていない．

今回は実測で終了文字となる回数の多かった 4 文字 a, y, o, q のみを残し，5 文字の終了可能文字を含む他の 20 文字を θ_{others} の 1 文字に圧縮した．すなわち，文字の集合は $\Theta = \{\mathrm{a}, \mathrm{b}, \cdots, \mathrm{u}, \mathrm{v}, \mathrm{y}, \mathrm{z}\}$ から $\Theta' = \{\mathrm{a}, \mathrm{y}, \mathrm{o}, \mathrm{q}, \theta_{\mathrm{others}}\}$ に変化し，圧縮の定義から θ_{others} の入次数，出次数は次の式を満たす．

$$k_{\mathrm{in},\theta_{\mathrm{others}}} = \sum_{\theta \in \Theta \setminus \{\mathrm{a,y,o,q}\}} k_{\mathrm{in},\theta},$$

$$k_{\mathrm{out},\theta_{\mathrm{others}}} = \sum_{\theta \in \Theta \setminus \{\mathrm{a,y,o,q}\}} k_{\mathrm{out},\theta}.$$

よって，圧縮した国名辞書は以下の次数ベクトル $\overrightarrow{k'_{\mathrm{in}}}$, $\overrightarrow{k'_{\mathrm{out}}}$ を持つ．

$$\overrightarrow{k'_{\mathrm{in}}} = (k_{\mathrm{in,a}}, k_{\mathrm{in,y}}, k_{\mathrm{in,o}}, k_{\mathrm{in,q}}, k_{\mathrm{in},\theta_{\mathrm{others}}}) = (72, 6, 11, 1, 103),$$
$$\overrightarrow{k'_{\mathrm{out}}} = (k_{\mathrm{out,a}}, k_{\mathrm{out,y}}, k_{\mathrm{out,o}}, k_{\mathrm{out,q}}, k_{\mathrm{out},\theta_{\mathrm{others}}}) = (11, 1, 1, 1, 179).$$

圧縮していない国名辞書における鎖長分布の実測値（10000 冊 × 各 100 回）と圧縮した国名辞書における鎖長分布の理論値を比較した結果が図 4.10 である．$C = 24$ から $C = 5$ に圧縮して計算したにも関わらず，実測値は理論値によく一致している．ただし，理論値は他の終了可能文字の分布をもたないため，どれだけ実測値の精度を上げても完全に一致することはない点に注意が必要である．

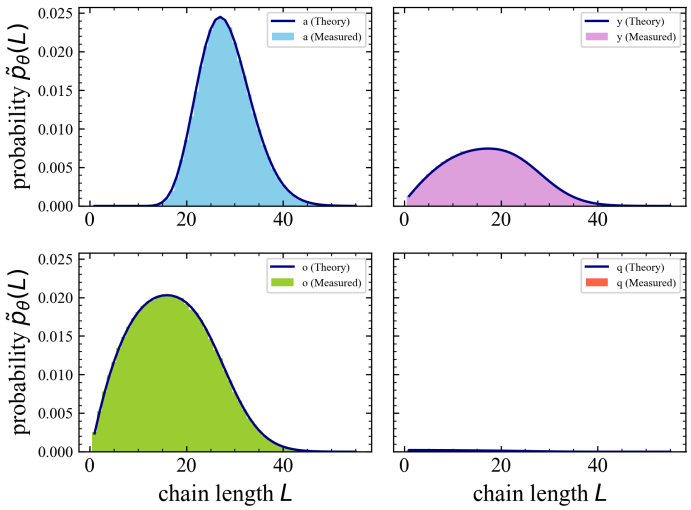


図 4.10: 国名辞書のシャッフル辞書群の鎖長分布（ヒストグラム：実測値，実線：理論値，a, y, o, q のみを表示.）

第 5 章

単一辞書

本研究の第一の目的は 1 つの辞書の鎖長分布を求めることである．単一辞書はシャッフル辞書群のように鎖長分布を数式で表すことは難しいが，アルゴリズムを構築して厳密に分布を導出することが可能である．本章では単一辞書の鎖長分布を求める方法について説明し，単一辞書で分布を求めるのが難しい理由について考察する．

5.1 経路行列

多重有向グラフの経路に対して経路行列と鎖経路行列を定義する．

定義 5.1.1. N 頂点の多重有向グラフ上の長さ l の経路に対して，$\theta\phi$ 成分が経路上の有向辺 $\theta\phi$ の数であるような N 次正方行列 $J^{(l)}$ を経路行列と呼ぶ．また，経路がそれ以上伸びないとき，経路行列を J で表し，鎖経路行列と呼ぶ．

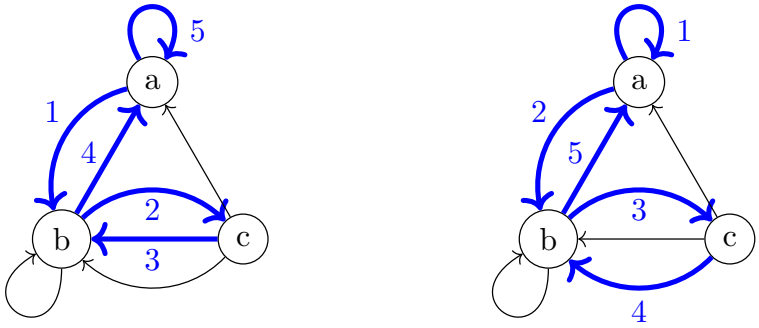
しりとりは単語辺型ネットワーク上の自己回避トレイルである．したがって，しりとりの l ステップ目の単語列は経路行列 $J^{(l)}$ で表され，鎖長 L で終了したときの単語列は鎖経路行列 $J = J^{(L)}$ で表される．例えば，図 5.1 は 2.1 節で登場した辞書ネットワーク上の異なる 2 つの経路を表しており，いずれの経路も鎖経路行列 J は以下のように求められる．

$$J = J^{(5)} = \begin{pmatrix} J^{(5)}_{\mathrm{aa}} & J^{(5)}_{\mathrm{ab}} & J^{(5)}_{\mathrm{ac}} \\ J^{(5)}_{\mathrm{ba}} & J^{(5)}_{\mathrm{bb}} & J^{(5)}_{\mathrm{bc}} \\ J^{(5)}_{\mathrm{ca}} & J^{(5)}_{\mathrm{cb}} & J^{(5)}_{\mathrm{cc}} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

しりとりで単語を 1 つ選択すると，辞書ネットワークは辺を 1 本失い，単語列に対応する経路がその辺を獲得する．よって，$A^{(l)}_{\theta\phi}$ が 1 減少すると，$J^{(l)}_{\theta\phi}$ は 1 増加する．この対応関係から次の式が成り立つ．

$$\text{任意の } \theta, \phi \in \Theta \text{ について}, 0 \leq J_{\theta\phi} \leq A_{\theta\phi}. \tag{5.1}$$

また，$J^{(l)}$ のすべての成分の和はステップごとに 1 ずつ増加するので，鎖長 L と鎖経路行列



(a) 経路 1　　(b) 経路 2

図 5.1: 辞書ネットワーク上の 2 つの経路

J の間に次の式が成り立つ.

$$L = \sum_{\theta \in \Theta} \sum_{\phi \in \Theta} J_{\theta\phi}. \tag{5.2}$$

さらに，経路行列が鎖経路行列となる条件（鎖形成条件）は，経路行列の θ 行が初期隣接行列の θ 行に一致した後，文字 θ で終わる単語が選択されたときと考えられる．以上より，しりとりにおける $J^{(l)}$ の成長の仕方は以下のようにまとめられる．

1. $J^{(0)} = O$（零行列）から始める．
2. 文字 θ で始まり文字 ϕ で終わる単語が選択されたら，$\theta\phi$ 成分を 1 増加させる．
3. いずれかの文字 θ で，$J^{(l)}$ の θ 行が A の θ 行に一致した後，末尾文字に θ を選択したらしりとりは終了し，そのときの経路行列 $J^{(L)}$ を鎖経路行列とする．

鎖長分布 $p(L)$ を求める単純な方法は単語列を列挙することである．つまり，辞書ネットワーク上の経路をすべて列挙して，各経路における鎖長と実現確率を求めればよい．しかし，この方法は効率的とは言えない．例えば，図 5.1 の経路 1 の実現確率は (3.9) 式を用いて，

$$\frac{1}{C} \times q_{\mathrm{ab}}^{(0)} \times q_{\mathrm{bc}}^{(1)} \times q_{\mathrm{cb}}^{(2)} \times q_{\mathrm{ba}}^{(3)} \times q_{\mathrm{aa}}^{(4)} = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{54}$$

と計算される．同様に経路 2 の実現確率は,

$$\frac{1}{C} \times q_{\mathrm{aa}}^{(0)} \times q_{\mathrm{ab}}^{(1)} \times q_{\mathrm{bc}}^{(2)} \times q_{\mathrm{cb}}^{(3)} \times q_{\mathrm{ba}}^{(4)} = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{1} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{54}$$

と計算される．両者の実現確率が一致していることがわかる．このように，経路を構成する有向辺が同じであれば，その有向辺を通過する順番が異なっていても実現確率は一致する．一般に，鎖経路行列が J であるような 1 本の経路の実現確率 p_{path} は以下のように表される．

$$p_{\mathrm{path}} = \frac{1}{C} \frac{\prod_{\theta \in \Theta} \prod_{\phi \in \Theta} A_{\theta\phi}{}^{[J_{\theta\phi}]}}{\prod_{\theta \in \Theta} k_{\mathrm{out},\theta}{}^{[\sum_{\phi \in \Theta} J_{\theta\phi}]}}. \tag{5.3}$$

ここで，$a^{[n]}$ は Pochhammer 記号 [38] であり，整数 a, n $(n \geq 0)$ について，$a^{[n]} = a(a-1)\cdots(a-n+1)$, $a^{[0]} = 1$ と定義される．つまり，ある経路の実現確率を計算するためには，鎖経路行列が分かれば十分であり，具体的な経路を知る必要はない．本章では，経路行列を用いて効率的に鎖長分布を計算する方法について説明する．

5.2 $C = 2$

$C = 2$ の単一辞書における鎖長分布の求め方を説明する．使用する文字は $\Theta = \{x, z\}$ であり，以下の隣接行列が与えられるとする．

$$A = \begin{pmatrix} A_{xx} & A_{xz} \\ A_{zx} & A_{zz} \end{pmatrix}. \tag{5.4}$$

また $k_{\mathrm{in},x} > k_{\mathrm{out},x}$ であり，しりとりは必ず x で終了する．

5.2.1 ブロック法

ブロック法は，与えられた隣接行列 A から実現しうる鎖経路行列 J を求め，各 J の鎖長と実現確率を計算する方法である．経路や経路行列の探索処理を必要としないため，非常に効率的に計算することができ，$C = 2$ の場合では鎖長分布を数式で表せるようになる．

初めに $\theta, \phi \in \Theta$ について，$J_{\theta\phi}$ のとりうる範囲を決定する．J_{xx}, J_{xz} については，5.1 節で述べた鎖形成条件から，

$$J_{xx} = A_{xx}, \quad J_{xz} = A_{xz} \tag{5.5}$$

と表せる．J_{zx} については，これから説明するブロック列を用いて考えることができる．

しりとりの単語列から各単語の末尾文字だけを抜き出した文字列を末尾文字列と呼ぶ．ただし，末尾文字列の先頭に $\theta^{(0)}$ すなわち 1 単語目の最初の文字を追加する．さらに，末尾文字列で同じ文字が 1 文字以上連続している部分を，1 つのブロックに置き換えた列をブロック列と呼ぶ．例えば，しりとりの単語列が “$x\cdots z$”（x で始まり z で終わる単語）→ “$z\cdots z$” → “$z\cdots x$” → “$x\cdots z$” → “$z\cdots x$” → “$x\cdots x$” → “$x\cdots x$” の場合，末尾文字列は $xzzxzxxx$ と表され，ブロック列は $\boxed{x}\boxed{z}\boxed{x}\boxed{z}\boxed{x}$ と表される．ブロック列の定義から隣接するブロックは必ず異なることがわかる．

しりとりは x で終了するので，ブロック列も $\boxed{x}$ で終了する．よって，ブロック列は以下の 2 パターンが考えられる．

$$\theta^{(0)} = x \text{ のとき}, \quad \boxed{x}\boxed{z}\cdots\boxed{z}\boxed{x},$$
$$\theta^{(0)} = z \text{ のとき}, \quad \boxed{z}\boxed{x}\cdots\boxed{z}\boxed{x}.$$

前者は辺 zx が辺 xz と同数使われ，後者は辺 zx が辺 xz より 1 つ多く使われる．したがっ

て，次の保存則が成立する．

$$J_{zx} - J_{xz} = \begin{cases} 0 & (\theta^{(0)} = x) \\ 1 & (\theta^{(0)} = z) \end{cases} \tag{5.6}$$

ゆえに，(5.5), (5.6) 式より，

$$J_{zx} = \begin{cases} A_{xz} & (\theta^{(0)} = x) \\ A_{xz} + 1 & (\theta^{(0)} = z) \end{cases} \tag{5.7}$$

が得られる．J_{zz} に関しては (5.1) 式以上のことはわからない．よって，実現しうる J は，

$$J = \begin{pmatrix} A_{xx} & A_{xz} \\ A_{xz} \text{ or } A_{xz}+1 & J_{zz} \end{pmatrix}, \quad J_{zz} = 0, 1, \cdots, A_{zz} \tag{5.8}$$

の $2(A_{zz}+1)$ 通りである．

鎖経路行列を求めることができたので，各々の鎖長と実現確率を計算する．鎖長 L は (5.2) 式により簡単に求めることができ，

$$L = \begin{cases} A_{xx} + 2A_{xz} + J_{zz} & (\theta^{(0)} = x) \\ A_{xx} + 2A_{xz} + J_{zz} + 1 & (\theta^{(0)} = z) \end{cases} \tag{5.9}$$

と計算される．

続いて，各経路行列 J が実現する確率を求める．J が与えられたとき，それを満たすブロック列は一意に定まる．$\theta^{(0)} = x$ の場合，ブロック列は

$$\boxed{x}\boxed{z}\boxed{x}\boxed{z}\cdots\boxed{z}\boxed{x}$$

であり，$\boxed{x}$ が $A_{xz}+1$ 個，$\boxed{z}$ が A_{xz} 個含まれている．また，末尾文字列で x が連続する部分は A_{xx} 個，z が連続する部分は J_{zz} 個存在する．したがって，A_{xx} 個の x の連続を $A_{xz}+1$ 個の $\boxed{x}$ に割り当て，J_{zz} 個の z の連続を A_{xz} 個の $\boxed{z}$ に割り当てることで末尾文字列が完成する．その場合の数は，

$$\binom{A_{xx}+A_{xz}}{A_{xx}}\binom{J_{zz}+A_{xz}-1}{J_{zz}}$$

と求められる．どの末尾文字列も同じ確率で実現し，その確率は (5.3) 式を用いて，

$$\begin{aligned} &\frac{1}{2}\frac{{A_{xx}}^{[J_{xx}]}{A_{xz}}^{[J_{xz}]}}{(A_{xx}+A_{xz})^{[J_{xx}+J_{xz}]}} \cdot \frac{{A_{zx}}^{[J_{zx}]}{A_{zz}}^{[J_{zz}]}}{(A_{zx}+A_{zz})^{[J_{zx}+J_{zz}]}} \\ =&\frac{1}{2}\frac{{A_{xx}}^{[A_{xx}]}{A_{xz}}^{[A_{xz}]}}{(A_{xx}+A_{xz})^{[A_{xx}+A_{xz}]}} \cdot \frac{{A_{zx}}^{[A_{xz}]}{A_{zz}}^{[J_{zz}]}}{(A_{zx}+A_{zz})^{[A_{xz}+J_{zz}]}} \\ =&\frac{1}{2\binom{A_{xx}+A_{xz}}{A_{xx}}} \cdot \frac{{A_{zx}}^{[A_{xz}]}{A_{zz}}^{[J_{zz}]}}{(A_{zx}+A_{zz})^{[A_{xz}+J_{zz}]}} \end{aligned}$$

