Discrete Mathematics and Theoretical Computer Science DMTCS vol. **17:2**, 2015, 235–250

On the Dynamics of Systems of Urns[*]

Jacek Cichoń Rafał Kapelko Marek Klonowski

Institute of Mathematics and Computer Science, Wrocław University of Technology, Poland

received 5th Sep. 2014, revised 30th Aug. 2015, accepted 2nd Oct. 2015.

In this paper we present an analysis of some generalization of the classic urn and balls model. In our model each urn has a fixed capacity and initially is filled with white balls. Black balls are added to the system of connected urns and gradually displace white balls. We show a general form of formulas for the expected numbers of black balls in a given urn and we analyze some special cases (parallel and serial configurations). We are mainly interested in a counterpart of the Coupon Collector Problem for the model considered.

The primary motivation for our research is the formal analysis of the mix networks (introduced by D. Chaum) and its immunity to so-called flooding (blending) attacks.

Keywords: urn and balls model, mix networks, Coupon Collector Problem

1 Introduction

In this paper we investigate the following process. We have a finite directed acyclic (i.e. without directed cycles) graph $\mathcal{G} = (V, E)$; we also have a family of urns $(U_a)_{a \in V}$ indexed by nodes of $\mathcal{G}$. Each urn contains some number of balls of two kinds - let us say, black and white balls. Initially, each urn contains only white balls. We say that a node $a \in V$ is a *source node* if there is no $b \in V$ such that $(b, a) \in E$. Similarly, a node $a \in V$ is a *sink node* if there is no $b \in V$ such that $(a, b) \in E$. At each round (enumerated by natural numbers),

1. we choose a random path $v_0, \ldots, v_k$ from a source to a sink,
2. we pick randomly balls $b_0, \ldots, b_k$ from urns $U_{v_0}, \ldots, U_{v_k}$,
3. we remove the ball b_k from U_{v_k},
4. for each $i < k$ we move the ball b_i to the urn $U_{v_{i+1}}$,
5. we put one black ball in the urn U_{v_0}.

[*]This paper was supported by Polish National Science Center – decision No. DEC-2013/09/B/ST6/02251

Notice that the number of balls in each urn is constant during the evolution of this process. We assume that all random choices during the execution of our process are done independently according to uniform distributions.

Notice that if $\mathcal{G} = (\{v\}, \emptyset)$, then our model is reduced to the standard, classic "urn model". In our investigations the classic model describes properties of the urns indexed by source nodes. The behavior of remaining urns U_b depends on the behavior of all urns from the family $\{U_a : (a, b) \in E\}$.

The first motivation of our research is the analysis of blending attacks on some kinds of mix networks analyzed in e.g., O'Connor (2005), Serjantov et al. (2002), Dingledine et al. (2006). Mix networks, introduced by D. L. Chaum in the paper Chaum (1981), are one of the most popular ways of protecting anonymous communication. More information about mix networks can be found in Section 7.

1.1 Organization of this paper

In Section 2 we investigate a general model of a system of urns. In Section 3 we present a special (but most important for applications) model wherein urns are arranged in a row and balls are consecutively moved from one urn to another. We present several results including some asymptotics. In Section 4 we investigate a system of urns arranged in a parallel. We compare some strategies of arranging urns in Section 5. In Section 7 we show how our results can be applied to analysis of so-called flooding (blending) attack against mix networks.

1.2 Notations and preliminary facts

In the paper, $\mathbf{E}[X]$ denotes the expected value of the random variable X. If $f(z) = \sum_i a_i \cdot z^i$ is a formal power series, then we define the coefficient extractor as $[z^i]f(z) = a_i$ (see Flajolet and Sedgewick (2009), Chapter 1). $|A|$ denotes the cardinality of a set A.

The partial exponential function $e_n(x)$ is defined by the formula $e_n(x) = \sum\limits_{k=0}^{n} \frac{x^k}{k!}$. The sequence $(e_n(n)/e^n)_{n\geq 0}$ is decreasing and

$$\frac{e_n(n)}{e^n} = \frac{1}{2} + \mathrm{O}\left(\frac{1}{\sqrt{n}}\right) \tag{1}$$

(see e.g. Weisstein (2013)).

