Calculation of Information Cascade Final Size Distributions

Kazumasa Oida[1]

Fukuoka Institute of Technology, Fukuoka, 811-0295 Japan
oida@fit.ac.jp,
WWW home page: https://www.fit.ac.jp/˜oida/index.html

1 Introduction

A large-scale information cascade is a phenomenon in which attractive content (viral content) spreads to a large number of online social network (OSN) users. The problem of predicting the final size of a cascade when its size is small has been discussed for more than 20 years, where the size is the total number of users who have shared information. If this problem can be solved, then (1) the content delivery system (CDS) can be made more efficient by moving viral contents to servers closer to the viewers, (2) useful information (stock trading, product development, etc.) can be retrieved earlier from viral contents, and (3) the forecasting technologies can then be applied to viral marketing, which intentionally creates large-scale cascades.

Large cascades rarely occur, so available datasets for prediction are currently limiting. Thus, it is not easy to apply machine learning techniques or to verify the statistical validity for proposed methods. On the other hand, simulation-based forecasting research is progressing. Bipolarization in the final size distribution of large information cascades has been reported in [2], which emerged independent of OSN topology (small-world, scale-free, and random), existence of various types of communities, and user behavior (social reinforcement and user response times). This phenomenon occurs only when the OSN size is finite, and was reproduced by the urn model, which is a simplified version of Twitter-type information diffusion. This study mathematically formulates the model to theoretically verify this novel discovery.

Fig. 1 (left) shows the urn model representing a tweet spread through followers. The model repeats a trial, in which a ball is randomly extracted from the urn and then a ball is returned to the urn, until $p = 0$, where p corresponds to the number of users who will receive the tweet and its initial value $f(\geq 1)$ corresponds to the number of followers. There are N balls in the urn, and each is either black or white (the initial condition is that all balls are white). A black (white) ball corresponds to a user who has (not) posted a retweet. When a white ball is taken out of the urn, a black ball is returned to the urn with retweet probability λ and p is increased by $f-1$. In all the other cases, the ball extracted is returned and p is decreased by one. The number of black balls B in the urn at $p = 0$ corresponds to the final size of the cascade.

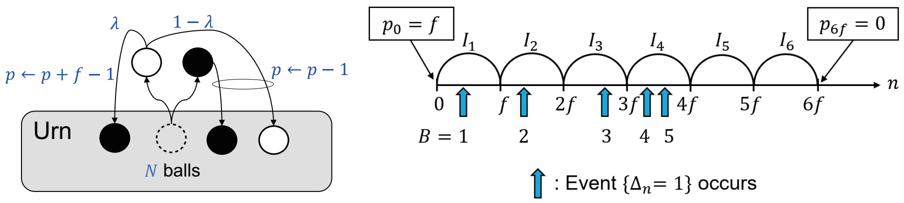

Fig. 1. Left: The urn model. Right: One of the cases where $B_\tau = 5$. $I_k = [(k-1)f+1, kf]$.

2 Formularisation

This section derives a stochastic process from the urn model. Let p_n and B_n be p and B immediately after the n-th trial, respectively. The stopping time τ of the trial is defined as

$$\tau = \inf\{n \geq 1 | p_n = 0\}. \tag{1}$$

The goal of this study is to verify the possibility of bipolarization using the probability distribution of B_τ. B_τ is determined by Markov chain $\{X_n\}$ given by $X_0 = (0, f)$, $X_n = (B_n, p_n)$, and $X_{n+1} = (B_n + \Delta_{n+1}, p_n + \Delta'_{n+1})$, where $\Delta_n \in \{0,1\}$ and $\Delta'_n \in \{-1, f-1\}$ represent increases in B and p due to the n-th trial, respectively, $\{\Delta_n = 0\} = \{\Delta'_n = -1\}$, $\{\Delta_n = 1\} = \{\Delta'_n = f-1\}$, and

$$P(\Delta_{n+1} = 1 | B_n = i) = 1 - P(\Delta_{n+1} = 0 | B_n = i) = \frac{N-i}{N}\lambda. \tag{2}$$

Proposition 1 *For any integers $k, f > 0$, $\{B_\tau = k\} = \{\tau = (k+1)f\}$.*

Proof. As shown in Fig. 1(right), event $\{p = 0\}$ occurs only at the end of one of intervals I_j, $j = 1, 2, \ldots$, where $I_j = [(j-1)f+1, jf]$. $B_\tau = k$ indicates that the trials are not performed at intervals I_j, $j > k+1$. If $p = 0$ occurs at I_j, $j \leq k$, B_τ must be smaller than k. Thus $\{B_\tau = k\} \subset \{\tau = (k+1)f\}$. If $\tau = (k+1)f$, $B_{(k+1)f} = k$. Thus $\{B_\tau = k\} \supset \{\tau = (k+1)f\}$.