と計算される．したがって，$\theta^{(0)} = x$ のとき J_{zz} が実現する確率は次のように求められる．

$$
\begin{aligned}
&\binom{A_{xx}+A_{xz}}{A_{xx}}\binom{J_{zz}+A_{xz}-1}{J_{zz}} \times \frac{1}{2\binom{A_{xx}+A_{xz}}{A_{xx}}} \cdot \frac{{A_{zx}}^{[A_{xz}]}{A_{zz}}^{[J_{zz}]}}{(A_{zx}+A_{zz})^{[A_{xz}+J_{zz}]}} \\
&= \frac{1}{2}\binom{A_{xz}+J_{zz}-1}{J_{zz}} \cdot \frac{A_{zx}!}{(A_{zx}-A_{xz})!} \cdot \frac{A_{zz}!}{(A_{zz}-J_{zz})!} \cdot \frac{(A_{zx}+A_{zz}-A_{xz}-J_{zz})!}{(A_{zx}+A_{zz})!} \\
&= \frac{1}{2}\binom{A_{xz}+J_{zz}-1}{J_{zz}} \cdot \frac{(A_{zx}+A_{zz}-A_{xz}-J_{zz})!}{(A_{zx}-A_{xz})!(A_{zz}-J_{zz})!} \cdot \frac{A_{zx}!A_{zz}!}{(A_{zx}+A_{zz})!} \\
&= \frac{1}{2}\frac{\binom{A_{xz}+J_{zz}-1}{J_{zz}}\binom{A_{zx}+A_{zz}-A_{xz}-J_{zz}}{A_{zz}-J_{zz}}}{\binom{A_{zx}+A_{zz}}{A_{zz}}}. \qquad (5.10)
\end{aligned}
$$

同様に，$\theta^{(0)} = z$ の場合に J_{zz} が実現する確率は次のように求められる．

$$
\begin{aligned}
&\binom{A_{xx}+A_{xz}}{A_{xx}}\binom{J_{zz}+A_{xz}}{J_{zz}} \times \frac{1}{2\binom{A_{xx}+A_{xz}}{A_{xx}}} \cdot \frac{{A_{zx}}^{[A_{xz}+1]}{A_{zz}}^{[J_{zz}]}}{(A_{zx}+A_{zz})^{[A_{xz}+J_{zz}+1]}} \\
&= \frac{1}{2}\frac{\binom{A_{xz}+J_{zz}}{J_{zz}}\binom{A_{zx}+A_{zz}-A_{xz}-J_{zz}-1}{A_{zz}-J_{zz}}}{\binom{A_{zx}+A_{zz}}{A_{zz}}}. \qquad (5.11)
\end{aligned}
$$

(5.10), (5.11) 式の J_{zz} を L に変換し，それらを足し合わせることで，$p_x(L)$ は以下のような負の超幾何分布の形で表すことができる．

$$
\begin{aligned}
p_x(L) &= \frac{1}{2}\frac{\binom{L-k_{\mathrm{out},x}-1}{L-k_{\mathrm{out},x}-A_{xz}}\binom{D-L}{k_{\mathrm{out},x}+k_{\mathrm{in},z}-L}}{\binom{k_{\mathrm{out},z}}{A_{zz}}} + \frac{1}{2}\frac{\binom{L-k_{\mathrm{out},x}-1}{L-k_{\mathrm{out},x}-A_{xz}-1}\binom{D-L}{k_{\mathrm{out},x}+k_{\mathrm{in},z}-L+1}}{\binom{k_{\mathrm{out},z}}{A_{zz}}}. \\
&= \frac{1}{2}NHG(L-k_{\mathrm{out},x}-A_{xz}; k_{\mathrm{out},z}, A_{zz}, A_{xz}) \\
&\quad + \frac{1}{2}NHG(L-k_{\mathrm{out},x}-A_{xz}-1; k_{\mathrm{out},z}, A_{zz}, A_{xz}+1). \qquad (5.12)
\end{aligned}
$$

ブロック法による求め方を以下にまとめる．

ブロック法

1. 実現しうる鎖経路行列 J を求める．
2. 各 J の鎖長と実現確率を計算する．
3. 鎖長が L となる J を集めて，それらの実現確率の和を $p_\theta(L)$ とする．

鎖長 L の統計量について述べておく．最大値 $L_{\max}$ と最小値 $L_{\min}$ は (5.9) 式から直ちに求めることができる．

$$
L_{\max} = A_{xx} + 2A_{xz} + A_{zz} + 1, \qquad (5.13)
$$

$$
L_{\min} = A_{xx} + 2A_{xz}. \qquad (5.14)
$$

平均値 L_{mean} は (5.12) 式から 4.2.3 節と同様の手順で計算することができ，

$$
L_{\mathrm{mean}} = \frac{(2A_{xz}+1)A_{zz}}{2(A_{zx}+1)} + A_{xx} + 2A_{xz} + \frac{1}{2} \qquad (5.15)
$$

と求められる．

5.2.2　数値計算

5.2.1 節のブロック法が正しいかどうかを，具体的な辞書「xz 単一辞書」を用いた数値計算により確認する．xz 単一辞書は $\Theta = \{x, z\}$，$D = 100$ の単一辞書で，隣接行列 A は以下のように設定される．

$$A = \begin{pmatrix} A_{xx} & A_{xz} \\ A_{zx} & A_{zz} \end{pmatrix} = \begin{pmatrix} 10 & 10 \\ 60 & 20 \end{pmatrix}.$$

この設定からしりとりは x で終了することがわかる．

xz 単一辞書でしりとりを 10 万回実行した．このときの鎖長分布の実測値と，(5.12) 式から計算される鎖長分布の理論値を比較した結果が図 5.2 である．実測値は理論値によく一致していることがわかる．

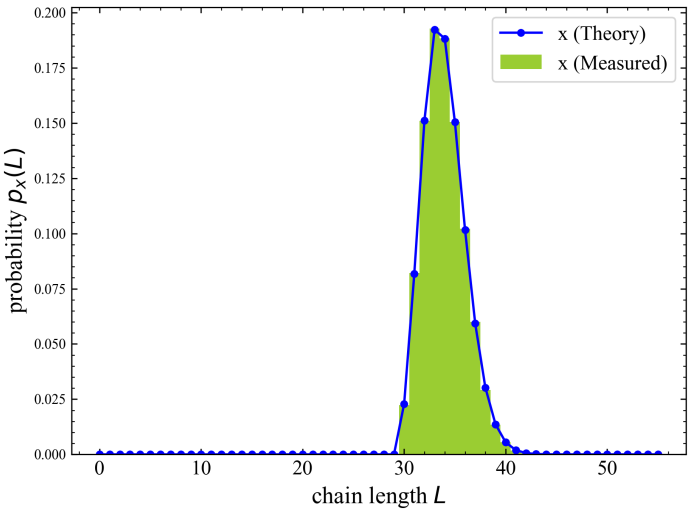


図 5.2: xz 単一辞書の鎖長分布（ヒストグラム：実測値，実線：理論値）

5.3　$C \geq 3$

続いて，$C \geq 3$ の単一辞書における鎖長分布の求め方を説明する．ここでは特に $C = 3$ に着目するが，$C > 3$ の場合にも同様の手法を適用することができる．使用する文字は $\Theta = \{x, y, z\}$ であり，以下の隣接行列が与えられるとする．

$$A = \begin{pmatrix} A_{xx} & A_{xy} & A_{xz} \\ A_{yx} & A_{yy} & A_{yz} \\ A_{zx} & A_{zy} & A_{zz} \end{pmatrix}. \tag{5.16}$$

また $k_{\mathrm{in},x} > k_{\mathrm{out},x}$, $k_{\mathrm{in},y} > k_{\mathrm{out},y}$ であり，しりとりは必ず x か y で終了する．

5.3.1 ブロック法の限界

$C = 3$ の場合はブロック法をそのまま適用するのが難しくなる．なぜなら，経路行列からブロック列が一意に定まらなくなるからである．このことについて 5.2.1 節と同じ手順を辿りながら説明する．なお，しりとりは x で終了する場合を考える．

初めに実現しうる鎖経路行列 J を求める．鎖形成条件から以下が成り立つ．

$$J_{xx} = A_{xx}, \quad J_{xy} = A_{xy}, \quad J_{xz} = A_{xz}. \tag{5.17}$$

また，次の 2 つの保存則が成立する．

$$J_{yx} + J_{zx} - J_{xy} - J_{xz} = \begin{cases} 0 & (\theta^{(0)} = x) \\ 1 & (\theta^{(0)} = y, z) \end{cases} \tag{5.18}$$

$$J_{yx} + J_{yz} - J_{xy} - J_{zy} = \begin{cases} 1 & (\theta^{(0)} = y) \\ 0 & (\theta^{(0)} = x, z) \end{cases} \tag{5.19}$$

ここで $\theta^{(0)} = x$ とすると，$J_{yy}, J_{zz}, J_{yx}, J_{yz}$ の 4 つを固定することで以下の J を得ることができる．

$$J = \begin{pmatrix} A_{xx} & A_{xy} & A_{xz} \\ J_{yx} & J_{yy} & J_{yz} \\ A_{xy} + A_{xz} - J_{yx} & J_{yx} + J_{yz} - A_{xy} & J_{zz} \end{pmatrix}. \tag{5.20}$$

$C = 2$ の場合は J を満たすブロック列がちょうど 1 つだけ存在していたが，$C = 3$ の場合はどうだろうか．(5.20) 式を満たすブロック列は，$\boxed{x}$ を $A_{xy} + A_{xz} + 1$ 個，$\boxed{y}$ を $J_{yx} + J_{yz}$ 個，$\boxed{z}$ を $A_{xz} + J_{yz}$ 個含んでいる必要がある．これらを最初と最後が $\boxed{x}$ で，同じブロックが連続しないように並べると，

$$\boxed{x}\,\boxed{y}\,\boxed{x}\cdots\boxed{x},$$
$$\boxed{x}\,\boxed{y}\,\boxed{z}\cdots\boxed{x},$$
$$\boxed{x}\,\boxed{z}\,\boxed{x}\cdots\boxed{x}$$

などのさまざまなブロック列を考えることができる．どのブロック列でも，そこからできる末尾文字列の数と実現確率は同じである．したがって，J から構成されるブロック列の数を求めることができれば，鎖長分布を計算することができると考えられる．

鎖経路行列を満たすブロック列の個数を求める方法は今のところ判明していない．この課題と類似の問題に，与えられた Euler グラフ上の異なる Euler 回路を数え上げる問題が存在する．その解がグラフ理論の BEST 定理 [39][40]（定理の内容は付録 E を参照）であることから，ブロック法に BEST 定理を適用できる可能性があり，現在調査を進めている．

5.3.2 ノーマル法とグッド法

経路行列に対応する末尾文字列の個数を縮退度と呼ぶ．ブロック法は効率的なアルゴリズムである一方で，縮退度がわからない場合，鎖長分布を計算することができなくなる．そこで効率性は一旦問題とせずに，経路行列 $J^{(l)}$ を総当たりで探索して $p_\theta(L)$ を求めることを考える．本節ではノーマル法のアルゴリズムについて説明した後，その改良版としてグッド法を提案する．

しりとりが x で終了する場合を考える．このとき，末尾文字列は必ず x で終了する．そこで，末尾文字列を x から始めて，末尾文字列の先頭に文字を追加していく操作を行う．つまり末尾文字列は，$l=1$ のとき xx, yx, zx の 3 通り，$l=2$ のとき xxx, xyx, xzx, yxx, yyx, yzx, zxx, zyx, zzx の 9 通り，$\cdots$ と l ステップ目に 3^l 通りの末尾文字列ができる．この過程をバックトレース (BT) 過程と呼ぶ．そして経路行列 $J^{(l)}$ は零行列から始まり，追加した単語に対応する成分が 1 ずつ増加して，図 5.3 のように遷移していく．

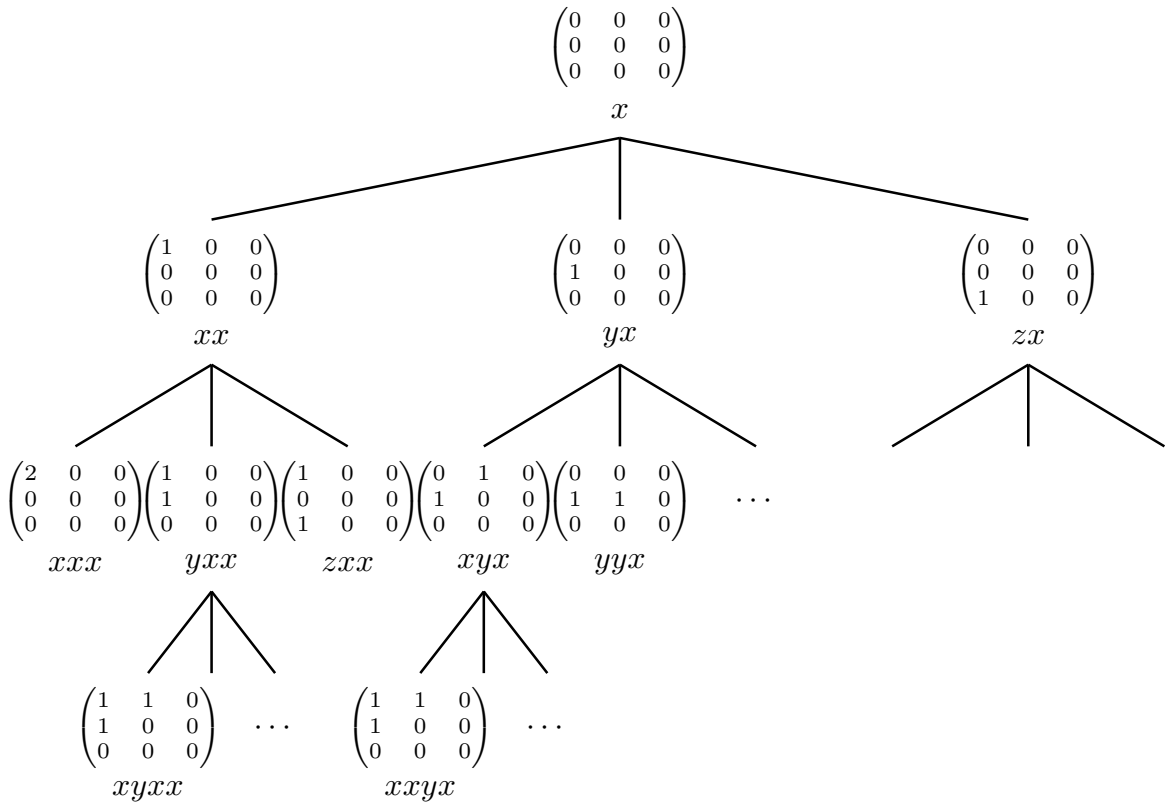


図 5.3: 経路行列の遷移図

(5.1) 式より，経路行列の各成分が隣接行列の各成分より大きくなることはない．よって，$J^{(l)}_{\theta\phi} > A_{\theta\phi}$ を満たす $\theta\phi$ 成分を持つような経路行列が発生したら，その行列の BT 過程を打ち切る．そのため，BT 過程は必ず D ステップ以内に終了する．こうして生き残った $J^{(l)}$ のうち，(5.17) 式の鎖形成条件を満たしたものが鎖経路行列 J となる．以降はブロック法と同

様であり，J ごとに鎖長と実現確率を計算して，鎖長分布を求めることができる．

ノーマル法

1. l ステップ目の経路行列 $J^{(l)}$ を列挙する．
2. 鎖形成条件を満たす $J^{(l)}$ を鎖経路行列 J とする．
3. J ごとに実現確率を計算して，その和を $p_\theta(l)$ とする．

ノーマル法は各ステップで最大 C の分岐を持つため，効率的であるとは言えない．そこで縮退度に着目する．図 5.3 で $xyxx$ と $xxyx$ は同じ経路行列を持っている．言い換えると，その経路行列の縮退度は 2 である．当然 $xyxx$ から派生する経路行列と，$xxyx$ から派生する経路行列は一致するため，$xyxx$ が縮退度 2 という情報を持ちながら探索を続け，$xxyx$ は打ち切るという手法を取れる．

末尾文字列が 3^l で指数的に増加する一方で，経路行列は 3^2 個の要素しか持たないため，l が大きくなるにつれてどんどんと重複するようになる．では，経路行列の種類数はどのように増加するのだろうか．表 5.1 は l ステップ目における末尾文字列と経路行列 $J^{(l)}$ の種類数を示しており，図 5.4 は l ごとの $J^{(l)}$ の種類数を両対数でプロットした結果である．なお，この図表は隣接行列を考えない場合であり，途中で打ち切ることなく BT 過程を繰り返したときの結果である．l が大きいところでは，$J^{(l)}$ の種類数はおよそ $l^{4.6}$ に比例しており，末尾文字列に比べて緩やかに増加することが確認された．したがって，重複している経路行列を探索しないだけで，計算量を大幅に減らすことができると期待される．

表 5.1: l ステップ目における末尾文字列，経路行列の種類数

l	末尾文字列の種類数	経路行列の種類数
1	3	3
2	9	9
3	27	25
4	81	62
5	243	138
6	729	280
7	2187	526
8	6561	928
9	19683	1554
10	59049	2492
11	177147	3852
12	531441	5771
13	1594323	8415