The Euler Gamma function for $z > 0$ is defined by the formula $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$. For natural numbers n we have $\Gamma(n+1) = n!$. The Euler Beta function is defined for $a, b > 0$ by the formula $\mathrm{B}(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. It is well known that $\mathrm{B}(a, b) = (\Gamma(a)\Gamma(b))/\Gamma(a+b)$. The incomplete regularized Beta function is defined by he formula

$$\mathrm{I}(z; a, b) = \frac{1}{\mathrm{B}(a, b)} \int_0^z x^{a-1}(1-x)^{b-1}dx.$$

We will use the following two recurrences for the incomplete beta function:

$$\mathrm{I}(z; a, b) = z\mathrm{I}(z; a-1, b) + (1-z)\mathrm{I}(z; a, b-1) \tag{2}$$

and

$$\mathrm{I}(z; a, b) = \mathrm{I}(z; a-1, b) - \frac{\Gamma(a+b-1)}{\Gamma(a)\Gamma(b)}(1-z)^b z^{a-1}\,. \tag{3}$$

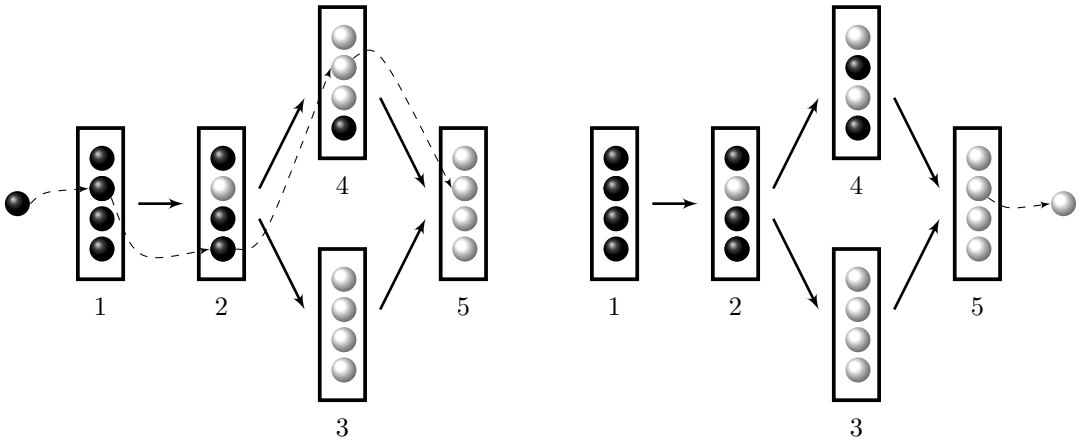

Fig. 1: Example of a system of connected urns. Graph $\mathcal{G} = (\{1,2,3,4,5\}, \{(1,2),(2,3),(2,4),(3,5),(4,5)\})$ just after round $t = 8$ (left side) and just after round $t = 9$ (right side). For each $a \in \{1,2,3,4,5\}$, we have $n_a = 4$. In this diagram we have $B_1(8) = B_1(9) = 4$, $B_4(9) = W_4(9) = 2$.

These formulas may be found in Weisstein (2013) and NIST (2013). They can also be proved by checking that derivatives (taken with respect to the variable z) of both sides of these equations are the same. Clearly, the initial values are the same in both sequences.

The incomplete beta function admits an analytic continuation for other values of parameters. We have $\mathrm{I}\,(z;a,0) = 0$ for $a > 0$ among others.

2 General Model

Let us fix a finite directed acyclic graph $\mathcal{G} = (V,E)$. Let $N = (n_a)_{a\in V}$ denote capacities of urns $(U_a)_{a\in V}$, i.e., n_a is the initial number of white balls in the urn U_a. We call the triple $\mathcal{U} = (V,E,N)$ a *system of connected urns*. Notice that the number of balls during each process is fixed for each urn. Indeed, each ball is replaced by exactly one ball.

Let us describe more precisely the choice of a random path in the graph $\mathcal{G}$. We start at one of the sources at random (equally likely). Then we take an edge out of it at random (all neighbors are equally likely). We do the same thing with every node we arrive at: choose one of the neighbors randomly. The walk is perpetuated till a sink is reached.

