

# 2ch.net の書き込みデータの統計的法則と Pitman 分布

Pólya urn and Pitman sampling formula in 2ch.net posting data

守真太郎 \*1  
Shintaro Mori

久門正人 \*2  
Masato Hisakado

\*1 北里大学理学部物理学科  
Department of Physics, Kitasato University

\*2 Fintech Lab.  
Fintech Lab.

2ch.net の投稿時系列データを解析する。過去  $r$  回の投稿に出現したスレッドの分布と次に投稿されたスレッドの関係  
を有限記憶のポリア壺過程としてモデル化する。このモデルでは過去  $r$  回の投稿に出現したスレッドの定常確率分布は  
Pitman 分布に従う。2ch.net のニュース系掲示板のデータで検証し、 $r$  が十分小さいとき、投稿がポリア壺過程で記述  
されることが分かった。

## 1. 複雑系経済学から経済物理学へ

1990 年前後、経済学が複雑系研究の 1 分野として盛んに研究されました。経済学が複雑系の研究対象となったのは、時間  
変化する環境での適応的な多体系というアイデアが、生物・生  
命現象だけでなく市場および市場参加者にも適用できたからで  
す。複雑系研究の中心はアメリカのサンタフェであり、複雑系  
経済学をリードしたのが Arthur でした。Arthur は、経済学で  
は負のフィードバックを仮定し均衡状態が 1 つとしてきたが、  
正のフィードバックが働く財も存在し、均衡状態が複数出現  
することがあることを示しました [Arthur 1989]。2 つの商品  
A, B があったとき、A, B の品質は同じでも、多数の消費者に  
選ばれることで優位となる外部性が働き、1 つの商品が市場を  
独占することがあります。Arthur が例として挙げたのが、ガ  
ソリン車と石炭車、タイプライターのキーボードの配列、そし  
て日本人には馴染みの深い、家庭用録画機の VHS とベータマッ  
クスのビデオデッキの規格です。こうした市場の選択による外  
部性をモデル化するとき Arthur が用いた数学がポリア壺で  
す。ポリア壺のモデルでは、A, B の 2 商品の場合、市場にお  
ける A 商品の比率を  $z$  として、次の消費者が A を選ぶ確率を  
 $f(z)$  という関数で表します。すると、消費者の選択とともに  
 $z$  は時間変化し、 $f(z)$  の安定な固定点、つまり  $f(z)$  のグラフ  
が対角線を上から下に横切る (downward crossing)  $f(z) = z$   
の解  $z_*$  に収束します。安定な固定点  $z_*$  が複数存在する場合、  
どの安定な固定点にも収束する確率は正です。Arthur は外部  
性によりフィードバックが働くとき  $f(z)$  はシグモイド関数の  
ように S 字カーブを描き、安定な固定点が 2 個存在するこ  
とを示唆したわけです。どちらの固定点に収束するかはランダ  
ムであり、経済学が想定する唯一の均衡状態のアイデアと対立  
することになります。この複雑系としての経済学研究の流れは  
進化経済学という名前で今でも活発な研究が行われています。

一方、正のフィードバックによる複数の安定固定点の可能性  
は経済物理学の問題として検証されました。経済物理学では、  
株価や為替などの高頻度の時系列データに見られる統計的な  
性質をエージェント間の相互作用をベースに理解することが 1  
つの目標となっています。我々はエージェント間の相互作用を  
実験室実験で計測し、系のミクロ・マクロを理解すること、また、  
単に統計的な性質を再現するモデルを提案するのではなく、統計物理学の文脈として新しい数理を見つけることが経済  
物理学が物理学であること必要条件であると考えました。S 字