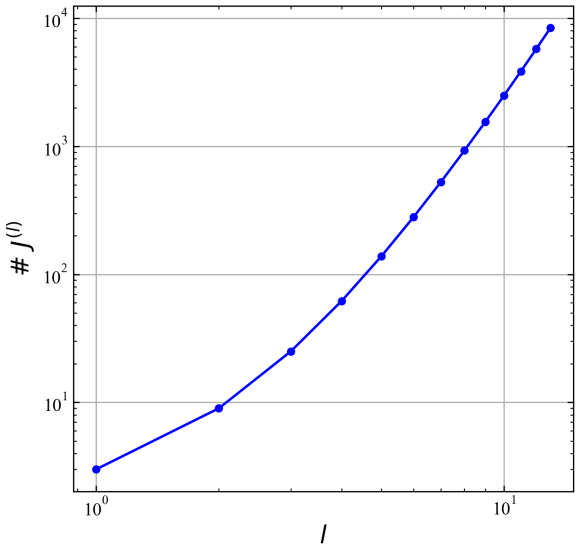


図 5.4: l ステップ目における経路行列 $J^{(l)}$ の種類数（両対数）

グッド法のアルゴリズムは以下のとおりである．BT 過程において，$J^{(l)}$ に加えてその縮退度 $\deg(J^{(l)})$ を以下のように計算して記録する．まず $J^{(0)}$ に対する縮退度を 1 とする．新たな経路行列 $J_1^{(l)}$ が現れたら，その縮退度 $\deg(J_1^{(l)})$ は親の経路行列の縮退度を継承する．そして，$J_1^{(l)}$ と同じ経路行列 $J_2^{(l)}$ が現れたら，$J_2^{(l)}$ の探索は打ち切りとし，$\deg(J_1^{(l)})$ に $\deg(J_2^{(l)})$ を加算する．これを繰り返して鎖形成条件を満たしたすべての経路行列 $J^{(l)}$ について，その実現確率と縮退度の積を計算し，足し合わせた値が $p_\theta(l)$ となる．このようにしてノーマル法よりも効率的に鎖長分布を計算することが可能となる．

グッド法

1. l ステップ目の経路行列 $J^{(l)}$ と縮退度 $\deg(J^{(l)})$ を重複なく列挙する．
2. 鎖形成条件を満たす $J^{(l)}$ を鎖経路行列 J とする．
3. J ごとに実現確率と縮退度の積を求めて，その和を $p_\theta(l)$ とする．

5.3.3 数値計算

5.3.2 節のグッド法が正しいかどうかを，具体的な辞書「xyz 単一辞書」を用いた数値計算により確認する．xyz 単一辞書は $\Theta = \{x, y, z\}$，$D = 13$ の単一辞書で，隣接行列 A は以下のように設定される．

$$A = \begin{pmatrix} A_{xx} & A_{xy} & A_{xz} \\ A_{yx} & A_{yy} & A_{yz} \\ A_{zx} & A_{zy} & A_{zz} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}.$$

この設定からしりとりは x か y で終了することがわかる．

xyz 単一辞書でしりとりを 10 万回実行した．このときの鎖長分布の実測値と，5.3.2 節の

グッド法で求めた鎖長分布の理論値を比較した結果が図 5.5 である．実測値は理論値によく一致していることがわかる．

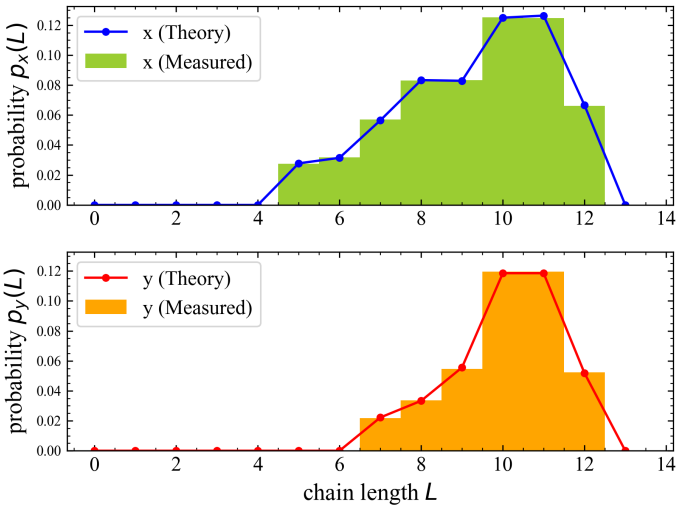


図 5.5: xyz 単一辞書の鎖長分布（ヒストグラム：実測値，実線：理論値）

5.4 実際の辞書への適用

グッド法はノーマル法に比べて計算時間が短縮されるが，文字数や単語数の大きな辞書では依然として膨大な計算時間を必要とする．そのため，単一の国名辞書 ($C = 24$, $D = 193$) の鎖長分布は未だに計算できていない．そこで，4.4 節と同様に複数の文字を 1 つの文字に圧縮することを考える．

文字数の圧縮とは，辞書ネットワークの複数の頂点を同一視することであった．シャッフル辞書群の場合は次数を足し合わせるだけでよかったが，単一辞書の場合は少し複雑になる．図 5.6 は $C = 4$ から $C = 3$ への圧縮の例で，頂点 c, d を 1 つの頂点 c+d に圧縮している．このとき，同一視される頂点間に有向辺があれば，それは自己ループに置き換えることとする．圧縮後の隣接行列は，圧縮前の隣接行列の青枠内の成分を足し合わせることで得ることができる．

同じ要領で国名辞書を $C = 5$ に圧縮してみる．実測で終了する回数の多い上位 4 文字 a, y, o, q を残し，他の 20 文字を θ_{others} の 1 文字に圧縮した．圧縮後の隣接行列 A' は以下のようになった．なお，行列の成分は a, y, o, q, θ_{others} の順に並べている．

$$A' = \begin{pmatrix} 9 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 63 & 6 & 11 & 1 & 98 \end{pmatrix}.$$

圧縮していない国名辞書における鎖長分布の実測値（100 万回実測）と $\Theta' =$

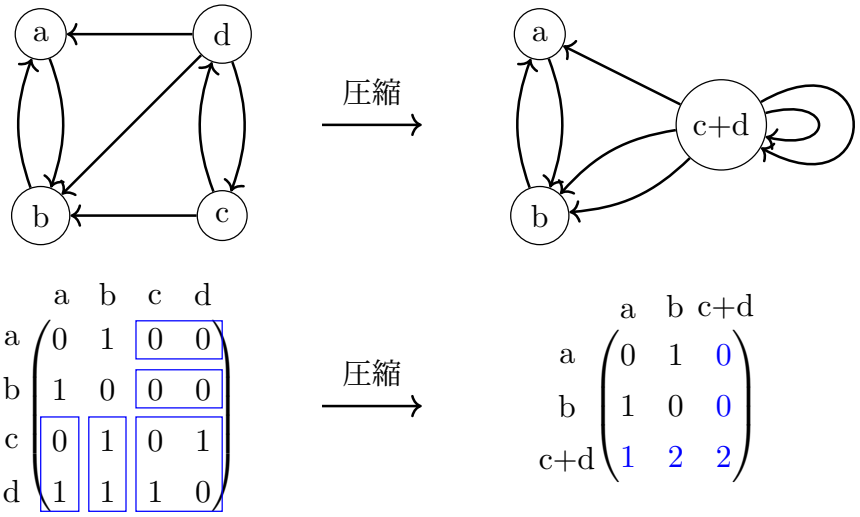


図 5.6: 文字数の圧縮と隣接行列の変化

$\{\mathrm{a}, \mathrm{y}, \mathrm{o}, \mathrm{q}, \theta_{\mathrm{others}}\}$ に圧縮した国名辞書における鎖長分布の理論値を比較した結果が図 5.7 である．シャッフル辞書群のときと異なり，全く一致しない結果となった．さらに，単一辞書の場合，終了不能文字どうしの圧縮でさえ鎖長分布が変化する可能性があり，単一辞書においては圧縮がそれほど有益な手段にならないと考えられる．ちなみに，実測値も $\Theta' = \{\mathrm{a}, \mathrm{y}, \mathrm{o}, \mathrm{q}, \theta_{\mathrm{others}}\}$ に圧縮した辞書で 100 万回しりとりを実行したところ，こちらは先ほどの理論値によく一致する結果となった (図 5.8).

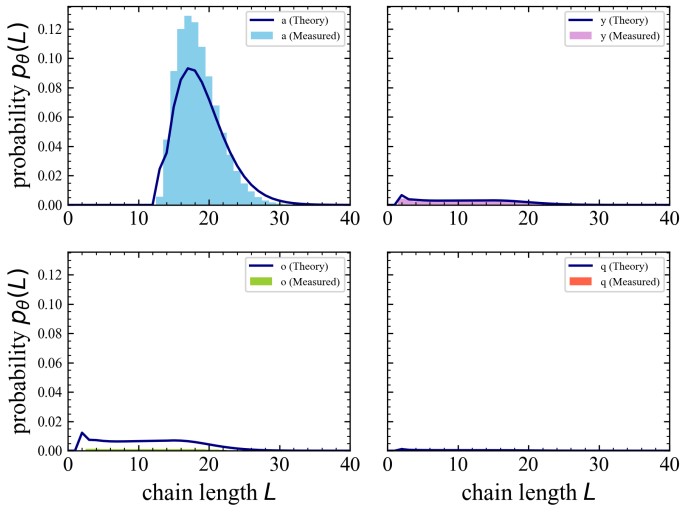


図 5.7: 国名辞書について，$C = 5$ に圧縮した鎖長分布の理論値（実線）と圧縮していない鎖長分布の実測値（ヒストグラム）の比較

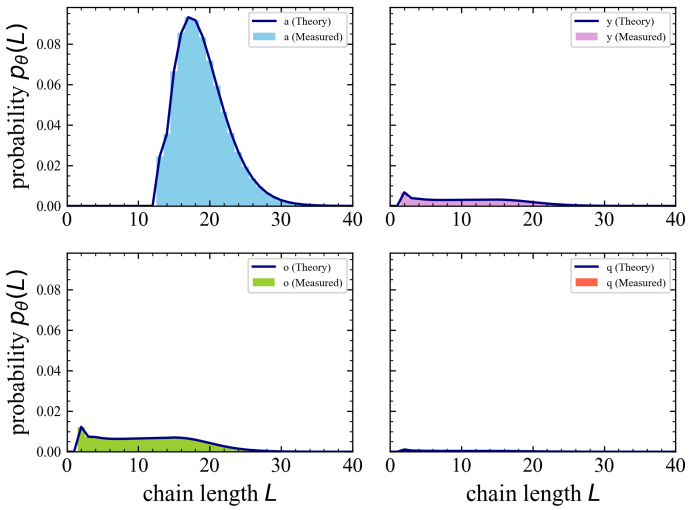


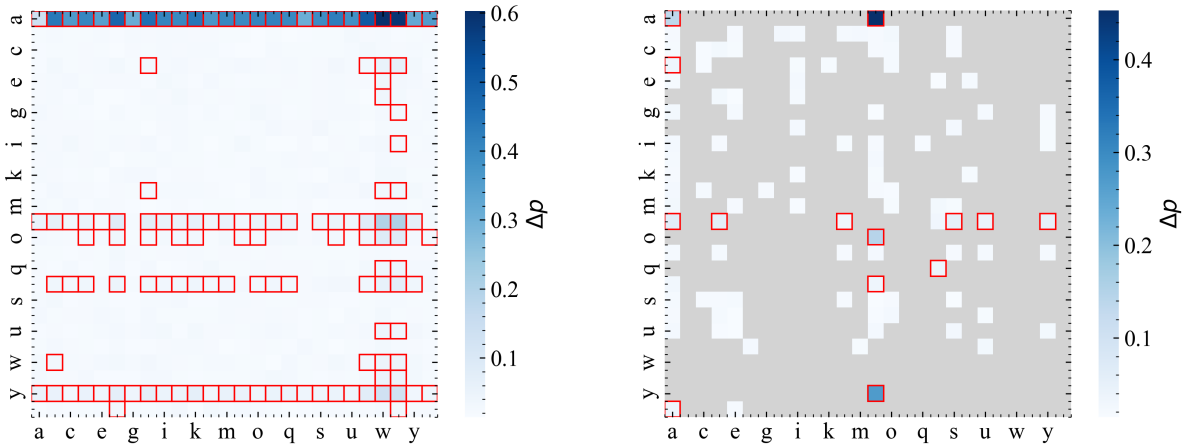
図 5.8: 国名辞書について，$C = 5$ に圧縮した鎖長分布の理論値（実線）と $C = 5$ に圧縮した鎖長分布の実測値（ヒストグラム）の比較

5.5 単一辞書が難しい理由

国名辞書のシャッフル辞書群における鎖長分布を良い精度で求めることができた一方で，国名辞書そのものにおける鎖長分布を求めることはできていない．単一辞書における鎖長分布の導出はなぜ難しいのだろうか．

国名辞書に単語を 1 つ追加したときの鎖長分布の変化を調べてみる．国名辞書の鎖長分布を $p(L)$ とし，θ で始まり ϕ で終わる単語を追加した辞書での鎖長分布を $p_{+\varepsilon}(L)$ とする．任意の $\theta, \phi \in \Theta$ について，$p_{+\varepsilon}(L)$ と $p(L)$ の全変動距離 $\Delta p_{+\varepsilon} = d_{TV}(p_{+\varepsilon}, p)$ を測定した結果が図 5.9 (a) である．$p(L)$，$p_{+\varepsilon}(L)$ はしりとりを 1 万回実行して求めた．赤枠は特に $\Delta p_{+\varepsilon}$ が大きい箇所を表している．例えば，多くの単語は追加しても分布がほとんど変化しないが，a や y から始まる単語は追加すると鎖長分布が大きく変化することがわかる．同様にして，単語を 1 つ削除したときの鎖長分布 $p_{-\varepsilon}(L)$ と $p(L)$ との全変動距離 $\Delta p_{-\varepsilon} = d_{TV}(p_{-\varepsilon}, p)$ は図 5.9 (b) のように測定される．灰色の箇所は対象の単語が辞書にない（削除できない）ことを表している．この場合も，多くの単語は削除による変化がほとんどないが，“a$\cdots$n”（a で始まり n で終わる単語）や “y$\cdots$n” を削除すると鎖長分布が大きく変化することがわかる．

実際に単語を追加，削除したときの鎖長分布を調べてみる．図 5.10 は国名辞書に対して，$\Delta p_{+\varepsilon}$ が大きい単語（“a$\cdots$w”, “n$\cdots$w”）と小さい単語（“b$\cdots$w”）を追加した場合の鎖長分布（中段），および $\Delta p_{-\varepsilon}$ が大きい単語（“a$\cdots$n”, “y$\cdots$n”）と小さい単語（“b$\cdots$n”）を削除した場合の鎖長分布（下段）を，元の鎖長分布（上段）と比較した結果である（いずれも 10 万回の実測により求めた）．“a$\cdots$w”, “n$\cdots$w” を追加した場合はいずれも終了文字 w が出



(a) $\Delta p_{+\varepsilon}$　　(b) $\Delta p_{-\varepsilon}$

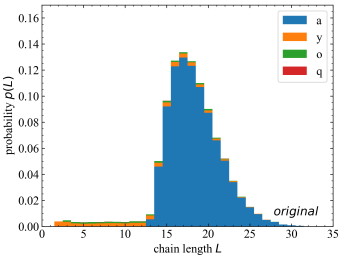
図 5.9: 国名辞書に 1 単語追加，削除したときの鎖長分布の変化

現する[*1]．しかし，“n · · · w” では依然として a で終了する確率が最も高い一方で，“a · · · w” では a で一度も終了しないという違いが生じた．また，“a · · · n” を削除した場合は分布全体が左側へ移動するが，“y · · · n” を削除した場合すぐに終了する確率が高くなる結果となった．このように追加，削除する単語によって変化の仕方が大きく異なることが確認された．

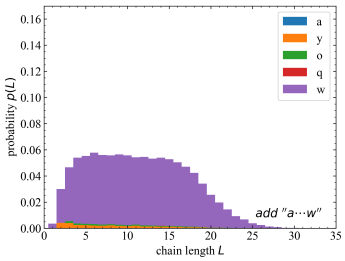
以上より，単語の選択の仕方によって鎖長分布は大きく変化し，その変化の仕方も様々であることが判明した．1 つの単語を取り除くという操作がしりとりの 1 ステップである[*2]ことを考えると，次にどの単語を選択するかによって，その先の分布が大きく変動するか否かが決まるのである．この複雑さこそが単一辞書の理論的な取り扱いを困難にする一因であると同時に，しりとりの面白さの源ではないかと考えられる．

[*1] w と x は元々の文字集合 Θ に含まれていない文字なので，“a · · · w” を追加することは C を増やすことにもなる．さらに，$k_{\mathrm{out,w}} = 0$ であることからこの単語を選ぶとすぐに終了してしまう．日本語の「ん」に対応すると言ってもよいだろう．