We denote by p_a the probability that a randomly chosen path from some source to some sink goes through the node a. For each $(a,b) \in E$, we denote by p_{ab} the probability of the event where the edge (a,b) is on the randomly chosen path. Notice that if a is a source node, then $p_a = \frac{1}{|S|}$, where S is the set of source nodes in the graph $\mathcal{G}$. Moreover, $p_{ab} = \frac{p_a}{out(a)}$, where $out(a) = |\{c \in V : (a,c) \in E\}|$ and

$$p_a = \sum_{(b,a)\in E} p_{ba}\ ,$$

when a is not a source node.

Let $B_a(t)$ denote the number of black balls in the urn U_a after the round t. Notice that $B_a(0) = 0$ for each $a \in V$. Our goal is to determine a difference equation for the sequence $(\mathbf{E}[B_a(t)])_{t\geq 0}$ for each $a \in V$.

For the sake of clarity of presentation, we add one artificial node $\bullet$ to our graph. We consider an extended system with nodes $V_\bullet = \{\bullet\} \cup V$, edges $E_\bullet = E \cup (\{\bullet\} \times S)$, where S is the set of source nodes in (V, E). We set capacity $n_\bullet = 1$ and we put initially one black ball to the urn $U_\bullet$. Then $B_\bullet(t) = 1$ for each t and $p_{\bullet s} = \frac{1}{|S|}$ for each $s \in S$.

Let us fix a node $a \in V$, the round number t and suppose that we know $\mathbf{E}[B_b(t)]$ for each $b \in V$. Then,

- at the $(t+1)^{\text{st}}$ round the number of black urns in the node a may increase by 1 if there is b such that (b, a) is on the chosen path, a black ball is selected from U_b and a white ball is selected from U_a. This happens with probability $p_{ba} \cdot \frac{B_b(t)}{n_b} \cdot \frac{n_a - B_a(t)}{n_a}$.
- at the $(t+1)^{\text{st}}$ round the number of black urns in the node a may decrease by 1 if there is b such that (b, a) is on the chosen path, a white ball was selected from U_b and a black ball was selected from U_a. This happens with probability $p_{ba} \cdot \frac{n_b - B_b(t)}{n_b} \cdot \frac{B_a(t)}{n_a}$.
- with the remaining probability the state is not changed. This is the case if a black ball is replaced by another black ball or a white ball by a white one.

Putting these facts together, we deduce that

$$\mathbf{E}[B_a(t+1)] = \sum_{(b,a)\in E} \frac{p_{b,a}}{n_b} \mathbf{E}[B_b(t)] + \left(1 - \frac{p_a}{n_a}\right) \mathbf{E}[B_a(t)] \ . \tag{4}$$

Let $W_a(t)$ denote the number of white balls in the urn U_a after the round t. Notice that $W_a(t) = n_a - B_a(t)$ for each $a \in V_\bullet$. In particular, $W_\bullet(t) = 0$ for each t. More generally, if $d(a)$ denotes the minimal distance of the node a from some source link from V, then for each $t \leq d(a)$ we have $W_a(t) = n_a$.

After a simple transformation we get

$$\mathbf{E}[W_a(t+1)] = \sum_{(b,a)\in E} \frac{p_{b,a}}{n_b} \mathbf{E}[W_b(t)] + \left(1 - \frac{p_a}{n_a}\right) \mathbf{E}[W_a(t)] \ . \tag{5}$$

Notice that recurrences (4) and (5) are the same. The solutions of these recurrences differ, as they have different initial conditions. Let

$$F_a(x) = \sum_{t\geq 0} \mathbf{E}[W_a(t)]\, x^t$$

be the generating function for the sequence $(\mathbf{E}[W_a(t)])_{t\geq 0}$. From (5) and from the fact that $F_a(0) = n_a$, we get the following equation for each $a \in V$:

$$F_a(x) = \frac{n_a + x \sum_{(b,a)\in E_\bullet} \frac{p_{ba}}{n_b} F_b(x)}{1 - \Delta_a x} \ ,$$

where $\Delta_a = 1 - \frac{p_a}{n_a}$.