カーブの検証は情報カスケード実験により行われました。情報  
カスケードとは、二択クイズなどの離散的な選択において多数  
のヒトが順番に不確かな情報と、過去に回答したヒトの選択情  
報をもとに選択したとき、自分が正しいと考える選択肢ではな  
く、多数派の選択肢を選ぶ傾向のことです。ある選択肢の過去  
の選択での比率  $z$  の関数  $f(z)$  としてヒトがその選択肢を選ぶ  
確率を記述すると、情報カスケードでの選択の時系列はポリア  
壺となります。実験では、二択クイズの難易度の変化により、  
安定固定点の個数が 1 個から 2 個へと変化することが観測さ  
れました [Mori 2012]。また、ポリア壺の安定固定点の個数の  
変化は統計物理学での非平衡相転移であり、平衡系の相転移と  
は異なる普遍類に属するだけではなく、非平衡相転移としても新  
しいことを示しました [Mori 2015]。

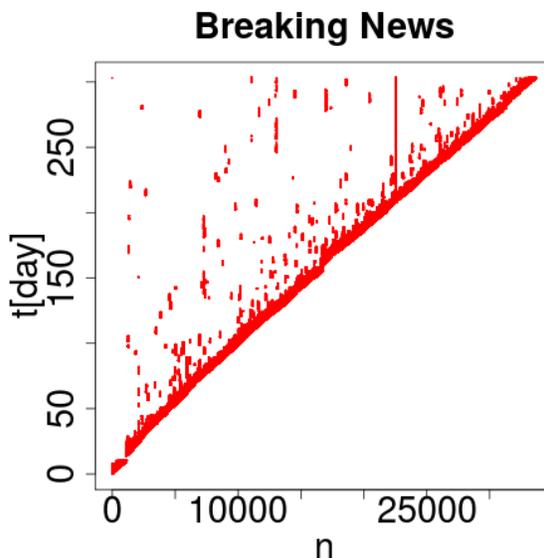


図 1: ニュース速報 (Breaking News) の 10ヶ月約 2000 万回  
の投稿の散布図。x 軸は  $s$  回目の投稿でのスレッド番号  $n(s)$ 、  
y 軸は投稿時刻  $t(s)$  (日)。この時系列データを y 軸に垂直に  
薄切りにすると Pitman 分布が現れます。図 3 参照してくだ  
さい。

こうして、経済現象、社会現象といったヒトの集団挙動の研

究は経済学、心理学から始まり物理学の対象にもなってきました。近年の情報技術、特に Wwww の発展によりヒトの詳細な行動データが集積され、ビッグデータと呼ばれる膨大なデータが研究対象に加わり、経済物理学、社会物理学に加えて情報科学で盛んに研究が行われるようになってきました。本講演では「2ちゃんねる」と呼ばれる電子掲示板 2ch.net の投稿過程のモデル化の研究について報告します。この研究はヒトの選択に影響するファクターを明らかにすることで、ヒトの相互作用、他人の選択のヒトの選択への社会的影響の法則性について知る経済物理学的な興味と、有限記憶ポリア壺のパラメータの推定法、モデルの妥当性の検証方法に関する数理統計的な興味との2つの興味に関するものです。

## 2. ポリア壺と Pitman 分布

生態学データの基本データは観測数  $r$ 、観測した種の数  $K$ 、捕獲した種  $i$  の個体数  $c_i, i \in \{1, 2, \dots, K\}$  の3つの整数です。  $r = \sum_{i=1}^K c_i$ 。この  $\{c_1, c_2, \dots, c_K\}$  を  $r$  の確率分割とみなし、Pitman 分布、または Pitman 確率分割に従うとすると、 $\theta, \alpha$  の2つのパラメータを用いて  $K$  および  $\vec{c} = (c_1, c_2, \dots, c_K), c_i > 0$  の確率分布は

$$P_r(\vec{c}) = \frac{n! \theta^{[K:\alpha]} }{\theta^{[r]}} \prod_{k=1}^K (1 - \alpha)^{[c_k - 1]} \quad (1)$$