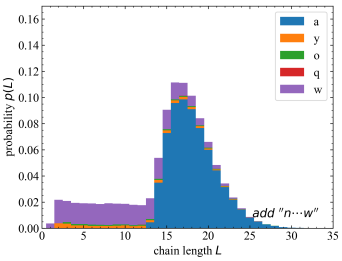
[*2] 厳密には初期条件に違いが出る．例えば，“a · · · n” を削除した辞書では最初の文字がランダムに決まるが，“a · · · n” を使用した後だと次は n から始めなければならない．



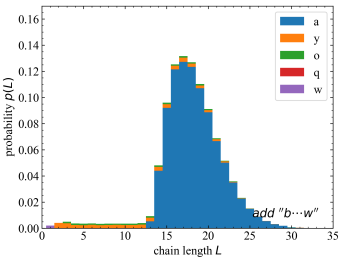
(a) 元の鎖長分布



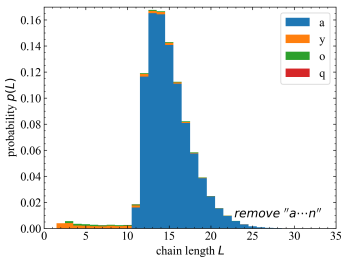
(b) “a · · · w” を追加した場合



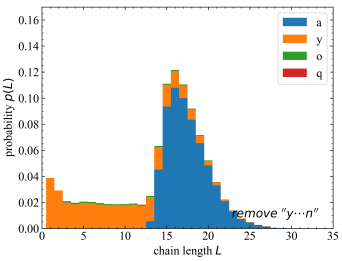
(c) “n · · · w” を追加した場合



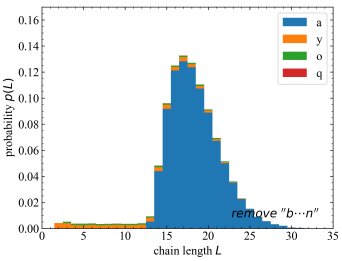
(d) “b · · · w” を追加した場合



(e) “a · · · n” を削除した場合



(f) “y · · · n” を削除した場合



(g) “b · · · n” を削除した場合

図 5.10: 国名辞書に 1 単語追加，削除したときの鎖長分布の比較（各 10 万回実測）
(b)(c) および (d) は 1 単語を追加することで鎖長分布が大きく変動する 2 例とほとんど変動しない 1 例．(e)(f) および (g) は 1 単語を削除することで鎖長分布が大きく変動する 2 例とほとんど変動しない 1 例．

第 6 章

結論

6.1 まとめ

本研究では勝敗や戦略を考えないランダムなしりとりに着目し，辞書ネットワーク上の自己回避的なランダムウォークとして定式化した上で解析を行った．得られた解析結果を，(i) しりとり全体について，(ii) シャッフル辞書群について，(iii) 単一辞書についての 3 つの観点からまとめる．特に鎖長分布の解析結果に関しては表 6.1 に記載している．

(i) しりとり全体については，終了文字 θ は必ず $k_{\mathrm{in},\theta} \geq k_{\mathrm{out},\theta}$ を満たすことが判明した．言い換えると，$k_{\mathrm{in},\theta} < k_{\mathrm{out},\theta}$ を満たす文字 θ でしりとりは終了しない．この性質は $k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面さえ描けば，しりとりを実行しなくてもどの文字で終了可能なのかがわかるという点で強力な性質であると言える．

(ii) シャッフル辞書群については，鎖長分布を解析的に求めることができた．$C = 2$ の場合，鎖長分布は負の超幾何分布に従うことが判明した．さらに，カバーレートの最大値が終了可能文字の出次数と入次数の差で決まり，最小値が出次数のみで決まり，平均値が出次数と入次数の比で決まることが明らかとなった．$C \geq 3$ の場合，鎖長分布は多変量逆超幾何分布に従うことが判明した．分布は文字の種類数 C に対して $C-2$ 重和を含むため，大きな C については膨大な計算時間を必要とする．文字数が大きい場合には，文字数の圧縮を行うことで，高い精度で近似された鎖長分布を効率的に得られることが確認された．例えば，国名辞書 ($C = 24$) のシャッフル辞書群に関して，$C = 5$ に圧縮したときの鎖長分布の理論値は，圧縮していないときの実測値によく一致する結果となった．なお，$C \geq 3$ の分布統計量は今のところ解析的に求められていない．

(iii) 単一辞書については，鎖長分布を計算するアルゴリズムを構築した．$C = 2$ の場合のみ解析解を得ることができ，シャッフル辞書群と同様に負の超幾何分布に従うことが判明した．さらに，分布統計量も計算することができた．$C \geq 3$ の場合，鎖長分布を求めるには単語列の総当たりが必要となるが，経路行列の縮退度を利用することで計算時間を短縮できることが明らかとなった．しかし，文字数や単語数が大きい場合，鎖長分布を計算するのに依然として時間がかかってしまい，未だに国名辞書そのものにおける分布を求めることはでき

ていない．さらに，単一辞書では文字数の圧縮が役に立たず，国名辞書を $C = 5$ に圧縮した鎖長分布が，圧縮していないときの分布から大きくずれる結果となった．単一辞書の計算が困難である理由の一つとして，単語の選択の仕方によって分布が大きく変化してしまうことが影響しているのではないかと考えられる．

表 6.1: 鎖長分布の解析結果

	単一辞書	シャッフル辞書群
$C = 2$	負の超幾何分布	負の超幾何分布
$C \geq 3$	経路行列の総当たりにより計算可能 （C, D が小さい場合）	多変量逆超幾何分布

6.2 今後の課題

今後の課題としては，一つ目に $C \geq 3$ のシャッフル辞書群における分布統計量の導出が挙げられる．$C = 2$ で計算したカバーレートの統計量だが，$C \geq 3$ の場合にどうなるのかはまだ判明していない．それに加えて，$C \geq 3$ では各終了可能文字における終了確率を考えることができ，こちらも求めることはできていない．もし，終了確率を解析的に導き出すことができれば，$k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面の情報から終了可能文字のみならず，どの文字で終了しやすいかまで予言できることになる．

二つ目に国名辞書そのものの鎖長分布を求めることが挙げられる．今のところブロック法は $C \geq 3$ の単一辞書に適用できない．BEST 定理などを利用して $C \geq 3$ でもブロック法が使えるようになれば，$C = 2$ と同様に鎖長分布を数式で表現できるようになる可能性がある．ただし，国名辞書は $C = 24$ であり，文字数の圧縮が単一辞書に不向きであることを踏まえると，やはり厳しい問題設定であるかもしれない．

三つ目に鎖長分布の形状を理解することが挙げられる．国名辞書の鎖長分布はどの文字もある値で急激に立ち上がった後，緩やかに減衰していく様子が見られる．一方で，Moby-Dick 辞書の鎖長分布は比較的緩やかに増大し，緩やかに減少している．このような立ち上がり方の個別性と減衰の仕方の普遍性がなぜ生じるのかを今後明らかにしたい．

四つ目に先行研究との比較が挙げられる．先行研究では，いくつかの典型的なネットワーク上で SAW を実行したときの経路長分布が計算されている（第 2 章）．それらとしりとりの鎖長分布を比較することで，しりとりや辞書ネットワークの特徴を見出すことができるかもしれない．

五つ目に単一辞書の各単語が持つ影響力の分析が挙げられる．5.5 節では，隣接行列のどの成分を変化させれば分布が大きく変化するのかを数値的に調査したが，その構造の解明には至らなかった．また，辞書ネットワークの自己ループは，各文字の終了確率に影響しないが，分布の形状に影響を与えることが数値的に確認され，そのロジックは現在調査中である（付

録 F を参照)．このように単語ごとの影響力を把握することは，しりとりのゲーム的観点から非常に興味深い課題である．

六つ目にシャッフル辞書群と単一辞書の関係を理解することが挙げられる．国名辞書において単一辞書とシャッフル辞書群の鎖長分布間の全変動距離は 0.547 と大きな値を示す．したがって，国名辞書については平均場近似がよい近似になっているとはいえない．それではどのような辞書で平均場近似はよい近似となるのだろうか．これを突き止めるには，シャッフル辞書群と単一辞書がどのような関係にあるのかを理解する必要があるだろう．今後は両者の関係性をより鮮明なものにし，平均場近似が妥当な場合とはどのような場合か調査したいと考えている．

七つ目に辞書の単語数と鎖長の関係性が挙げられる．語彙数が増えるほどしりとりが長く続くことは容易に想像されるが，それらはどのような関数形で記述されるのだろうか．付録 G の結果を見ると，鎖長の最大値と平均値は単語数に対して劣線形な依存性を持つことが数値的に確認された．その冪指数がどのように決まるのかなどの理論研究が今後の課題として残されている．しりとりの長さから語彙数が測定できるようになれば，外国語教育や言語機能のリハビリに活用できるかもしれない．

最後に品詞や言語ごとの鎖長分布の特徴が挙げられる．辞書を構成する品詞が変化すると各文字の $k_{\mathrm{in}} - k_{\mathrm{out}}$ の関係が大きく変化し，それにより鎖長分布も変化すると考えられる．例えば，付録 H では英語の副詞からなる辞書で鎖長分布を実測したが，しりとりはほとんど y で終了する結果となった．これには，英語の副詞が圧倒的に y で終わりやすいことが関係していると見られる．さらに，言語を変えると $k_{\mathrm{in}} - k_{\mathrm{out}}$ の分布だけでなく，文字の種類数 C も変化する場合がある．例えば日本語の場合 $C \geq 46$ となり，付録 H の計測では 11 の終了文字が出現した．このように品詞や言語を変えたときにしりとりの統計的性質がどのようになるのかは興味深いテーマである．

謝辞

本研究の遂行にあたり，多くの方にお力添えをいただきました．指導教員の水口毅先生には，研究の進め方をはじめ，論文の読み方，発表資料の作り方，解析の仕方に至るまで，研究における読み書き算盤を一からご教示いただきました．本論文を形にできたのは，ひとえに先生の熱心なご指導の賜物です．ここに深く感謝の意を表します．高千穂大学の鈴木岳人先生には，論文執筆や発表練習といった研究活動の様々な場面で，数多くの有益なご助言を賜りました．ここに厚く御礼申し上げます．また，学内・学外の研究集会において，有意義な議論を交わしてくださった先生方や学生の皆様に，心より感謝いたします．非線形物理研究室の先輩方，同期，後輩の皆様とは，研究室セミナーや発表練習，さらにはネットワーク科学の勉強会で数多くの議論を交わしました．多様な視点から自身の研究を深めることができ，大変感謝しております．最後に学生生活を温かく見守り支援してくれた家族への感謝の意を表し，本論文の結びといたします．

付録 A

国名辞書について

本付録では国名辞書の構成方法を詳しく述べ，辞書ネットワークとその次数分布を記載する．

A.1　国名辞書の構成方法

国名辞書は 2025 年 5 月時点での国際連合加盟国 193 カ国の英語名 [31] からなる辞書である．ただし，すべての単語に対して次の 2 つの処理を施した．

(1) 大文字は小文字に変換する．　**例**．Japan → japan
(2) 括弧とその中身は削除する．　**例**．netherlands (kingdom of the) → netherlands

以下の表は国名辞書を構成する単語の一覧である．

表 A.1: 国名辞書の単語一覧

afghanistan	bahrain	brunei darussalam
albania	bangladesh	bulgaria
algeria	barbados	burkina faso
andorra	belarus	burundi
angola	belgium	cabo verde
antigua and barbuda	belize	cambodia
argentina	benin	cameroon
armenia	bhutan	canada
australia	bolivia	central african republic
austria	bosnia and herzegovina	chad
azerbaijan	botswana	chile
bahamas	brazil	china

colombia	guatemala	malawi
comoros	guinea	malaysia
congo	guinea bissau	maldives
costa rica	guyana	mali
côte d'ivoire	haiti	malta
croatia	honduras	marshall islands
cuba	hungary	mauritania
cyprus	iceland	mauritius
czechia	india	mexico
democratic people's republic of korea	indonesia	micronesia
democratic republic of the congo	iran	monaco
denmark	iraq	mongolia
djibouti	ireland	montenegro
dominica	israel	morocco
dominican republic	italy	mozambique
ecuador	jamaica	myanmar
egypt	japan	namibia
el salvador	jordan	nauru
equatorial guinea	kazakhstan	nepal
eritrea	kenya	netherlands
estonia	kiribati	new zealand
eswatini	kuwait	nicaragua
ethiopia	kyrgyzstan	niger
fiji	lao people's democratic republic	nigeria
finland	latvia	north macedonia
france	lebanon	norway
gabon	lesotho	oman
gambia	liberia	pakistan
georgia	libya	palau
germany	liechtenstein	panama
ghana	lithuania	papua new guinea
greece	luxembourg	paraguay
grenada	madagascar	peru

philippines	singapore	tunisia
poland	slovakia	türkiye
portugal	slovenia	turkmenistan
qatar	solomon islands	tuvalu
republic of korea	somalia	uganda
republic of moldova	south africa	ukraine
romania	south sudan	united arab emirates
russian federation	spain	united kingdom of great britain and northern ireland
rwanda	sri lanka	united republic of tanzania
saint kitts and nevis	sudan	united states of america
saint lucia	suriname	uruguay
saint vincent and the grenadines	sweden	uzbekistan
samoa	switzerland	vanuatu
san marino	syrian arab republic	venezuela, bolivarian republic of
sao tome and principe	tajikistan	viet nam
saudi arabia	thailand	yemen
senegal	timor-leste	zambia
serbia	togo	zimbabwe
seychelles	tonga	
sierra leone	trinidad and tobago	

A.2 国名辞書の次数分布

国名辞書の辞書ネットワークは図 A.1 のように表される．(a) が単語頂点型ネットワーク，(b) が単語辺型ネットワークである．このうち，単語頂点型ネットワークが典型的な次数分布を持っていれば，2.4 節で紹介した先行研究が直接的に利用できると考えられる．

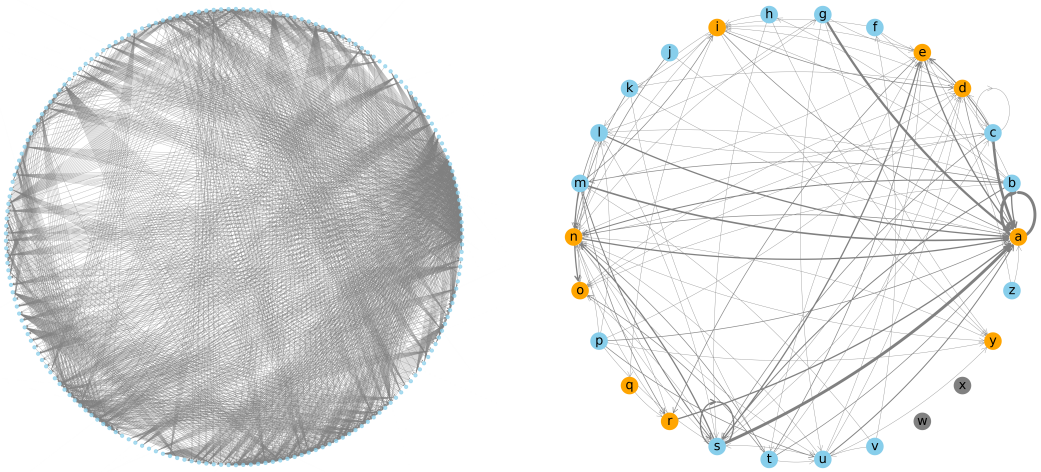
国名辞書の単語頂点型ネットワークにおける入次数分布 $P_{\kappa_{\mathrm{in}}}$ と出次数分布 $P_{\kappa_{\mathrm{out}}}$ は図 A.2 のように表される．いずれもポアソン分布や冪乗則のようなよく見られる分布にはならず，とびとびの値をとっていることが確認できる．この構造は辞書によらず出現する．

文字 θ で始まり文字 ϕ で終わる単語の入次数を $\kappa_{\mathrm{in},\theta\phi}$，出次数を $\kappa_{\mathrm{out},\theta\phi}$ とすると，

$$\kappa_{\mathrm{in},\theta\phi} = k_{\mathrm{in},\theta}, \quad \kappa_{\mathrm{out},\theta\phi} = k_{\mathrm{out},\phi} \quad (\theta \neq \phi), \tag{A.1}$$