3 Calculation of $P(B_\tau)$

3.1 Infinite User Size

As shown in Fig. 1(right), $\{B_\tau = k\}$ is determined by the first $(k+1)f$ trials $\Delta_1, \ldots, \Delta_{(k+1)f}$. Because $\Delta_n \in \{0,1\}$, $\omega = (\Delta_1, \ldots, \Delta_{(k+1)f})$ takes one of $2^{(k+1)f}$ different binary sequences. Let $\left(\begin{smallmatrix} & a & \\ b & & c \end{smallmatrix}\right)$ be the number of events ω that satisfy the condition that a trials generate b events $\{\Delta_n = 1\}$ and $\tau = c$. Then, $|\{\omega | B_{\tau(\omega)} = k\}| = \left(\begin{smallmatrix} & (k+1)f \\ k & (k+1)f \end{smallmatrix}\right)$. If N (the number of OSN users) is infinite, the right-hand side of (2) is $\lim_{N\to\infty} \frac{N-i}{N}\lambda = \lambda$, so $P(\Delta_{n+1} | B_n)$ is constant λ regardless of B_n. Thus, by using $X_0 = (0, f)$, $P(B_\tau = 0) = P(B_0 = 0)P(\Delta_1 = 0 | B_0 = 0) \cdots P(\Delta_f = 0 | B_{f-1} = 0) = (1-\lambda)^f$ and for $k > 0$,

$$P(B_\tau = k) = \begin{pmatrix} & (k+1)f \\ k & (k+1)f \end{pmatrix} \lambda^k (1-\lambda)^{(k+1)f-k}. \tag{3}$$

Proposition 2 *For any integers $k, f > 0$,*

$$\binom{kf}{k-1\ \ kf} = \binom{(k-1)f}{k-1} - \sum_{m=1}^{k-2} \binom{(k-1)f}{k-1\ \ mf}. \tag{4}$$

Proof. Because p_n becomes zero only at $n = f, 2f, 3f, \ldots$, $\{B_{kf} = k-1\} = \{B_{kf} = k-1\} \cap \bigcup_{m=1}^{k} \{\tau = mf\}$. From Proposition 1, $\{B_\tau = k-1\} = \{B_{kf} = k-1\} \cap \{\tau = kf\}$; therefore, $\{B_{kf} = k-1\} \setminus \{B_\tau = k-1\} = \{B_{kf} = k-1\} \cap \bigcup_{m=1}^{k-1} \{\tau = mf\}$. Because $|\{B_{kf} = k-1\}| = \binom{(k-1)f}{k-1}$, $\sum_{m=1}^{k-1} \binom{(k-1)f}{k-1\ \ mf} = |\{B_{kf} = k-1\} \cap \bigcup_{m=1}^{k-1} \{\tau = mf\}|$, and $\binom{(k-1)f}{k-1\ \ (k-1)f} = 0$, (4) holds.

Proposition 3 *For any integers k, f, m satisfying $2 \leq m < k$, $f > 0$,*

$$\binom{kf}{k\ \ mf} = \binom{mf}{m-1\ \ mf} \binom{(k-m)f}{k-m+1}. \tag{5}$$

Proof. $\binom{kf}{k\ \ mf}$ is the number of binary sequences $\omega = (\Delta_1, \ldots, \Delta_{kf})$ that satisfy $\{\tau = mf\}$. The number of binary sequences $(\Delta_1, \ldots, \Delta_{mf})$ satisfying $\{\tau = mf\}$ is $\binom{mf}{m-1\ \ mf}$, and the number of combinations in which events $\{\Delta_n = 1\}$ occur $k-(m-1)$ times out of sequences $(\Delta_{mf+1}, \ldots, \Delta_{kf})$ is $\binom{(k-m)f}{k-m+1}$.

Propositions 2 and 3 indicate that $\binom{kf}{k-1\ \ kf}$ is obtained using $\binom{mf}{m-1\ \ mf}$, $m = 1, 2, \ldots, k-2$, where $\binom{f}{0\ \ f} = 1$. Therefore, from (3), $P(B_\tau = k)$ can be derived in ascending order of k.

3.2 Finite User Size

Let us next consider the case N is finite and let $\lambda_i = \frac{N-i}{N}\lambda$. Probability $P(B_\tau = k)$ becomes a complicated equation when k is large, and its calculation time becomes long. Thus, the upper and lower bounds of $P(B_\tau = k)$ are derived.

Proposition 4 *Let $C = \binom{(k+1)f}{k\ \ (k+1)f} \prod_{i=0}^{k-1} \lambda_i$. $P(B_\tau = k)$ satisfies*

$$C \left(\prod_{i=0}^{k-1} (1-\lambda_i)^{f-1} \right) (1-\lambda_k)^f < P(B_\tau = k) < C(1-\lambda_k)^{(k+1)f-k}. \tag{6}$$