となります。ここで、 $\alpha, \theta$  は  $\alpha \geq 0$  のときは、 $\theta > -\alpha, \alpha < 0$  のときは、ある正の整数  $M > 0$  を用いて  $\theta = -M\alpha$  とします。また、 $x^{[j]} = x(x+1)\dots(x+j-1), x^{[j:\alpha]} = x(x+\alpha)\dots(x+(j-1)\alpha), x^{[j:1]} = x^{[j]}$  です。 $\vec{c}$  の順序を無視して、寸法指標と呼ばれる同じ大きさの  $c_k$  の数  $s_j = \sum_i \delta_{c_i, j}$  に関する確率分布  $P(\vec{s}), \vec{s} = (s_1, s_2, \dots, s_r)$  としたものを Pitman 分布、Pitman 確率分割 (Ewens-Pitman sampling formula)  $\text{EPSF}(\theta, \alpha)$  と呼びますが、ここでは式 (1) の  $P_r(\vec{c})$  を Pitman 分布  $\text{ESPF}(\vec{c}|\theta, \alpha)$  と呼ぶことにします。また、種の数  $K$  の確率分布  $P_r(K)$  は、次の漸化式を解くことにより求めることができます。

$$P_{r+1}(K) = \frac{r - K\alpha}{\theta + r} \cdot P_r(K) + \frac{\theta + (K-1)\alpha}{\theta + r} \cdot P_r(K-1) \quad (2)$$

初期条件は  $P_1(1) = 1$  であり、境界条件は  $P_r(0) = 0, P_r(r+1) = 0$  です。この漸化式の解を  $\text{ESPF}_K(K|\theta, \alpha)$  と書きます。

Pitman 分布のパラメータ  $\theta, \alpha$  の意味を理解するために、ポリア壺過程として Pitman 分布をモデル化します [Sibuya 2001, Yamato 2003]。まず、番号のついた無限個の壺  $U_1, U_2, \dots$  と番号のついた玉  $B_1, B_2, \dots$  を用意し、番号の順番に壺にいれていくとします。まず、玉  $B_1$  は壺  $U_1$  にいれます。 $r = 1$  で  $c_1 = 1$  であり、確率分布  $P(\vec{c}) = P(c_1)$  は  $P_1(1) = 1$  となります。次に  $r$  個の玉  $B_1, \dots, B_r$  を  $K$  個の壺にいれた状態  $\vec{c} = (c_1, \dots, c_K), \sum_j c_j = r, c_j > 0$  で、次の  $r+1$  の玉  $B_{r+1}$  を新しい壺  $B_{K+1}$  に確率  $(\theta + K\alpha)/(\theta + r)$  でいれるか、

$$\text{Pr}(K \rightarrow K+1) = \frac{\theta + K\alpha}{\theta + r},$$

すでに玉の入っている壺  $U_j, j = 1, \dots, K$  の何れかに確率  $(c_j - \alpha)/(\theta + r)$  で入れます。

$$\text{Pr}(c_j \rightarrow c_j + 1) = \frac{c_j - \alpha}{\theta + r}.$$

$\alpha < 0$  のときは、壺  $U_{M+1}, U_{M+2}, \dots$  には玉は入りません。このとき、 $r$  個玉をいれたときの  $\vec{c}$  の確率分布は、Pitman 分布  $P_r(\vec{c})$  に、また、玉の入っている壺の数  $K$  の確率分布は  $P_r(K)$  となります。この確率法則から、ボールを新しい壺へ入れる確率は  $\alpha > 0$  のとき  $K$  の増加関数です。また  $\alpha = 0$  のとき、新しい壺への玉の追加は確率  $\theta/(\theta+r)$  でおき、 $K-1$  の分布は二項分布  $\text{Bi}(r-1, \theta/(\theta+r))$  となります。 $\theta$  が大きいときは新しい壺にボールを入れる確率は大きくなります。玉の入っている壺  $U_j, j \leq K$  への玉を追加確率は  $c_j$  に比例し、 $(\theta+r), \alpha$  の減少関数となっています。 $\alpha = 0$  のとき、追加確率は  $c_j/(\theta+r)$  なので、 $\theta$  が大きくなると減少します。