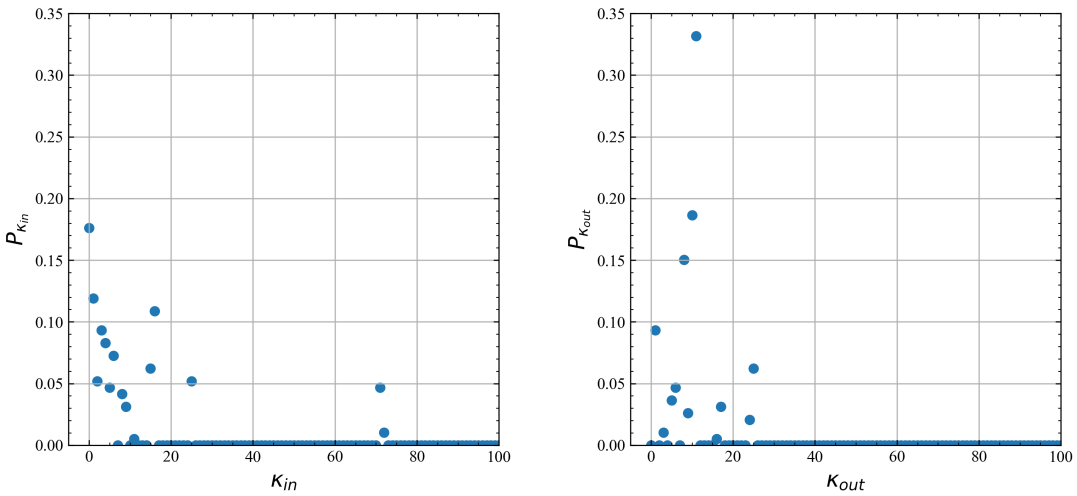
$$\kappa_{\mathrm{in},\theta\phi} = k_{\mathrm{in},\theta} - 1, \quad \kappa_{\mathrm{out},\theta\phi} = k_{\mathrm{out},\phi} - 1 \quad (\theta = \phi) \tag{A.2}$$

が成り立つ．したがって，単語の入次数や出次数は最大で $2C$ 個の値しか持たないため，単語頂点型の次数分布がとびとびの構造をとることがわかる．



(a) 単語頂点型ネットワーク　　(b) 単語辺型ネットワーク

図 A.1: 国名辞書の辞書ネットワーク



(a) 入次数分布 $P_{\kappa_{\mathrm{in}}}$　　(b) 出次数分布 $P_{\kappa_{\mathrm{out}}}$

図 A.2: 国名辞書の単語頂点型ネットワークの次数分布

付録 B

Moby-Dick 辞書について

辞書が変わると鎖長分布は全く別物となる．本付録では「Moby-Dick 辞書」を導入し，その鎖長分布が国名辞書の分布と大きく異なることを確認する．

B.1　Moby-Dick 辞書とその性質

Moby-Dick 辞書は米国の文学作品 “Moby-Dick” [32] に登場する英語の名詞 19088 語からなる辞書である．鎖長分布を見る前に，Moby-Dick 辞書の性質を確認しておく．図 B.1 は Moby-Dick 辞書の辞書ネットワークである．国名辞書と異なり孤立点は存在せず，$C = 26$ である．また，図 B.2 は各文字の $k_{\mathrm{in}}, k_{\mathrm{out}}$ の関係を表している．この図から終了可能文字は黄色の領域にある 8 文字であることがわかる．

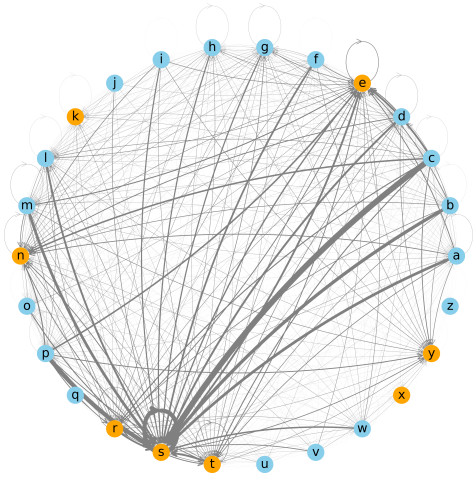


図 B.1: Moby-Dick 辞書の辞書ネットワーク　辺の太さは有向辺の数に比例している．

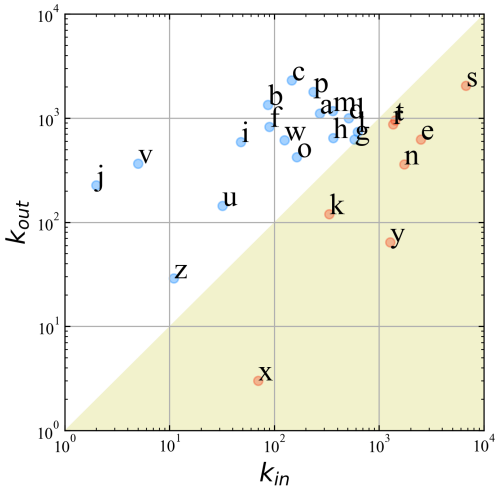
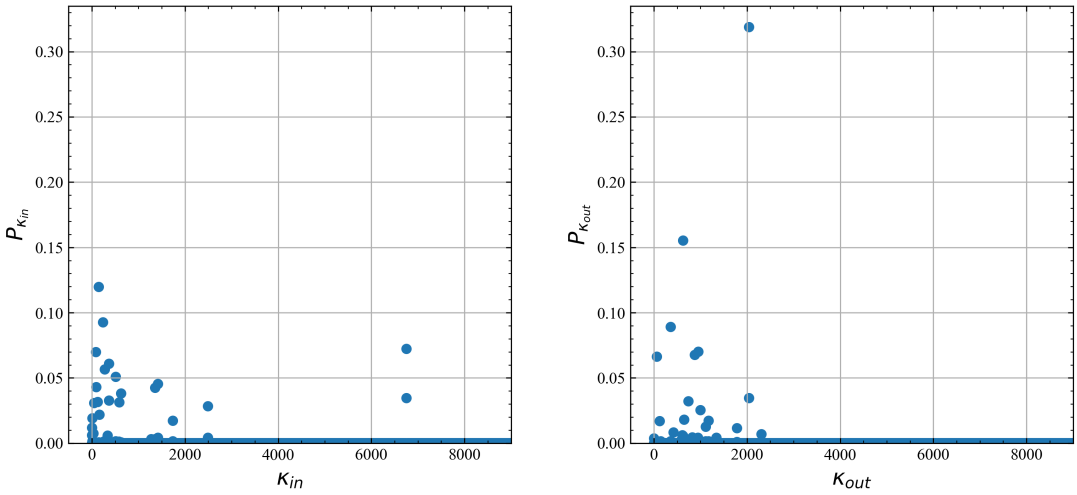


図 B.2: Moby-Dick 辞書の $k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面　両対数プロットである点に注意．

Moby-Dick 辞書の単語頂点型ネットワークにおける入次数分布 $P_{\kappa_{\mathrm{in}}}$ と出次数分布 $P_{\kappa_{\mathrm{out}}}$ は図 B.3 のようになった．国名辞書と同様にとびとびの構造を持つことが確認できる．

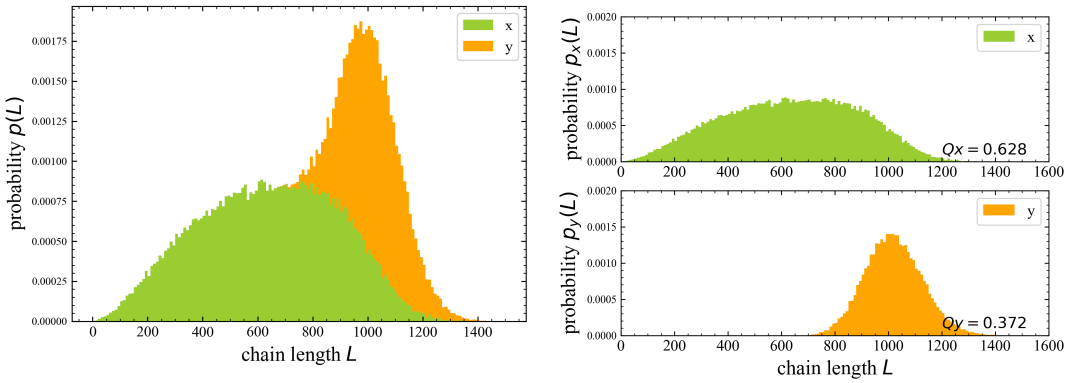


(a) 入次数分布 $P_{\kappa_{\mathrm{in}}}$　　(b) 出次数分布 $P_{\kappa_{\mathrm{out}}}$

図 B.3: Moby-Dick 辞書の単語頂点型ネットワークの次数分布

B.2　Moby-Dick 辞書の鎖長分布

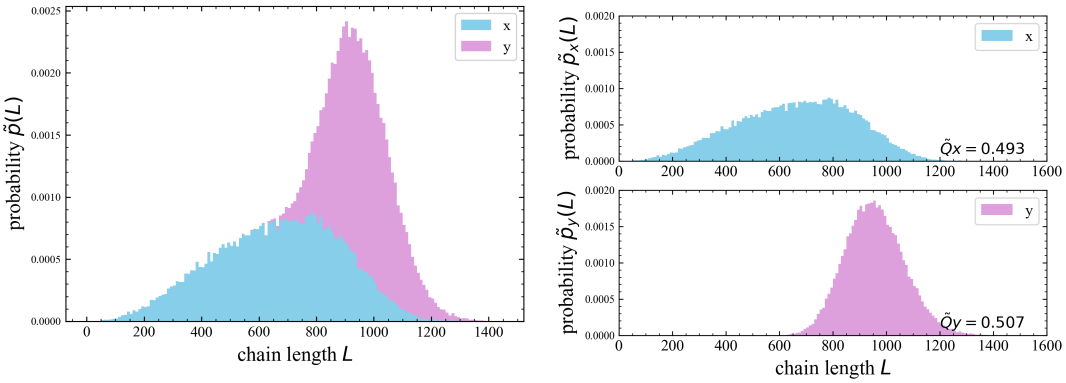
Moby-Dick 辞書を用いて 10 万回ランダムなしりとりを実行したときの鎖長分布は図 B.4 のようになった．国名辞書の鎖長分布と形状が大きく異なることがわかる．そして，10 万回実行して現れた終了文字は x, y の 2 文字のみであった．



(a) 終了文字ごとに色分けした分布　　(b) 終了文字ごとの分布

図 B.4: Moby-Dick 辞書の鎖長分布（10 万回実測）

Moby-Dick 辞書のシャッフル辞書を 1000 冊作成し，それぞれで 100 回ランダムなしりとりを実行したときの平均鎖長分布は図 B.5 のようになった．国名辞書と違い，現れた終了文字の種類は変化しなかった．ただし，単一辞書では x で終了しやすかったが，シャッフル辞書群ではわずかに y で終了しやすくなる．



(a) 終了文字ごとに色分けした分布　　(b) 終了文字ごとの分布

図 B.5: Moby-Dick 辞書のシャッフル辞書群の鎖長分布（1000 冊 × 各 100 回実測）

Moby-Dick 辞書のシャッフル辞書群の解析解は，$\Theta = \{\mathrm{x}, \mathrm{y}, \text{他}\}$ に圧縮することで，図 B.6 のようにかなりよい精度で求めることができた．一方で，単一辞書における鎖長分布の厳密解は未だに求めることができていない．

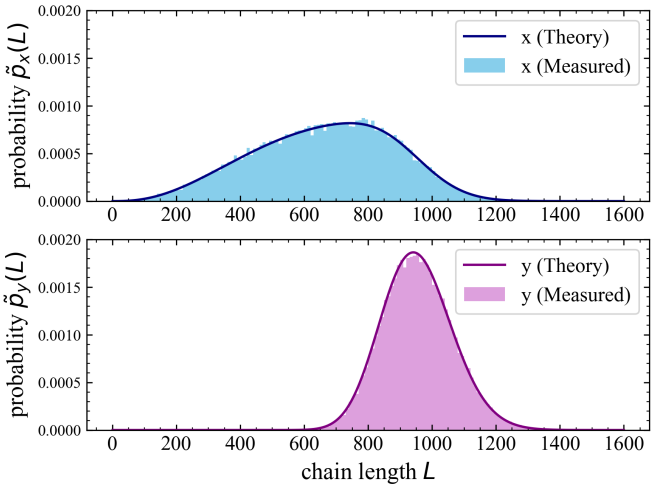


図 B.6: Moby-Dick 辞書のシャッフル辞書群の鎖長分布（ヒストグラム：実測値，実線：理論値）

付録 C

2 部グラフの射影

本付録では 2 部グラフの射影について説明する．そして，辞書ネットワークを単語と文字の有向 2 部グラフで表現したとき，その射影が本論文で扱っている 2 つの辞書ネットワークに対応することを記述する．

頂点集合を 2 つの部分集合に分割したとき，すべての辺が異なる集合にある頂点同士を結んでいるようなグラフを，2 部グラフと呼ぶ．図 C.1 の中央のグラフが 2 部グラフの例であり，頂点集合 {1, 2, 3, 4, 5, 6, A, B, C, D} を $U = \{1, 2, 3, 4, 5, 6\}$, $V = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ に分割すると，すべての辺は U の頂点と V の頂点を結んでおり，同じ部分集合内の頂点を結ぶ辺は存在していない．具体例として，U に俳優の集合，V に映画の集合をとり，俳優が映画に出演していたらそれらの映画と俳優の間に辺を張ることで，2 部グラフを構築することができる．

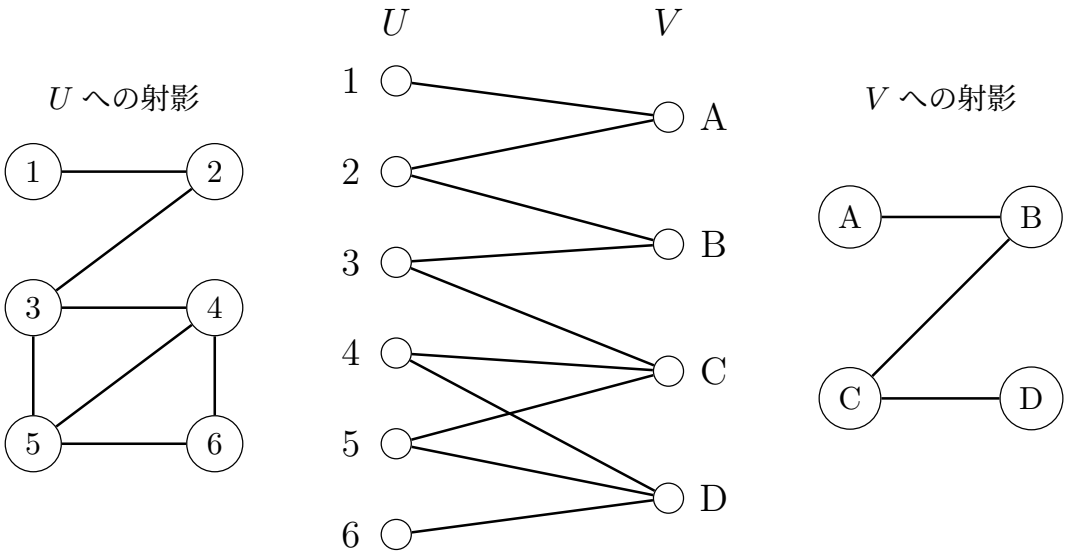


図 C.1: 無向 2 部グラフの射影の例

2 部グラフから 2 つの射影を得ることができる [42]．U への射影は U の頂点をすべての頂点とし，U に属する 2 頂点が同じ V の頂点につながっているとき，その 2 頂点の間に辺を

張ったグラフである．同様にして V への射影も構築される．図 C.1 の左右のグラフは，中央の 2 部グラフの U, V への射影である．先ほどの例において，俳優の集合への射影は同じ映画に出演した俳優が接続される「俳優の共演ネットワーク」となり，映画の集合への射影は同じ俳優が出演している映画が接続される「映画ネットワーク」となる．

ここからは辞書から構成される有向な 2 部グラフを考える．辞書に使われる単語と文字を頂点にとり，文字からはその文字で始まるすべての単語へ，単語からはその末尾文字への有向辺を張る．このグラフは，単語の集合 W と文字の集合 Θ で分割される 2 部グラフとなっている．例えば，辞書 {area, absorb, bacteria, bomb, basic, camera, climb, club} から構成される 2 部グラフは図 C.2 のように表される．

次に辞書の 2 部グラフの射影を考える．W を頂点集合とし，任意の頂点 $w_1, w_2 \in W$ について，w_1 を始点とする有向辺の終点でありかつ w_2 を終点とする有向辺の始点であるような Θ の頂点が 2 部グラフに存在していれば，w_1 から w_2 への有向辺を張る．こうして構成されるグラフを W への射影と呼ぶことにする（Θ への射影も同様）．そうすると，辞書の 2 部グラフの W への射影は単語頂点型ネットワークに対応し，Θ への射影は単語辺型ネットワークに対応することがわかる（図 C.2）．ただし，通常の 2 部グラフと異なり，いずれの射影も多重辺を許容し，Θ への射影についてのみ自己ループも認めるものとする．