Let $prec(a)$ denote the set of all nodes from V for which there is some oriented path to the node a (we include the node a in $prec(a)$). For any real number x and a path $\sigma = (b_1, \ldots, b_n)$ from some source to node a we define ct function that counts the number of nodes on σ with Δ parameter equal x:

$$ct_a(x, \sigma) = |\{k \in \{1, \ldots, n\} : x = \Delta_{b_k}\}|$$

and, finally, we put

$$ct_a(x) = \max\{ct_a(x, \sigma) : \sigma \text{ is a path from some source to } a\} .$$

If $p(t)$ is a polynomial, then $deg(p)$ denotes the degree of p. We shall formulate a theorem which reduces the problem of finding closed formulas for $\mathbf{E}[W_a(t)]$ to problems of linear algebra. This theorem may be treated as a specialized form of a theorem about expansion of rational functions (see e.g. Theorem IV.9 from Flajolet and Sedgewick (2009)).

Theorem 1 *Let $a \in V$ and let $D(a) = \{\Delta_b : b \in prec(a)\}$. Then there are polynomials $(p_\Delta(t))_{\Delta \in D(a)}$ such that*

$$(\forall t \geq 0) \left(\mathbf{E}[W_a(t)] = \sum_{\Delta \in D(a)} p_\Delta(t) \Delta^t \right)$$

and $deg(p_\Delta) < ct_a(\Delta)$ for each $\Delta \in D(a)$.

Proof: We claim that for each $a \in V$ there are polynomials $(\alpha^a_\Delta)_{\Delta \in D(a)}$ and integers $(k^a_\Delta)_{\Delta \in D(a)}$ such that $deg(\alpha^a_\Delta) < k^a_\Delta \leq ct_a(\Delta)$ and

$$F_a(x) = \sum_{\Delta \in D(a)} \frac{\alpha^a_\Delta(x)}{(1 - \Delta x)^{k^a_\Delta}} .$$

Observe that if a is a source node, then $F_a(x) = \frac{n_a}{1 - \Delta_s}$, so the claim is true for any source node. Hence, suppose that the claim is true for all $b \in prec(a) \setminus \{a\}$. From the recurrence (5) we get

$$F_a(x) = \frac{n_a}{1 - \Delta_a x} + \sum_{(b,a) \in E} \sum_{\Delta \in D(b)} \frac{x \cdot \alpha^b_\Delta(x)}{(1 - \Delta x)^{k^b_\Delta}(1 - \Delta_a \cdot x)} .$$

Let us consider a term

$$\tau = \frac{x \cdot \alpha(x)}{(1 - \Delta x)^k (1 - \Delta_a x)} ,$$

where $deg(\alpha) < k$. If $\Delta_a = \Delta$, then $\tau = \frac{x \cdot \alpha(x)}{(1 - \Delta x)^{k+1}}$ and $deg(x \cdot \alpha) < k + 1$. If $\Delta_a \neq \Delta$, then we decompose τ into partial fractions and we find a constant B and a polynomial β such that $deg(\beta) < k$ and

$$\tau = \frac{B}{1 - \Delta_a} + \frac{\beta(x)}{(1 - \Delta x)^k} .$$

This proves the claim.

Let us fix $\Delta \in D(a)$ and let us consider the term $\frac{\alpha(x)}{(1-\Delta)^k}$, where $\alpha(x)$ is some polynomial of degree less than k. Then

$$\frac{\alpha(x)}{(1-\Delta)^k} = \sum_{l=0}^{k-1} \alpha_l \frac{x^l}{(1-\Delta)^k}$$

for some constants $\alpha_0, \ldots, \alpha_{k-1}$. Observe that

$$\frac{x^l}{(1-\Delta)^k} = \sum_{m \geq 0} \binom{m+k-1}{k-1} \Delta^m x^{m+l} .$$

Therefore,

$$[x^t]\frac{x^l}{(1-\Delta)^k} = \Delta^{t-l}\binom{t-l+k-1}{k-1} = \Delta^t \frac{\binom{t-l+k-1}{k-1}}{\Delta^l} = \Delta^t \beta(t) ,$$

where β is a polynomial and $deg(\beta) = k-1$. The term $\frac{\alpha(x)}{(1-\Delta)^k}$ is of the same form. □

Theorem 1 gives us a general form of a solution of the recurrence equation (5) for a fixed structure $\mathcal{U} = (V, E, N)$. In order to find required coefficients of polynomials $(p_\Delta)_{\Delta \in D}$ we may use the equations $W_a(0) = \ldots = W_a(d(a)) = n_a$ and, if $d(a)$ is too small, we may solve explicitly the equation (5) for sufficiently many small values of the time parameter t.