上記のポリア壺では、1 個の壺から出発し、 $r$  個の玉を加えたときの壺  $U_j$  に入る玉の個数  $c_j$  が Pitman 分布に従うというものです。つまり、非定常分布として Pitman 分布を導いています。ポリア壺の確率過程を次のように変更することで、定常分布として Pitman 分布を導くことができます [Hisakado 2017]。まず、 $r$  個の番号のついた玉  $B_1, \dots, B_r$  を用意し、 $r$  を  $K$  個に分割して  $\vec{c} = (c_1, c_2, \dots, c_K), \sum_{j=1}^K c_j = r$  に合わせて  $K$  個の壺に玉をいれて初期状態を用意します。そして、次の玉  $B_{r+1}$  を上記のポリア壺過程と同じ確率で壺に入れます。そして、一番番号が小さい玉  $B_1$  を取り除き、壺に入っている玉の総数を  $r$  に保ちます。同様に、次の玉  $B_{r+2}$  もポリア壺過程に従って壺に追加し、玉  $B_2$  を取り除きます。この操作を繰り返したとき、 $U_j$  に入る玉の個数  $c_j > 0$  の定常確率分布は Pitman 分布  $P_r(\vec{c})$  に従い、また、玉の入っている壺の個数  $K$  は確率分布  $P_r(K)$  に従うことを示すことができます。このポリア壺では、 $B_t, t \geq r+1$  を追加するとき、 $B_{t-r}, \dots, B_{t-1}$  の  $r$  個の玉の情報のみを参照していることになるので、有限記憶のポリア壺過程と呼ぶことにします。このポリア壺過程では、壺の実力には差はなく、入っている玉の個数で次の玉が入る確率が決まります。以下、2ch.net の解析では玉を投稿、壺をスレッドと読み替えますが、ポリア壺過程ではスレッドの実力 (投稿を促す力) を無視する単純化を行っていることに注意してください。

## 3. 2ch.net の投稿時系列データの解析

2ch.net は日本最大の電子掲示板であり、非常に幅の広い話題を扱っています。各々の掲示板はニュース、食、文化などのカテゴリーとよばれる大きな分類で区別され、さらにジャンルにあたる板により細分化されます。板は数百のスレッドと呼ばれる電子掲示板を持ち、ユーザーはスレッドの投稿を読んだり投稿します。ここでは、ニュースジャンルの5つの板、ビジネスニュース、東アジアニュース、ライブニュース、音楽ニュース速報+、ニュース速報+、の投稿の時系列データを解析します。各々の板は数百のスレッドを持ち、管理者は古いスレッドを削除しながら新しいスレッドを追加して板のスレッドの数をほぼ一定に保ちます。ニュースジャンルの板ではユーザーがスレッドに投稿できるのは数日間です。1つのスレッドの最大投稿数は1000、または容量が500KBなので、それを超えた場合、管理者が続き番号で同じスレッドタイトルのスレッドを追加します。この場合、最大投稿数1000、投稿可能期間の制限はなくなるようになります。表1は、解析で扱った板の基本的な統計情報です。2009年の約半年から10ヶ月の間、毎日1度2ch.net サーバーにアクセスし、全スレッドを読み込んで10桁のスレッド番号 (スレッドID) 投稿日時、ユーザーIDを記録したものを使っています。1000回投稿のあった続きのスレッドの場合、最も古い同じスレッドタイトルのスレッドの

レッド ID も記録しますが、各スレッドの最初の投稿でそのスレッドが続きのスレッドかどうかを判定するため、続きスレッドであるにも関わらず新規スレッドと記録している可能性は排除できません。