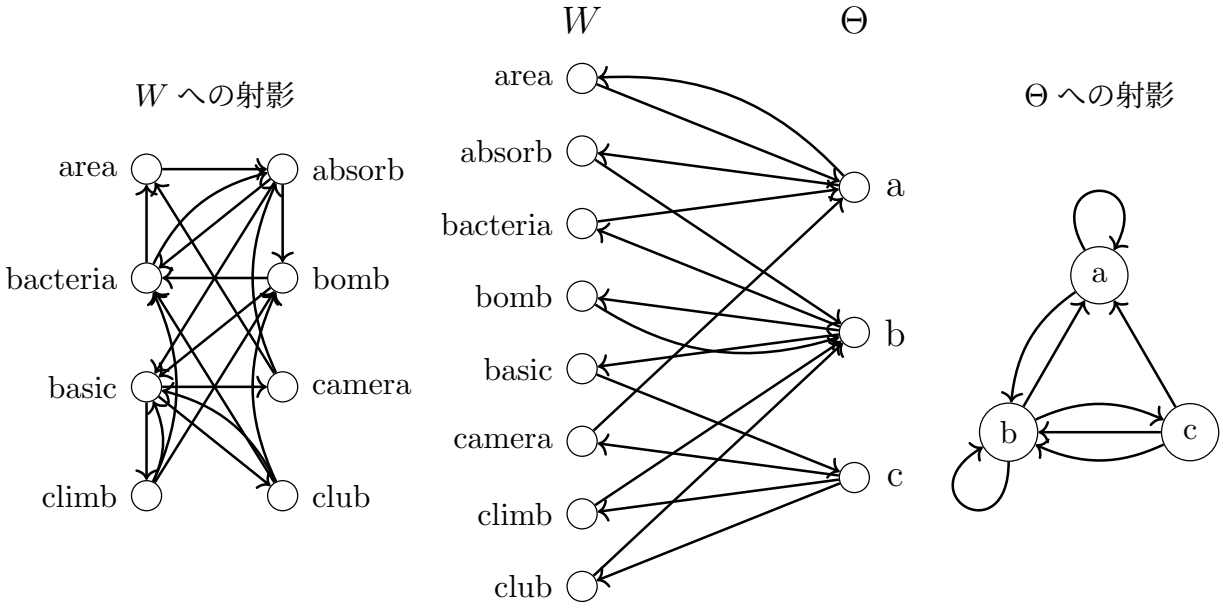


図 C.2: 有向 2 部グラフの射影の例

付録 D

負の超幾何分布の諸性質

本付録では負の超幾何分布に関する 2 つの性質 (4.11), (4.12) 式が成立することを証明する．なお，以下の証明は [41] を参考にした．これとは別の証明方法が [37] [38] に記載されている．まずは定理を証明するための公式を示しておく．

補題 D.0.1. 次の 3 つの公式が成り立つ．

(a) $\binom{n}{k} = (-1)^k \binom{k-n-1}{k}$

(b) $\sum_{j=0}^{k} \binom{m}{j}\binom{n-m}{k-j} = \binom{n}{k}$

(c) $\sum_{j=0}^{k} \binom{j+m}{j}\binom{n-m-j}{k-j} = \binom{n+1}{k}$

証明. (a)

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!} \\
&= \frac{(-k+n+1)(-k+n+2)\cdots(-k+n+k)}{k!} \\
&= (-1)^k \frac{(k-n-1)(k-n-2)\cdots(k-n-k)}{k!} \\
&= (-1)^k \frac{(k-n-1)!}{k!(-n-1)!} \\
&= (-1)^k \binom{k-n-1}{k}.
\end{aligned}$$

(b) m 次多項式と n 次多項式の積は次のようにかける．ただし，$i > m$ のとき $a_i = 0$, $j > n$ のとき $b_j = 0$ とする．

$$
\begin{aligned}
&\left(\sum_{i=0}^{m} a_i x^i\right)\left(\sum_{j=0}^{n} b_j x^j\right) \\
&= (a_0x^0 + a_1x^1 + \cdots + a_mx^m)(b_0x^0 + b_1x^1 + \cdots + b_nx^n) \\
&= a_0b_0x^0 + (a_0b_1 + a_1b_0)x^1 + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots + a_mb_nx^{m+n} \\
&= \sum_{r=0}^{m+n}\left(\sum_{k=0}^{r} a_k b_{r-k}\right)x^r.
\end{aligned}
$$

また二項定理より以下が成り立つ.

$$
(1+x)^{m+n} = \sum_{r=0}^{m+n}\binom{m+n}{r}x^r.
$$

これらの等式から,

$$
\begin{aligned}
\sum_{r=0}^{m+n}\binom{m+n}{r}x^r &= (1+x)^{m+n} \\
&= (1+x)^m(1+x)^n \\
&= \left(\sum_{i=0}^{m}\binom{m}{i}x^i\right)\left(\sum_{j=0}^{n}\binom{n}{j}x^j\right) \\
&= \sum_{r=0}^{m+n}\left(\sum_{k=0}^{r}\binom{m}{k}\binom{n}{r-k}\right)x^r
\end{aligned}
$$

が導かれ，x^r の係数を比較することで,

$$
\binom{m+n}{r} = \sum_{k=0}^{r}\binom{m}{k}\binom{n}{r-k}
$$

が得られる．ゆえに,

$$
\sum_{j=0}^{k}\binom{m}{j}\binom{n-m}{k-j} = \binom{n}{k}.
$$

(c)

$$
\begin{aligned}
\sum_{j=0}^{k}\binom{j+m}{j}\binom{n-m-j}{k-j} &= \sum_{j=0}^{k}(-1)^j\binom{-m-1}{j}(-1)^{k-j}\binom{k-n+m-1}{k-j} \\
&= (-1)^k\sum_{j=0}^{k}\binom{-m-1}{j}\binom{k-n-2-(-m-1)}{k-j} \\
&= (-1)^k\binom{k-n-2}{k} = (-1)^k\binom{k-(n+1)-1}{k} \\
&= \binom{n+1}{k}.
\end{aligned}
$$

1 つ目と 5 つ目の等号で公式 (a) を，3 つ目の等号で公式 (b) を用いた. □

準備ができたので，負の超幾何分布に関する 2 つの性質を証明する．

定理 D.0.1. 負の超幾何分布の確率質量関数は以下の規格化条件を満たす．

$$\sum_{m_2=0}^{n_2} NHG(m_2; N, n_2, m_1) = \sum_{m_2=0}^{n_2} \frac{\binom{m_2+m_1-1}{m_2}\binom{N-m_2-m_1}{n_2-m_2}}{\binom{N}{n_2}} = 1. \tag{D.1}$$

証明.

$$\begin{aligned}\sum_{m_2=0}^{n_2} \frac{\binom{m_2+m_1-1}{m_2}\binom{N-m_2-m_1}{n_2-m_2}}{\binom{N}{n_2}} &= \frac{1}{\binom{N}{n_2}} \sum_{m_2=0}^{n_2} \binom{m_2+(m_1-1)}{m_2}\binom{(N-1)-(m_1-1)-m_2}{n_2-m_2} \\ &= \frac{1}{\binom{N}{n_2}}\binom{N}{n_2} = 1.\end{aligned}$$

2 つ目の等号で補題 D.0.1 の公式 (c) を用いた． □

定理 D.0.2. 確率変数 m_2 が $NHG(m_2; N, n_2, m_1)$ に従うとき，その平均値 $E[m_2]$ は以下のように与えられる．

$$E[m_2] = m_1 \frac{n_2}{N - n_2 + 1} \tag{D.2}$$

証明.

$$\begin{aligned}E[m_2] &= \sum_{m_2=0}^{n_2} m_2 \frac{\binom{m_2+m_1-1}{m_2}\binom{N-m_2-m_1}{n_2-m_2}}{\binom{N}{n_2}} \\ &= \frac{m_1}{\binom{N}{n_2}} \sum_{m_2=0}^{n_2} \frac{m_2}{m_1}\binom{m_2+m_1-1}{m_2}\binom{N-m_2-m_1}{n_2-m_2} \\ &= \frac{m_1}{\binom{N}{n_2}} \sum_{m_2=0}^{n_2} \left(\frac{m_2}{m_1}+1\right)\binom{m_2+m_1-1}{m_2}\binom{N-m_2-m_1}{n_2-m_2} \\ &\quad - \frac{m_1}{\binom{N}{n_2}} \sum_{m_2=0}^{n_2} \binom{m_2+m_1-1}{m_2}\binom{N-m_2-m_1}{n_2-m_2} \\ &= \frac{m_1}{\binom{N}{n_2}} \sum_{m_2=0}^{n_2} \frac{m_2+m_1}{m_1}\binom{m_2+m_1-1}{m_1-1}\binom{N-m_2-m_1}{n_2-m_2} - m_1.\end{aligned}$$

最後の等号で定理 D.0.1 を用いた．さらに，変形を続けると，

$$\begin{aligned}E[m_2] &= \frac{m_1}{\binom{N}{n_2}} \sum_{m_2=0}^{n_2} \binom{m_2+m_1}{m_2}\binom{N-m_2-m_1}{n_2-m_2} - m_1 \\ &= \frac{m_1}{\binom{N}{n_2}}\binom{N+1}{n_2} - m_1 \\ &= m_1 \frac{n_2}{N-n_2+1}\end{aligned}$$

が得られる．2 つ目の等号で補題 D.0.1 の公式 (c) を用いた． □

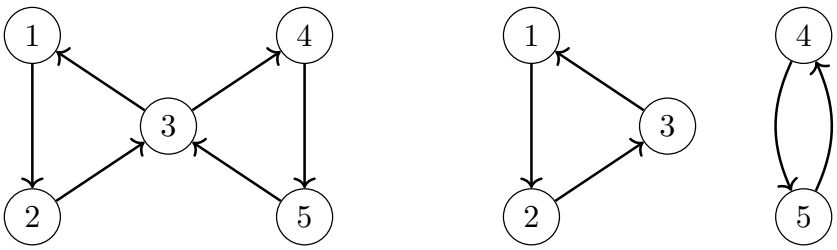
付録 E

BEST 定理

Euler グラフに含まれる Euler 回路の個数を求める定理が存在する．その定理は Smith と Tutte が発見 [39] し，Ehrenfest と de Bruijn が一般化 [40] したことから，4 人の名前の頭文字をとって「BEST 定理」と呼ばれている．本付録では BEST 定理とそれに関連する定理について紹介する．なお，本付録の内容は [43] を参考にした．最初に定理に登場するグラフ理論の用語について説明する．

E.1 BEST 定理について

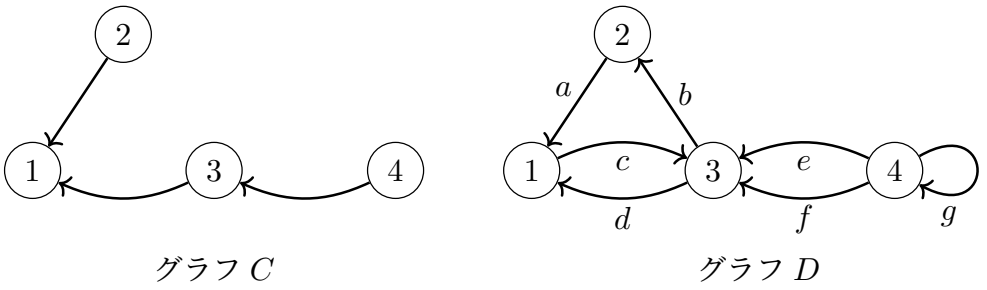
まずは BEST 定理が対象とするグラフの性質について整理する．多重有向グラフの任意の頂点について，入次数と出次数が等しいグラフを平衡なグラフと呼ぶ．図 E.1 のグラフ A, B はいずれも平衡なグラフの例である．次に，多重有向グラフのすべての有向辺を一度ずつ含む回路（始点と終点が一致する経路）のことを Euler 回路と呼び，Euler 回路を持つグラフのことを Euler グラフと呼ぶ．多重有向グラフ G が Euler グラフであるための必要十分条件は，G が弱連結かつ平衡であることである．図 E.1 の例では，グラフ A は弱連結かつ平衡であるため Euler グラフであるが，グラフ B は平衡ではあるものの弱連結ではないため Euler グラフではないと判断できる．BEST 定理は Euler グラフを含むすべての平衡なグラフに対し．Euler 回路の個数を導き出す定理である．



グラフ A　　グラフ B

図 E.1: Euler グラフと平衡なグラフ

続いて，全域有向木について説明する．多重有向グラフ G の頂点 r について，任意の頂点から r への路（頂点の重複がない経路）が存在し，かつ G を無向にしたグラフが閉路（閉じた路）を持たないとき，G を r を終根とする有向木と呼ぶ．例えば，図 E.2 のグラフ C は頂点 1 を終根とする有向木である．さらに，G の全域部分グラフ（頂点を保持したままいくつかの辺を除去して得られるグラフ）であって，r を終根とする有向木であるものを，r を終根とする G の全域有向木と呼ぶ．図 E.2 の有向木 C はグラフ D の辺 b, c, e, g を除去することによって得られるため，C は頂点 1 を終根とする D の全域有向木であると言える．



グラフ C　　グラフ D

図 E.2: 有向木と全域有向木

今から紹介する BEST 定理は，Euler 回路の個数を全域有向木の個数に紐付ける．

定理 E.1.1 (BEST 定理)**.** すべての頂点が 1 以上の出次数を持つ平衡な多重有向グラフを G，G の頂点集合を V とする．また，G の有向辺 a を任意にとり，a の始点を r とする．r を終根とする全域有向木の個数を $\tau(G, r)$ と表し，最初の有向辺が a である G の Euler 回路の個数を $\varepsilon(G, a)$ と表すとき，次の等式が成り立つ．

$$\varepsilon(G, a) = \tau(G, r) \prod_{u \in V} (k_{\mathrm{out},u} - 1)!. \tag{E.1}$$

証明は [43] を参照されたい．重要な点として，「最初の有向辺が a である G の Euler 回路の個数」は「回転して一致する回路を同一視するときの G が持つ Euler 回路の個数」に等しいことを述べておく．さらに，$\tau(G, r)$ が頂点 r の取り方によらず同じ値をとることが知られている（この性質は定理 E.1.1 から導かれる）．

E.2　行列木定理について

BEST 定理により，Euler 回路の個数を求める問題は全域有向木の個数を求める問題に置き換わった．実は後者を計算する定理も存在しており，行列木定理と呼ばれている．この定理の紹介に先立って，多重有向グラフの Laplacian を定義する．

定義 E.2.1. G を N 頂点の多重有向グラフとし，A を G の隣接行列とする．頂点 i, j につ

いて，次の式で定義される $N \times N$ 行列 L を G の Laplacian と呼ぶ．

$$L_{ij} = -A_{ij} \quad (i \neq j), \tag{E.2}$$
$$L_{ij} = k_{\mathrm{out},i} - A_{ij} \quad (i = j). \tag{E.3}$$

例えば，図 E.2 のグラフ D の Laplacian L は次のように求められる．

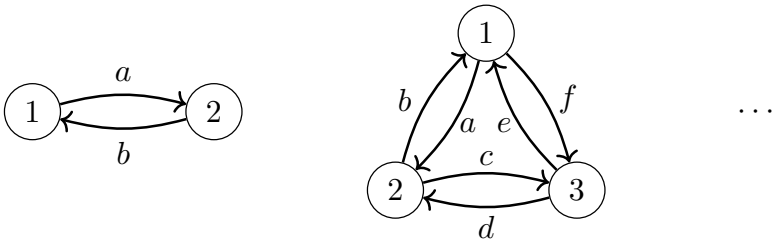
$$\begin{aligned} L &= \begin{pmatrix} k_{\mathrm{out},1} - A_{11} & -A_{12} & -A_{13} & -A_{14} \\ -A_{21} & k_{\mathrm{out},2} - A_{22} & -A_{23} & -A_{24} \\ -A_{31} & -A_{32} & k_{\mathrm{out},3} - A_{33} & -A_{34} \\ -A_{41} & -A_{42} & -A_{43} & k_{\mathrm{out},4} - A_{44} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & -2 & 2 \end{pmatrix}. \end{aligned}$$

全域有向木に関する以下の定理が成り立つ．ここで，$M_{\sim i \sim j}$ は行列 M から i 行目と j 列目を取り除いた行列を表している．

定理 E.2.1 (行列木定理)**.** G を多重有向グラフ，L を G の Laplacian，r を G の任意の頂点とする．このとき以下の等式が成り立つ．

$$(\#\ r \text{ を終根とする } G \text{ の全域有向木}) = \det(L_{\sim r \sim r}). \tag{E.4}$$

証明は [43] を参照されたい．BEST 定理と行列木定理を組み合わせることによって，平衡なグラフにおける Euler 回路の数を計算することが可能となる．1 つ例を見ておこう．図 E.3 のようなグラフ K_n^{bidir}（n は 2 以上の整数）の Euler 回路の個数を求めてみる．K_n^{bidir} は n 頂点完全グラフ K_n の各辺を双方向の有向辺に置き換えた多重有向グラフである．