Example Let $\mathcal{G} = (\{0,1,2\}, E, N)$, where E=$\{(0,1),(1,2)\}$ and N_0 = N_1 = N_2 = $n > 1$. Then, $prec(2) = \{0,1,2\}$, $p_0 = p_1 = p_2 = 1$, so $\Delta_0 = \Delta_1 = \Delta_2 = 1 - \frac{1}{n}$. Hence, $D(2) = \{1 - \frac{1}{n}\}$ and $deg_2(1-\frac{1}{n}) = 3$. From Theorem 1 we deduce that $\mathbf{E}[W_2(t)] = p(t)(1-\frac{1}{n})^t$ for some polynomial $p(t)$ of degree less than 3. Let $p(t) = a + bt + ct^2$. Notice that $\mathbf{E}[W_2(0)]$ = $\mathbf{E}[W_2(1)]$ = $\mathbf{E}[W_2(2)]$ = n, hence $n = p(0)(1-\frac{1}{n})^0 = a$, $n = p(1)(1-\frac{1}{n})$ = $(a+b+c)(1-\frac{1}{n})$ and $n = p(2)(1-\frac{1}{n})^2$ = $(a+2b+4c)(1-\frac{1}{n})^2$. The solution of this system of linear equations (in the variables a, b and c) is given by $a = n$, $b = -\frac{3n-2n^2}{2(n-1)^2}$, $c = \frac{n}{2(n-1)^2}$, hence

$$\mathbf{E}[W_2(t)] = n\left(1 + \frac{2n-3}{2(n-1)^2}t + \frac{1}{2(n-1)^2}t^2\right)\left(1-\frac{1}{n}\right)^t .$$

In a similar way we may show that $\mathbf{E}[W_1(t)] = n\left(1+\frac{t}{n-1}\right)\left(1-\frac{1}{n}\right)^t$ and, of course, that $\mathbf{E}[W_0(t)]$ = $n\left(1-\frac{1}{n}\right)^t$.

It is clear that in a similar way we can analyze any system of urns with the underlying graph of the form $(\{0,\ldots,n+1\}, E)$ where $E = \{(k,k+1) : k = 0,\ldots,n\}$. However, in the next section we show a more uniform approach to this class of graphs. Theorem 1 will be explicitly used in Sections 4 and 6.2.

3 Serial System of Urns

Let us fix two parameters n and k. Let $\mathcal{U}_k = (\{0,\ldots,k\}, \{(a,a+1) : a < k\}, N)$, where $n_a = n$ for each $a \in \{0,\ldots,k\}$. That is, each urn has the same capacity. Therefore, at each round we move one ball from the urn U_a to U_{a+1} (if $a < k$), we remove one ball from the k^{th} urn and we add one black ball to the

urn U_0. As before, we assume that we choose the balls independently, using uniform distributions. Notice that this model strictly corresponds to so-called *MIX-cascade* described in Section 7.

The evolution of this model is described by the vector $B(t) = (B_0(t), \ldots, B_k(t))$ of black balls in consecutive urns. Notice that $B(0) = (0, \ldots, 0)$. Let us observe that the random variable B_0 describes the classic urn and balls problem and $\mathbf{E}[B_0(t)] = n\left(1 - \left(1 - \frac{1}{n}\right)^t\right)$. Clearly, $B_0(0) \leq B_0(1) \leq B_0(2) \leq \ldots$. But if $k > 0$, then for $0 < a \leq k$ the processes $(B_a(t))_{t \geq 0}$ are not monotonic with probability 1 - for example $\Pr[B_1(2) = 1, B_1(3) = 0] = \frac{n-1}{n^3} > 0$.