Proof. Let ω^* (ω_*) be one of ω values in $\{\omega | B_{\tau(\omega)} = k\}$ that maximizes (minimizes) probability $P(\{\omega\})$. From (3), $\binom{(k+1)f}{k\ \ (k+1)f} P(\omega_*) < P(B_\tau = k) < \binom{(k+1)f}{k\ \ (k+1)f} P(\omega^*)$. In the following, $P(\omega^*)$ and $P(\omega_*)$ are obtained. Assume that events $\{\Delta_n = 1\}$ occur k times at $n = n_1, n_2, \ldots, n_k$ ($n_1 < n_2 < \cdots$). From (2), $P(\Delta_{n_1} = 1 | B_{n_1-1} = 0) \cdots P(\Delta_{n_k} = 1 | B_{n_k-1} = k-1) = \prod_{i=0}^{k-1} \lambda_i$, which is independent of $(n_1, \ldots, n_k)$. For any ω, $P(\omega)$ is given as $P(\omega) = \delta \prod_{i=0}^{k-1} \lambda_i$, where $\delta = \prod_{i \in W} P(\Delta_i = 0 | B_{i-1})$ and $W = \{1, 2, \ldots, (k+1)f\} \setminus \{n_1, \ldots, n_k\}$. From (2), for any $n_i \in \{n_1, \ldots, n_k\}$, δ strictly decreases as n_i increases. Thus, δ is minimized (maximized) at $(n_1, \ldots, n_k) = (f, \ldots, kf)$ ($(n_1, \ldots, n_k) = (1, \ldots, k)$). Accordingly, $\prod_{i=0}^{k-1} (1-\lambda_i)^{f-1} (1-\lambda_k) \leq \delta \leq (1-\lambda_k)^{(k+1)f-k}$.

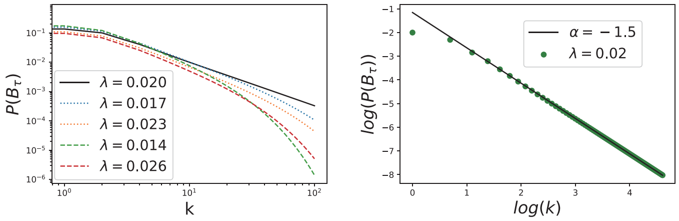

Fig. 2. Left: $P(B_\tau)$ at $N = \infty$ and $f = 50$. Right: Linear approximation with slope α at $f = 50$. The correlation coefficient of samples $(\log k, \log(P(B_\tau = k)))$, $k \geq 95$, is less than -0.999999999.

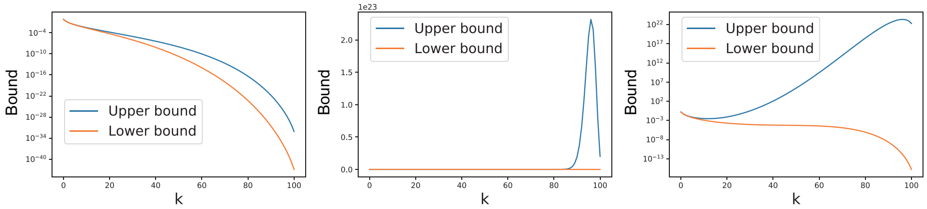

Fig. 3. Bounds of $P(B_\tau = k)$ at $N = 100$ and $f = 50$. Left: $\lambda = 0.01$. Center and right: $\lambda = 0.035$.

4 Numerical Results

Fig. 2 shows numerical results when $N = \infty$. As shown in Fig. 2 (left), $P(B_\tau = k)$ strictly decreases as k increases regardless of $f\lambda$, where $f\lambda$ is the expected number of followers who post retweets among f followers. If $f\lambda > 1$ ($f\lambda < 1$), the cascade is expected to grow forever (eventually stop growing). Fig. 2 (left) also indicates that $P(B_\tau = k)$ decreases slowly only if $f\lambda = 1$. Fig. 2 (right) verifies that the tail of $P(B_\tau = k)$ follows a power law $P(B_\tau = k) \propto k^{-1.5}$, which is similar to the result in [1]. Thus, Fig. 2 suggests that $f\lambda = 1$ is the necessary condition for large-scale cascade emergence when $N = \infty$.

Fig. 3 shows the upper and lower bounds of $P(B_\tau = k)$. Fig. 3(left) suggests that $P(B_\tau = k)$ is a decreasing function of k at a low retweet probability λ. In Figs. 3(center) and (right), the upper bound shows bipolarization. These figures demonstrate that $P(B_\tau = k)$ is not a decreasing function because the upper bound has a local minimum value (0.0031) at $k = 12$ and the sum of the lower bound for all $k \in \{0, 1, \ldots, 100\}$ is less than 0.3, so that $P(B_\tau = k)$ must be greater than 0.007 at least one $k > 12$. In other words, $P(B_\tau = k)$ has at least one peak at $k > 12$. In summary, $P(B_\tau = k)$ monotonically decreases if $N = \infty$, while it does not if N is finite and $f\lambda$ is large.

References

1. Gleeson, J.P., Onaga, T., Fennell, P., Cotter, J., Burke, R., OfSullivan, D.J.: Branching process descriptions of information cascades on twitter. Journal of Complex Networks 8(6), cnab002 (2020)
2. Oida, K.: Bi-polarization in cascade size distributions. IEEE Access 9, 72867–72880 (2021)