K_2^{bidir}　　K_3^{bidir}

図 E.3: 有向完全グラフ K_n^{bidir}

まずは全域有向木の個数 $\tau(K_n^{\mathrm{bidir}}, r)$ を求める．平衡なグラフにおいて $\tau(K_n^{\mathrm{bidir}}, r)$ は頂点 r によらず同じ値をとるため，終根に頂点 1 をとることにする．K_n^{bidir} の Laplacian L は，

$$L = \begin{pmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & n-1 \end{pmatrix}$$

と書けるので，$\tau(K_n^{\mathrm{bidir}}, 1)$ は (E.4) 式を用いて以下のように計算される．

$$
\begin{aligned}
\tau(K_n^{\mathrm{bidir}}, 1) &= \det(L_{\sim 1 \sim 1}) \\
&= \det \begin{pmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & n-1 \end{pmatrix} \\
&= \det \begin{pmatrix} n-1 & -1 & -1 & -1 & \dots & -1 \\ -n & n & 0 & 0 & \dots & 0 \\ -n & 0 & n & 0 & \dots & 0 \\ -n & 0 & 0 & n & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -n & 0 & 0 & 0 & \dots & n \end{pmatrix} \\
&= n^{n-2} \det \begin{pmatrix} n-1 & -1 & -1 & -1 & \dots & -1 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 \\ -1 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \\
&= n^{n-2} \det \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 \\ -1 & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix} \\
&= n^{n-2}.
\end{aligned}
$$

行列式は行基本変形を繰り返すことにより求めている．なお，式中に登場する行列は $(n-1) \times (n-1)$ 行列であることに注意．さらに (E.1) 式を適用することにより，K_n^{bidir} の任意の有向辺 a から始まる Euler 回路の個数 $\varepsilon(K_n^{\mathrm{bidir}}, a)$ は，

$$
\begin{aligned}
\varepsilon(K_n^{\mathrm{bidir}}, a) &= \tau(K_n^{\mathrm{bidir}}, 1) \prod_{u=1}^{n} (k_{\mathrm{out},u} - 1)! \\
&= n^{n-2} \prod_{u=1}^{n} (n-2)! \\
&= n^{n-2} (n-2)!^n
\end{aligned}
\tag{E.5}
$$

と求められる．図 E.3 において，a から始まる Euler 回路がいくつ存在するのかを数えてみると，$n = 2$ のときは $a \to b$ の 1 つだけであり，$n = 3$ のときは $a \to b \to f \to d \to c \to e$, $a \to c \to d \to b \to f \to e$, $a \to c \to e \to f \to d \to b$ の 3 つ存在している．そして，この結果は (E.5) 式に一致することが確かめられる．

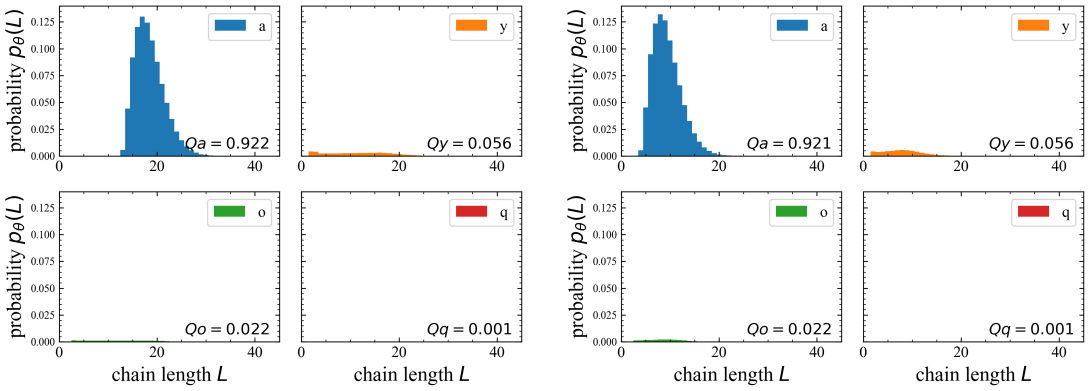
付録 F

自己ループの効果

本付録では，辞書ネットワークの自己ループの有無によって，鎖長分布や終了確率がどのように変化するのかを測定し，自己ループがしりとりに与える影響について考察を行う．

辞書ネットワークの自己ループは，先頭文字と末尾文字が一致する単語（国名辞書における “albania”, “seychelles” など）を表しており，隣接行列では対角成分がその数を表す．勝敗を考えるしりとりでは，局面そのままで手番だけが入れ替わる性質から，自己ループが戦略的に使用されることが多い．それでは，戦略を考えないしりとりでは自己ループがどのような意味を持つのだろうか．

自己ループがある場合とない場合の鎖長分布，終了確率を比較する．図 F.1 の (a) は従来の国名辞書を用いた結果であり，(b) は国名辞書からすべての自己ループを除去した辞書を使用した結果である．自己ループを除くと鎖長分布は大きく変化し，a の分布のピークが左へ 9 移動するなど，鎖長が全体的に小さくなる傾向が見られた．一方で，各文字の終了確率は自己ループを除去する前後でほとんど変化しなかった．以上の結果から，自己ループは鎖長に影響を及ぼす一方で，終了確率には影響を及ぼさないことが示唆される．



(a) 自己ループあり　　(b) 自己ループなし

図 F.1: 自己ループの有無による鎖長分布，終了確率の変化（国名辞書，100 万回実測）

付録 G

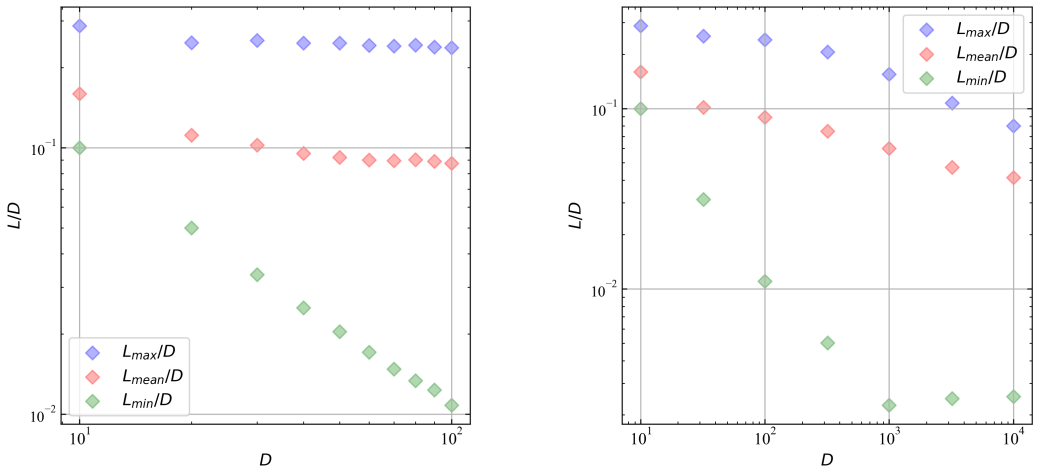
鎖長の単語数依存性

本付録では辞書の単語数が増加したときに，鎖長の統計量がどのように変化するのかを調べる．

国名辞書（193 語）では単語数が少ないため，付録 B の Moby-Dick 辞書（19088 語）を用いて測定する．Moby-Dick 辞書から D 単語をランダムに選択した辞書を各 D で 1000 冊作成した．そして，各辞書で 1000 回しりとりを実行し，カバーレートの最大値，最小値，平均値を求め，D ごとに 1000 冊の平均をとった．その結果が図 G.1 である．

カバーレートはいずれも D^{-1} よりは緩やかに減少している．$L/D \sim D^{-\gamma}$ に従うと考えると，最大値と平均値はいずれも $\gamma \approx 0.19$ 程度で減少する結果となった．最小値は $D = 10^3$ まで $\gamma \approx 0.82$ 程度で減少するが，それ以降はほとんど一定の値をとるようになる．これは辞書ネットワークが繋がり始めたことが関係している可能性もあり，今後詳細に調査したいと考えている．

D が大きい極限で，L/D がどのような関数形に漸近するのかはよくわかっていない．この世に存在する単語は有限であるため，現実的な問題ではないものの，数学的な問題として興味深い課題である．



(a) $D = 10, 20, \cdots, 100$　　(b) $D = 10^1, 10^{1.5}, \cdots, 10^4$

図 G.1: Moby-Dick 辞書の単語数 D をランダムに増やしたときの，鎖長の最大値 $L_{\max}$，最小値 $L_{\min}$，平均値 L_{mean} の変化（各 D で辞書を 1000 冊作成し，各辞書で 1000 回しりとりを実行）

付録 H

言語・品詞ごとの鎖長分布

これまで，英語の名詞でランダムなしりとりを実行したときの鎖長分布を解析してきた．それでは，言語や品詞を変えたときにしりとりの統計的性質はどのように変化するのだろうか．本付録では言語や品詞を変えたときの鎖長分布を数値的に求め，その変化の仕方について考察を行った．

H.1　言語による変化

本節では日本語でしりとりを実行した場合の鎖長分布を計測する．ただし 3.1 節で述べたように，日本語のしりとりは英語に比べていくらか任意性が生じる．そこで以下のルールを追加する．

- 濁点や半濁点は付けたり外したりできる．
- 拗音や促音で終わった場合，次の単語は末尾を大きくした文字から始める．
- 長音で終わった場合，次の単語は長音の前の文字から始める．

辞書として，2026 年 2 月時点で国際連合に加盟している国の日本語名 193 語 [44] を用いた．ただし，単語内の括弧とその中身は削除している．また上記のルールに従って，単語内の濁音や半濁音は清音に変換し，拗音や促音は大きな文字に置き換え，長音は取り除いている．以下，この辞書を「日本語国名辞書」と呼ぶ．図 H.1 は日本語国名辞書の辞書ネットワークである．清音の大きな文字は 46 種類あるが，このうち「ぬ」,「わ」,「を」は国名の先頭文字と末尾文字のどちらにも使われないため，この辞書は $C = 43$ である．図 H.2 は日本語国名辞書の各文字の k_{in}，k_{out} の関係を表している．日本語国名辞書では 19 文字で終了可能となることが判明した．ただし，図 H.2 は両対数プロットであり，k_{in} や k_{out} が 0 となる文字は表示されないことに注意．例えば，「ん」は $(k_{\mathrm{in},\,ん}, k_{\mathrm{out},\,ん}) = (26, 0)$ である[*1]ため表示されない．

[*1] 「ん」で終わる単語を選択したら終了するというルールが自動的に成り立つ．

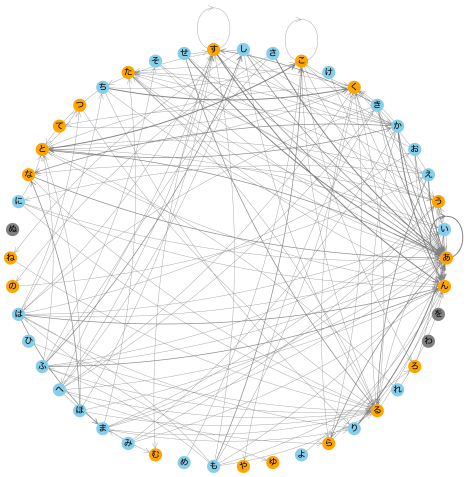


図 H.1: 日本語国名辞書の辞書ネットワーク
辺の太さは有向辺の数に比例している．

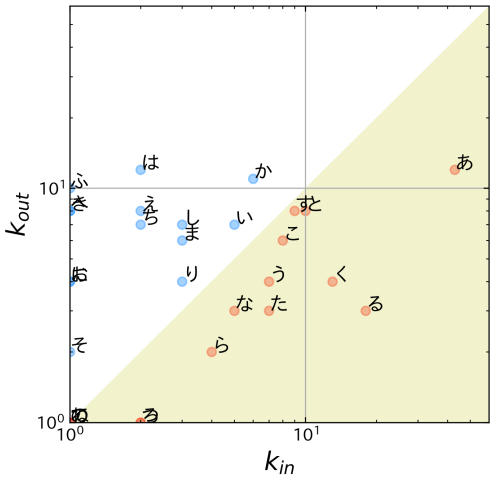


図 H.2: 日本語国名辞書の $k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面
両対数プロットである点に注意．

日本語国名辞書でランダムなしりとりを 100 万回実行したときの鎖長分布は図 H.3 のようになった．鎖長の最小値は 1，最大値は 35，平均値は 5.9 ± 4.6 であり，$L = 1$ でいきなりピークを持つ単調減少な分布となった．英語の場合に比べて鎖長が短く，分布の形状も大きく異なっていることがわかる．また，96.1% が「ん」で終了し，残る 3.9% の単語列は 10 種類の文字で終了する結果となった．これは「ん」で始まる単語が存在しない一方で，「ん」で終わる単語がたくさん存在することが影響していると考えられ，国名辞書に限らず多くの日本語の辞書では「ん」で終了しやすいことが予想される．

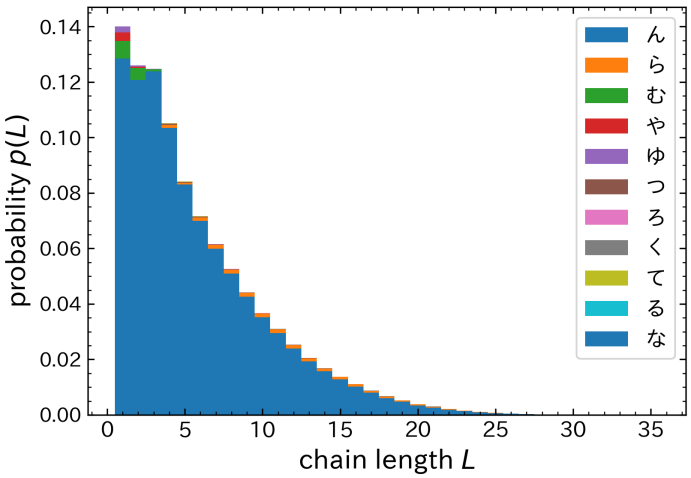
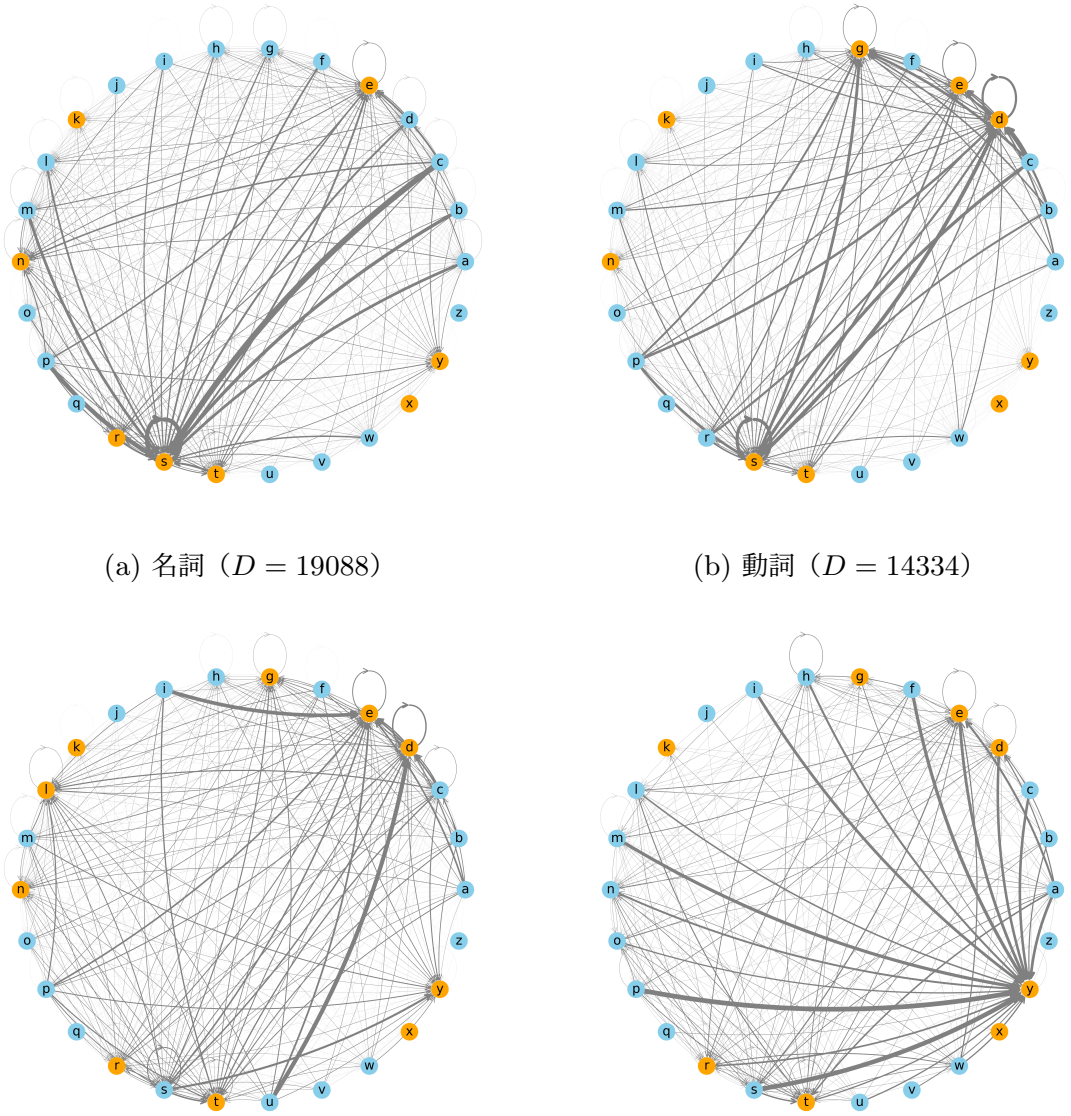