表 1 の  $N$  は、続きのスレッドを同一視したときの各板に現れたスレッド数です。 $S$  は投稿総数ですが、表では 1 つのスレッドあたりの平均の投稿数  $S/N$  と、スレッドの投稿数の最大値を示しています。スレッド数は数千から数万、投稿数は 80 万から 2000 万程度の巨大なデータであることが分かります。また、寿命はスレッド毎の最初の投稿と最後の投稿の時間差の平均値を示しています。

表 1: 2ch.net の投稿データの統計。全スレッド数  $N$ , スレッドあたりの投稿数  $S/N$  とその標準偏差, 最大投稿数, スレッドの寿命 (日)。

板名	$N$	$S/N$	最大投稿数	寿命
ビジネスニュース	8248	140	7707	7.5
東アジアニュース	8225	388	27966	7.5
ライブニュース	15307	53	30443	2.0
音楽ニュース	23000	332	78388	2.8
ニュース速報	33677	658	113220	2.9

最大投稿数と  $S/N$  の差が大きいことから予想されるように、各スレッドの投稿数の分布は裾の厚い分布になりますが、べき則には従わず対数正規的な振る舞いを示します。板ごとにスレッドを  $n \in \{1, \dots, N\}$  でラベルし、 $s$  回目の投稿が  $t[s]$  にスレッド  $n$  に行われたとき、 $(t(s), n(s))$  と表します。Figure 1 はニュース速報の  $(t(s), n(s)), s = 1, \dots, S$  の散佈図です。スレッドは有限の寿命を持っているので、帯状のパターンを確認できます。非常に長い寿命のスレッドも存在しますが、それは続きのスレッドを同一視したためです。

### 3.1 確率法則の検証とパラメータ推定

2ch.net の電子掲示板の投稿過程を有限記憶のポリア壺過程を用いて記述します。掲示板には数百のスレッドが存在し、ユーザーはスレッドを選択して投稿を行います。 $s$  回目の投稿において、 $s-r, s-r+1, \dots, s-1$  回の過去  $r$  回の投稿を調べ、投稿が行われたスレッド数を  $K$ 、またスレッドに古い投稿が行われた順にスレッドに番号  $1, 2, \dots, K$  をつけます。スレッド  $j$  の投稿数を  $c_j, j = 1, \dots, K$  とすると、 $\sum_j c_j = r$  が成立します。 $s$  回目の投稿が有限記憶のポリア壺過程となっており、スレッド  $j$  に行われる確率は  $(c_j - \alpha)/(\theta + r)$ 、過去  $r$  回の投稿には現れていないスレッドへの投稿確率は  $(\theta + K\alpha)/(\theta + r)$  です。 $s = 10^4 \dots, 101 \times 10^4$  の 100 万回の投稿データ  $(n(s), t(s))$  を用いてパラメータ  $\theta, \alpha$  を最尤法で評価しました。

確率法則をデータから直接計算したものと最尤法で有限記憶ポリア壺の確率法則でパラメータをデータから推定したものを比較します。図 2 が結果です。 $r \in \{10, 30, 70\}$  の 3 つのケースでのニュース速報の確率法則をプロットしています。新規スレッドへの書き込み確率の  $K$  依存性は  $r$  が大きくなると傾きが小さくなり、 $\alpha$  が減少することが分かります。 $r$  を数百から 1000 程度にすると  $\alpha$  の推定値はほぼゼロになり、確率法則の  $K$  依存性も消滅することを確認しました。十分大きな  $r$  では、新規スレッドへの投稿はポアソンのように行われることを示唆します。また、既存スレッドへの投稿の確率の  $c$  依存性も  $r$  と