3.1 Difference equations

Since we are going to analyze behavior of the system $\mathcal{U}_k$ for arbitrary k, instead of using Theorem 1 we shall solve directly the recurrence 4 adapted to this case. Notice that for each $a \in \{0, \ldots, k\}$ we have $p_a = 1$ and that $p_{a,a+1} = 1$ for each $a = 0, \ldots, k-1$. For $a \in \{1, \ldots, k\}$ we obtain the following recurrence from equation 4:

$$\mathbf{E}[B_a(t+1)] = \frac{1}{n}\mathbf{E}[B_{a-1}(t)] + (1 - \frac{1}{n})\mathbf{E}[B_a(t)]\,. \tag{6}$$

Let $y_a(t) = \mathbf{E}[B_a(t)]$, $\delta = \frac{1}{n}$ and $\Delta = 1 - \frac{1}{n}$. Then the initial observation and the equation (6) can be rewritten as

$$\left\{ \begin{array}{rcl} y_0(t) & = & n(1 - \Delta^t), \\ y_{a+1}(t+1) & = & \Delta \cdot y_{a+1}(t) + \delta \cdot y_a(t)\,. \end{array} \right. \tag{7}$$

Let us also recall that $y_a(0) = 0$ for each a. It is also clear that $y_a(t) = 0$ for each $t \leq a$.

3.2 Closed formula

In this section we will show that there exists a closed formula for the expected value of the random variable $B_a(t)$ for arbitrary a and t.

Theorem 2 *For each $a \geq 0$ and $t \geq 0$ we have*

$$\mathbf{E}[B_a(a+t)] = n \cdot \mathrm{I}\left(\frac{1}{n}; a+1, t\right)\,.$$

Proof: Let $z_a(t) = y_a(a+t)$ and $\Delta = 1 - \frac{1}{n}$ and $\delta = \frac{1}{n}$. Notice that $z_a(0) = 0$ for each a, $z_0(t) = n \cdot (1 - \Delta^t)$ and

$$z_{a+1}(t+1) = y_{a+1}((a+t+1)+1) = \Delta \cdot y_{a+1}(a+1+t) + \delta \cdot y_a(a+t+1) = \Delta \cdot z_{a+1}(t) + \delta \cdot z_a(t+1)\,.$$

Therefore, the equations (7) may be rewritten as follows:

$$\left\{ \begin{array}{rcl} z_0(t) & = & n(1 - \Delta^t), \\ z_{a+1}(t+1) & = & \Delta \cdot z_{a+1}(t) + \delta \cdot z_a(t+1)\,. \end{array} \right. \tag{8}$$

Plain calculations show that

$$n \cdot \mathrm{I}\left(\frac{1}{n}; 1, t\right) = n\frac{\Gamma(t+1)}{\Gamma(1)\Gamma(t)} \int_0^{\frac{1}{n}} (1-x)^{t-1} dx = n\left(1 - \left(1 - \frac{1}{n}\right)^t\right) ,$$

so $z_0(t) = n\mathrm{I}\left(\frac{1}{n}; 1, t\right)$. From equation (2) applied for $z = \frac{1}{n}$ we get

$$n \cdot \mathrm{I}\left(\frac{1}{n}; a+2, t+1\right) = \Delta \cdot n \cdot \mathrm{I}\left(\frac{1}{n}; a+2, t\right) + \delta \cdot n \cdot \mathrm{I}\left(\frac{1}{n}; a+1, t+1\right) . \tag{9}$$

Moreover, $n \cdot \mathrm{I}\left(\frac{1}{n}; a+1, 0\right) = 0$, so the sequences $(z_a(t))_{a,t}$ and $(n \cdot \mathrm{I}\left(\frac{1}{n}; a+1, t\right))_{a,t}$ satisfy the same recurrence relations. □

Theorem 2 gives us a closed formula for the expected number of black balls after a given number of steps in a given urn. However, we need another formula for $\mathbf{E}\left[B_a(a+t+1)\right]$ that is convenient for the investigation of properties of a fixed urn and for various values of t.