図 H.3: 日本語国名辞書の鎖長分布（100 万回実測）

H.2　品詞による変化

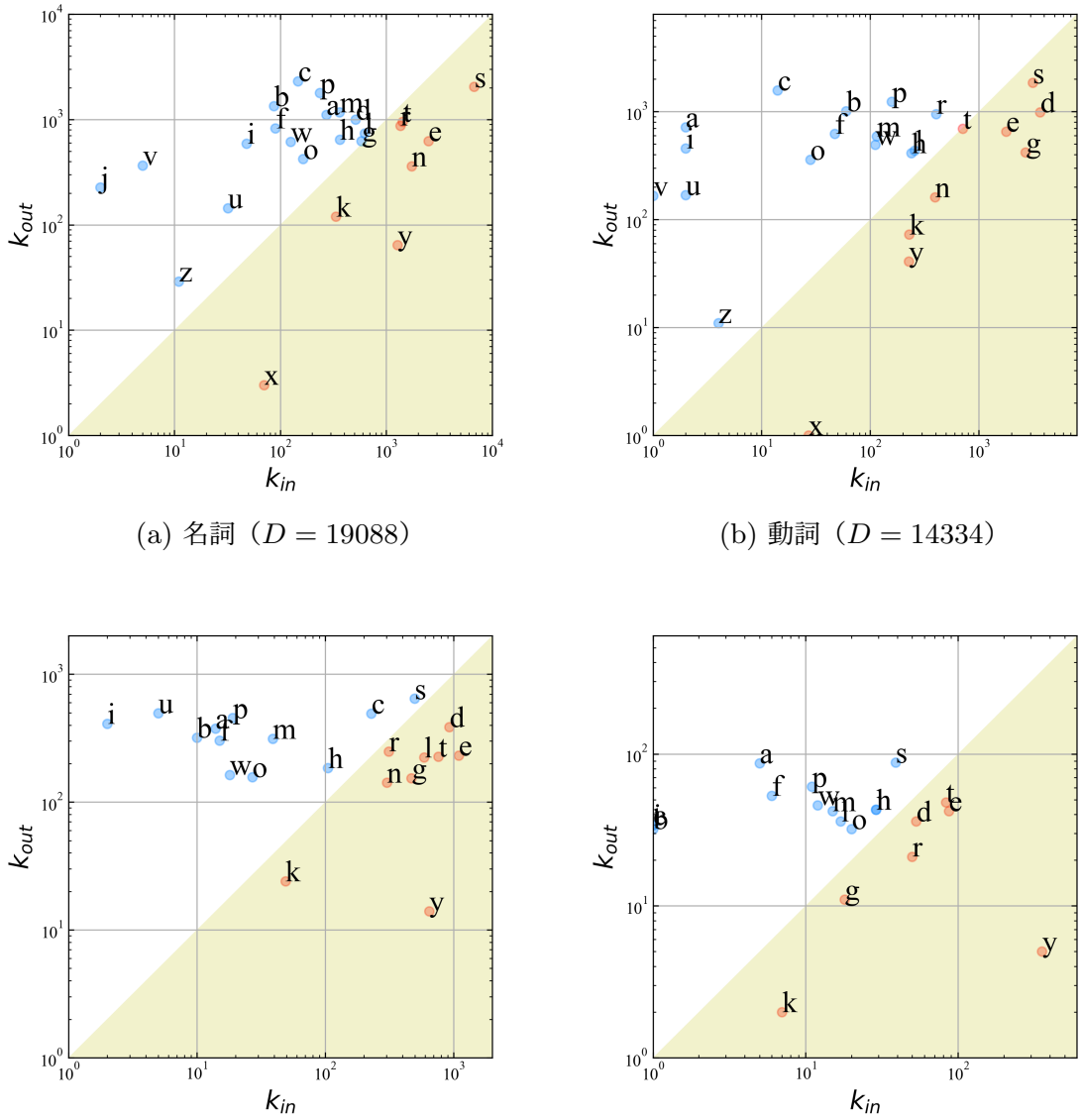
続いて，品詞ごとの鎖長分布の違いを観察する．特に本節では英語の名詞，動詞，形容詞，副詞に着目する．Moby-Dick 辞書は “Moby-Dick” に登場する名詞 19088 語を集めた辞書であるが，同じように動詞 14334 語，形容詞 6137 語，副詞 839 語を集めて，それぞれの辞書でしりとりで実行してみる．しりとりを実行する前に，これらの辞書ネットワークを図 H.4，$k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面を図 H.5 に示した．いずれも孤立点は存在せず，$C = 26$ であった．



(a) 名詞（$D = 19088$）　(b) 動詞（$D = 14334$）

(c) 形容詞（$D = 6137$）　(d) 副詞（$D = 839$）

図 H.4: “Moby-Dick” に登場する単語の品詞ごとの辞書ネットワーク
辺の太さは有向辺の数に比例しており，その比例定数は品詞ごとに異なる点に注意．

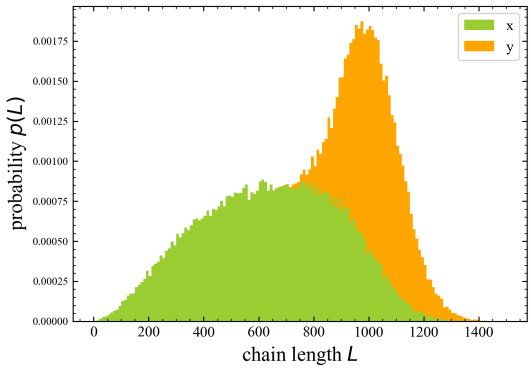


(a) 名詞（$D = 19088$）　(b) 動詞（$D = 14334$）

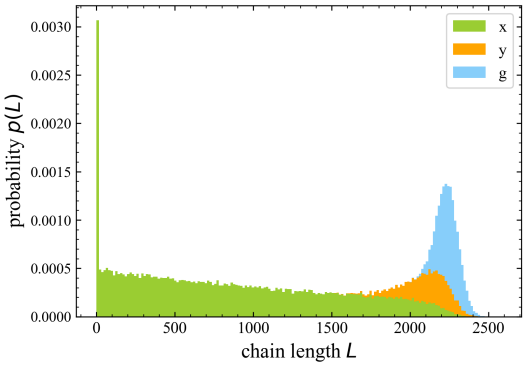
(c) 形容詞（$D = 6137$）　(d) 副詞（$D = 839$）

図 H.5: “Moby-Dick” に登場する単語の品詞ごとの $k_{\text{in}} - k_{\text{out}}$ 平面
両対数プロットである点に注意.

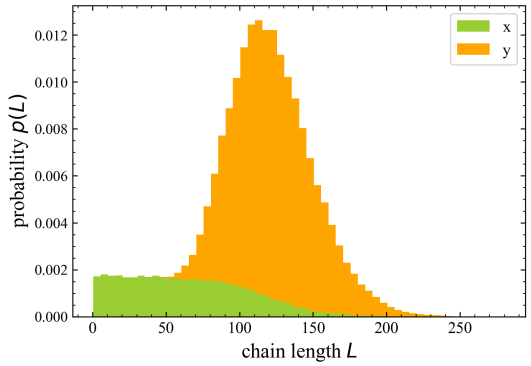
各辞書で 10 万回しりとりを実行したときの鎖長分布を図 H.6 に示した．分布の形状は品詞ごとに様々であり，分布の多様性が確認される．終了文字の種類に関してはいずれも似通っており，x, y がすべての品詞に現れる結果となった．しかし，その終了確率は大きく異なっており，名詞と動詞は過半数が x で終了する一方で，形容詞は過半数が y で終了し，副詞に至ってはほとんどが y で終了する結果となった．副詞が y で終了しやすい理由としては，actually や recently など y で終了する単語が突出して多いからと考えられる．このように，言語や品詞の特徴は鎖長分布や終了確率に反映される．言語的特徴としりとりの統計的性質の詳細な関係は，今後調査したいと考えている．



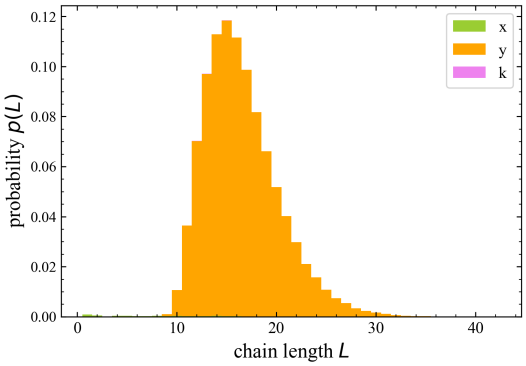
(a) 名詞（$D = 19088$）



(b) 動詞（$D = 14334$）



(c) 形容詞（$D = 6137$）



(d) 副詞（$D = 839$）

図 H.6: “Moby-Dick” に登場する単語の品詞ごとの鎖長分布（10 万回実測）

参考文献

[1] N. Inui, et al., “Solving the Longest Word-Chain Problem”, *Proceedings of the First International Conference on Informatics in Control, Automation and Robotics*, (2004) pp.214–221.

[2] M. Murata and T. Shirado, “Statistical Investigation of a Japanese Word Chain Game”, *International Information Institute (Tokyo). Information*, **18** (2015) pp.1631–1640.

[3] I. Tishby, O. Biham, and E. Katzav, “The distribution of path lengths of self avoiding walks on Erdős-Rényi networks”, *Journal of Physics A: Mathematical and Theoretical*, **49** (2016) 285002.

[4] C. P. Herrero, “Self-avoiding walks on scale-free networks”, *Physical Review E*, **71** (2005) 016103.

[5] 藤田悠朔, 鈴木岳人, 水口毅, 「ランダムなしりとりの平均場解析」, 第 31 回交通流と自己駆動粒子系のシンポジウム 論文集, (2025) pp.33–36.

[6] 伊藤隆ほか, 「しりとりゲームの数理的解析」, 情報処理学会論文誌, **43** (2002) pp.3012–3020.

[7] 乾伸雄, 品野勇治, 小谷善行, 「単語長を考慮した最長しりとり問題の実験的考察」, 情報処理学会論文誌, **46** (2005) pp.131–142.

[8] 田中駿ほか, 「しりとりを通してみた幼児の言語発達」, LD 研究, **33** (2024) pp.187–195.

[9] 高橋登, 「幼児のことば遊びの発達：“しりとり” を可能にする条件の分析」, 発達心理学研究, **8** (1997) pp.42–52.

[10] M. F. F. Abbas, “Applying Word Chain Game To Improve Students’ Vocabulary Mastery”, *ELT-Lectura*, **1** (2014) pp.44–48.

[11] W. Ramadani, W. Naro and N. A. Nur, “The Influence of Word Chain Game on Increase the Eighth Grade Students’ Vocabulary at Mts Barana Jeneponto”, *English Language Teaching for EFL Learners Journal*, **2** (2020) pp.1–11.

[12] L. Euler, “Solutio problematis ad geometriam situs pertinentis”, *Commentarii academiae scientiarum Petropolitanae*, **8** (1741) pp.128–140. `https://scholarlycommons.pacific.edu/euler-works/53`.

[13] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks”, *Nature*, **393** (1998) pp.440–442.

[14] A. L. Barabási and R. Albert, “Emergence of Scaling in Random Networks”, *Science*, **286** (1999) pp.509–512.

[15] M. E. J. Newman, “The Structure and Function of Complex Networks”, *SIAM review*, **45** (2003) pp.167–256.

[16] R. F. I. Cancho and R. V. Solé, “The Small World of Human Language”, *Proceedings: Biological Sciences*, **268** (2001) pp.2261–2265.

[17] A. E. Motter, et al., “Topology of the conceptual network of language”, *Physical Review E*, **65** (2002) 065102.

[18] P. C. Hemmer and S. Hemmer, “Trapping of genuine self-avoiding walks”, *Physical Review A*, **34** (1986) 3304.

[19] N. Madras and G. Slade, “The Self-Avoiding Walk”, (Birkhäuser, Boston) (1996).

[20] A. R. Conway and A. J. Guttmann, “Enumeration of self avoiding trails on a square lattice using a transfer matrix technique”, *Journal of Physics A: Mathematical and General*, **26** (1993) 1535.

[21] A. Malakis, “Self-avoiding walks on oriented square lattices”, *Journal of Physics A: Mathematical and General*, **8** (1975) 1885.

[22] E. W. Weisstein, “Line Graph”, From MathWorld–A Wolfram Resource. `https://mathworld.wolfram.com/LineGraph.html`. （2026 年 1 月 17 日参照.）

[23] P. G. de Gennes, “Scaling Concepts in Polymer Physics”, (Cornell university press, Ithaca) (1979).

[24] S.-J. Yang, “Exploring complex networks by walking on them”, *Physical Review E*, **71** (2005) 016107.

[25] L. Page, et al., “The PageRank Citation Ranking: Bringing Order to the Web”, *Technical Report. Stanford InfoLab.*, (1999) pp.1–17.

[26] L. da F. Costa and G. Travieso, “Exploring complex networks through random walks”, *Physical Review E*, **75** (2007) 016102.

[27] S. Hemmer and P. C. Hemmer, “An average self-avoiding random walk on the square lattice lasts 71 steps”, *The Journal of Chemical Physics*, **81** (1984) pp.584–585.

[28] I. Majid and N. Jan, “Majid et al. Respond”, *Physical Review Letters*, **55** (1985) 2092.

[29] C. P. Herrero, “Kinetic-growth self-avoiding walks on small-world networks”, *The European Physical Journal B*, **56** (2007) pp.71–79.

[30] C. P. Herrero, “Kinetic growth walks on complex networks”, *Journal of Physics A: Mathematical and General*, **38** (2005) 4349.

[31] United Nations. https://www.un.org/en/about-us/member-states. （2025 年 5 月 5 日参照.）

[32] Project Gutenberg. https://www.gutenberg.org/ebooks/2701.

[33] K. G. Janardan and G. P. Patil, “A Unified Approach for a Class of Multivariate Hypergeometric Models”, *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **34** (1972) pp.363–376.

[34] M. J. Kronenburg, “The Binomial Coefficient for Negative Arguments”, *arXiv:1105.3689*, (2011) pp.1–6.

[35] J. A. Keats and F. M. Lord, “A theoretical distribution for mental test scores”, *Psychometrika*, **27** (1962) pp.59–72.

[36] M. S. Ridout, “Memory in Coal Tits: An Alternative Model”, *Biometrics*, **55** (1999) pp.660–662.

[37] S. N. Jones, “A Gaming Application of the Negative Hypergeometric Distribution”, *UNLV Theses, Dissertations, Professional Papers, and Capstones.*, (2013) 1846.

[38] R. A. Khan, “A Note on the Generating Function of a Negative Hypergeometric Distribution”, *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, **56** (1994) pp.309–313.

[39] W. T. Tutte and C. A. Smith, “On Unicursal Paths in a Network of Degree 4”, *The American Mathematical Monthly*, **48** (1941) pp.233–237.

[40] T. van Aardenne-Ehrenfest and N. G. de Bruijn, “Circuits and trees in oriented linear graphs”, *Simon Stevin : Wis- en Natuurkundig Tijdschrift*, **28** (1951) pp.203–217.

[41] Wikipedia. https://en.wikipedia.org/wiki/Negative_hypergeometric_distribution. （2025 年 9 月 26 日参照.）

[42] A. L. Barabási, 池田裕一・井上寛康・谷澤俊弘（監訳）, 京都大学ネットワーク社会研究会（訳）,「ネットワーク科学 –ひと・もの・ことの関係性をデータから解き明かす新しいアプローチ–」, 共立出版 (2019).

[43] D. Grinberg, “An introduction to graph theory”, *arXiv:2308.04512*, (2023) pp.1–422. （inzkyk（訳）, 2024. https://inzkyk.xyz/graph/.）

[44] 国際連合広報センター. https://www.unic.or.jp/info/un/un_organization/member_nations/. （2026 年 2 月 9 日参照.）