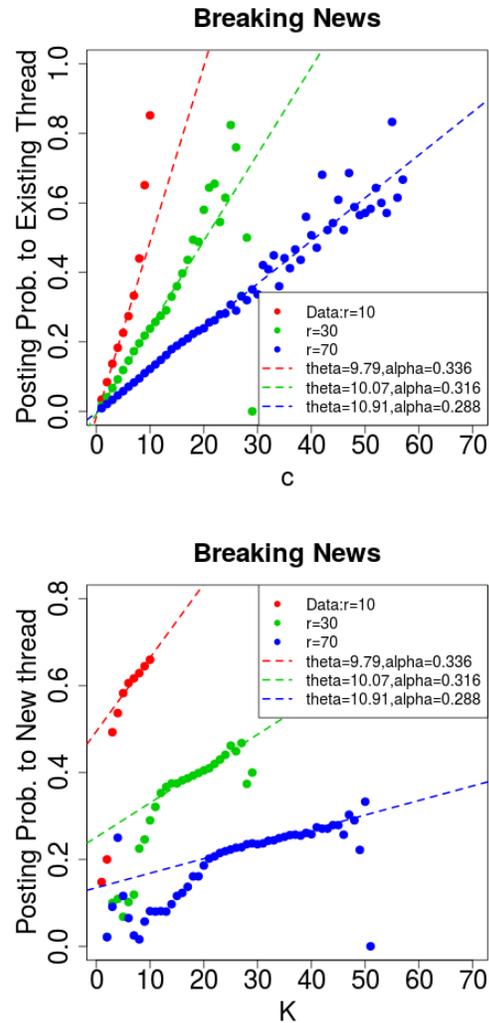


図 2: ニュース速報の確率法則：上図：投稿数  $c$  の既存スレッドへの投稿確率の  $c$  依存性を示したもの。下図：スレッド数  $K$  のときの新規スレッドへの投稿確率の  $K$  依存性を示したものの  $r \in \{10, 30, 70\}$ 。点線は理論モデルのフィッティングの結果、丸はデータの解析結果。

もに傾きが減少していますが、これは  $r$  の増加により  $c$  のレンジが広がったためです。実際、 $c/r$  に対してプロットすると  $r$  依存性は消滅し、対角線に乗ります。これらの結果から、 $r$  が小さいとき、つまり投稿仮定を薄切りにして断面を見た時に有限記憶のポリア壺の確率法則が確認できました。図 2 では、ニュース速報の  $r \in \{10, 30, 70\}$  の結果しか示していませんが、他の板でも同様の振る舞いが確認できています。このことはニュース系の板では一般的に成立することを示唆します。

### 3.2 $P_r(K)$ と $P_r(\vec{c})$ の検証

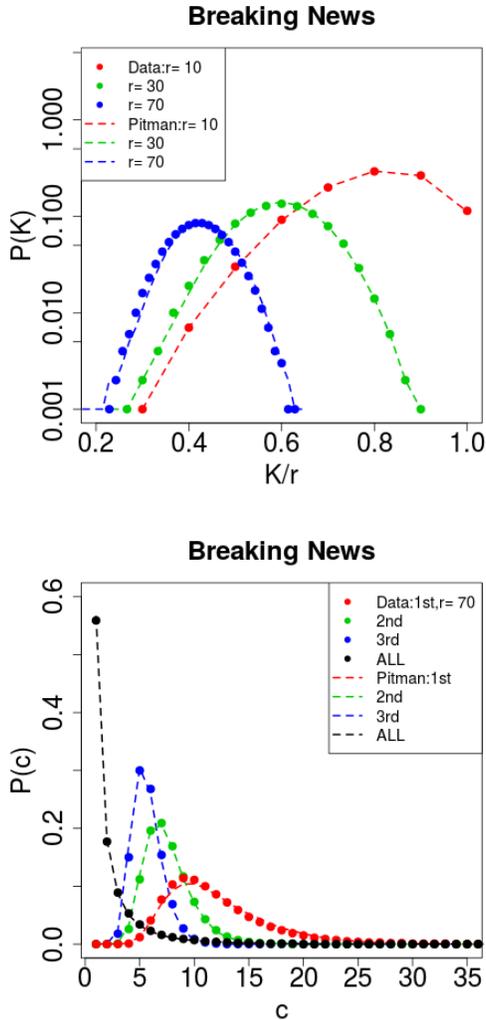


図 3: ニュース速報の  $K$  と  $\vec{c}$  の分布関数をプロット。上図:  $P_r(K)$  をプロット。下図:  $P_r(\vec{c})$  から、 $\vec{c}$  を降順で並べたときの  $c_1 \geq c_2 \geq c_3$  の分布と  $c_j$  の分布をプロット。点線は理論モデルの計算結果、丸はデータの解析結果。