Theorem 3 *For each $a \geq 0$ and $t \geq 0$ we have*

$$\mathbf{E}\left[B_a(a+t+1)\right] = n\left(1 - \left(1 - \frac{1}{n}\right)^{t+1} \sum_{k=0}^{a} \binom{k+t}{k} \frac{1}{n^k}\right) . \tag{10}$$

Proof: From formula (3) we deduce that

$$\mathrm{I}\left(z; a+1, t+1\right) = \mathrm{I}\left(z; a, t+1\right) - \binom{a+t}{a}(1-z)^{t+1} z^a .$$

We also know that

$$\mathrm{I}\left(z; 1, t+1\right) = 1 - (1-z)^{t+1} = 1 - \binom{0+t}{0}(1-z)^{t+1} z^0,$$

so we get

$$\mathrm{I}\left(z; a+1, t+1\right) = 1 - \sum_{k=0}^{a} \binom{k+t}{k}(1-z)^{t+1} z^k .$$

After putting $z = \frac{1}{n}$ into this formula and using Theorem 2, we get the desired identity. □

3.3 Asymptotic behavior

In this section we investigate asymptotic behavior of the system of urns. In particular, it is important for us when the first black ball appears in a particular urn, when the fixed urn is full of black balls and when a big portion of balls in a given urn is black.

3.3.1 When almost all balls are black

In this section we shall investigate the moment at which almost all balls in a given layer are black. More precisely, we fix the number n of balls in each layer and we want to approximate the moment when $n-1$ balls in a given urn are black.

Lemma 1 *Let* $t_n = n(\ln(n) + (a+\nu)\ln(\ln(n)) - \ln(a!))$ *for some* $\nu \geq 0$. *If* $0 \leq k < a+\nu$, *then*

$$n\left(1-\frac{1}{n}\right)^{t_n}\binom{k+t_n}{k}\frac{1}{n^k} = \mathrm{O}\left(\frac{1}{(\ln n)^{a+\nu-k}}\right) .$$

and for $k = a$ *and* $\nu = 0$ *we have*

$$n\left(1-\frac{1}{n}\right)^{t_n}\binom{a+t_n}{k}\frac{1}{n^a} = 1 + \mathrm{O}\left(\frac{\ln\ln n}{\ln n}\right) .$$

The proof of this lemma can be instantly deduced from the following equation

$$\ln\left(1-\frac{1}{n}\right)^{t_n} = -t_n \ln\frac{1}{1-\frac{1}{n}} = -\frac{t_n}{n}\left(1+\mathrm{O}\left(\frac{1}{n}\right)\right) . \tag{11}$$

Theorem 4 *If* a *is fixed and* $t_n = n\left(\ln(n) + a\ln(\ln(n)) - \ln(a!)\right)$, *then*

$$\mathbf{E}\left[B_a(a+t_n)\right] = n - 1 + \mathrm{O}\left(\frac{\ln\ln n}{\ln n}\right)$$

as n *approaches infinity.*

Proof: The theorem is a consequence of Lemma 1 (with $\nu = 0$) and Theorem 3. □

Notice that for $a = 0$ we get $\mathbf{E}\left[B_0(n\ln n)\right] = n - 1 + \mathrm{o}\left(n\right)$, so our result is consistent with the classic Coupon Collector Problem (see Feller (1965)). We shall come back to this observation in Sec. 6. Below we show that just after the time $n(\ln n + \ln((\ln n)^a/a!))$ all the balls in a^{th} urn are black with high probability.

Theorem 5 *If* a *and* $\nu > 0$ *are fixed and* $t_n = n(\ln(n) + (a+\nu)\ln(\ln(n)) - \ln(a!))$, *then*

$$\Pr[B_a(a+t_n) \neq n] = \mathrm{O}\left(\frac{1}{(\ln n)^{\nu}}\right) .$$

as n *approaches infinity.*

Proof: Notice that the number of white balls in the t_n-th round in the a^{th} urn is $W_a(t_n) = n - B_a(a+t_n)$. Thus, $W_a(t_n) \neq 0$ is equivalent to $B_a(a+t_n) \neq n$. However, $\Pr[W_a(t_n) \neq 0] \leq \mathbf{E}\left[W_a(t_n)\right] = n - \mathbf{E}\left[B_a(a+t_n)\right]$ as $W_a(t_n)$ is integer valued, non-negative random variable. To prove our theorem it is sufficient to show that $\mathbf{E}\left[B_a(a+t_n)\right] = n - \mathrm{O}\left(\frac{1}{(\ln n)^{\nu}}\right)$. This is the consequence of Lemma 1 for $\nu > 0$ and Theorem 3 . □

3.3.2 First black ball

Let us recall that $B_0(1) = 1$.