次はスレッド数  $K$  の確率分布  $P_r(K)$  と投稿数  $\vec{c}$  の確率分布  $P_r(\vec{c})$  を検証します。図 3 はニュース速報の投稿データから計算した  $K$  の分布と  $\vec{c}$  を降順に並べたときの、 $c_1 \geq c_2 \geq c_3$  の分布と  $c_j, j = 1, \dots, K$  の分布と  $P_r(K)$  および  $P_r(\vec{c})$  をプロットしたものです。上図は  $r \in \{10, 30, 70\}$  での  $K$  の分布、下図は  $r = 70$  での  $c_1, c_2, c_3$  および  $c$  の分布をプロットしています。この結果から、 $r \leq 70$  程度で薄切りした断面は Pitman 分布に従うことが分かります。

## 4. 結論

本稿では、2ch.net の投稿過程を有限記憶のポリア壺としてモデル化し、投稿データで検証しました。投稿の時系列データを薄切りして調べたとき、既存スレッドへの投稿確率は投稿数に比例し、ポジティブフィードバックが働いていることが分かりますが、線形なので情報カスケードで見られる S 字カーブのような均衡点が複数存在するものではありません。また、投稿数の分布、スレッド数の分布も Pitman 分布に従うことが分かり、2ch.net のニュース系掲示板のマイクロ・マクロの部分的な記述には成功していると考えます。

これからの課題のひとつが 2ch.net の記憶  $r$  の推定方法です。 $r$  が大きくなると  $\alpha$  がゼロに近づき、一方  $\theta$  は増加します。これはポアソンの的に確率  $\theta/(\theta+r)$  で新規スレッドへの投稿が行われることを示唆しますが、一方で  $P_r(K), P_r(\vec{c})$  はデータを用いて計算した  $K, \vec{c}$  の分布と一致しくなりません。記憶長  $r$  には限界があるか、または、ポリア壺過程で無視されているスレッドの実力が無視できなくなることを示唆します。有限記憶ポリア壺の  $r, \theta, \alpha$  の推定方法、また、 $\theta, \alpha$  の  $r$  依存性、壺の実力(適合度)など、確率モデルの数理として興味深い問題と、2ch.net における投稿過程というヒトの集団挙動の解明が渾然一体となって、本研究が実り多いものになることを期待しています。

## 参考文献

[Arthur 1989] W.B.Arthur, "Competing Technologies, Increasing Returns, and Lock-in by Historical Events." *Econ. J.*99,(1989),116-31.  
 [Mori 2012] S.Mori, M. Hisakado and T. Takahashi, "Phase transition to two-peaks phase in an information cascade voting experiment", *Phys.Rev.E*86(2012),026109-026118.  
 [Mori 2015] S. Mori and M. Hisakado, "Correlation function for generalized Polya urns: Finite-size scaling analysis", *Phys.Rev.E*92(2015),052112-052121.  
 [Sibuya 2001] M.Sibuya and H.Yamato, "Pitman's model of random partitions", *数理解析研究所講究録* 1240(2001),6473.  
 [Yamato 2003] 大和元、渋谷政昭,"ピットマン確率分割と関連する話題", *統計数理* (2003) 第 51 巻第 2 号 351-372.  
 [Hisakado 2017] M. Hisakado and S.Mori,"Pitman sampling formula in Equilibrium and Non-equilibrium processes", preprint.