Theorem 6 *If a is fixed and $t_n = ((a+1)!n^a)^{\frac{1}{a+1}}$, then*

$$\lim_{n\to\infty} \mathbf{E}\left[B_a(a+t_n)\right] = 1 \ .$$

Proof: Let us observe that for all $x \in [0, \frac{1}{n}]$ we have $x^a(1-\frac{1}{n})^t \leq x^a(1-x)^{t-1} \leq x^a$. Hence

$$\frac{n}{\mathrm{B}(a+1,t)}\left(1-\frac{1}{n}\right)^{t-1}\int_0^{\frac{1}{n}} x^a dx \leq n\mathrm{I}\left(\frac{1}{n};a+1,t\right)$$

and

$$n\mathrm{I}\left(\frac{1}{n};a+1,t\right) \leq \frac{n}{\mathrm{B}(a+1,t)}\int_0^{\frac{1}{n}} x^a dx \ ,$$

so

$$\frac{\Gamma(a+t+1)}{\Gamma(t)}\left(1-\frac{1}{n}\right)^{t-1}\frac{1}{(a+1)!n^a} \leq n\mathrm{I}\left(\frac{1}{n};a+1,t\right) \leq \frac{\Gamma(a+t+1)}{\Gamma(t)}\frac{1}{(a+1)!n^a} \ .$$

Notice that

$$\frac{\Gamma(a+t+1)}{\Gamma(t)} = t^{a+1}\prod_{j=0}^{a}\left(1+\frac{j}{t}\right) \ ,$$

so

$$\frac{\Gamma(a+t_n+1)}{\Gamma(t_n)} = (a+1)!\cdot n^a\left(1+\mathrm{O}\left(\frac{1}{n}\right)^{\frac{a+1}{a}}\right) \ .$$

Moreover, from equation (11) we get

$$\left(1-\frac{1}{n}\right)^{t_n} = 1+\mathrm{O}\left(\frac{1}{n}\right)^{\frac{1}{a+1}} \ ,$$

which accomplishes the proof. □

3.3.3 When nearly half the balls are black

Let us recall that $e_a(a)/e^a$ is close to $\frac{1}{2}$ (see Eq. (1)). We will show that at time $a \cdot n$ nearly half of the balls in a^{th} urn are black.

Theorem 7 *If a is fixed and $t_n = a \cdot n$, then*

$$\mathbf{E}\left[B_a(a+t_n+1)\right] = n\left(1-\frac{e_a(a)}{e^a}\right)+\mathrm{O}\left(1\right) \ .$$

Proof: Applying equation (11) we can easily prove that

$$\left(1-\frac{1}{n}\right)^{an+1} = e^{-a}(1+\mathrm{O}\left(\frac{1}{n}\right)) .$$

Moreover,

$$\binom{k+an}{k} = \frac{1}{k!}\prod_{j=0}^{k-1}(k+an-j) = \frac{a^k n^k}{k!}\prod_{l=1}^{k}(1+\frac{l}{an}) = \frac{a^k n^k}{k!}(1+\mathrm{O}\left(\frac{1}{n}\right)) ,$$

and

$$\left(1-\frac{1}{n}\right)^{an+1}\binom{k+an}{k}\frac{1}{n^k} = e^{-a}\frac{a^k}{k!}(1+\mathrm{O}\left(\frac{1}{n}\right)) .$$

From Theorem 3 we get

$$\mathbf{E}\left[Y_a(a+an+1)\right] = n\left(1-\sum_{k=0}^{a}e^{-a}\frac{a^k}{k!}\left(1+\mathrm{O}\left(\frac{1}{n}\right)\right)\right) =$$

$$n\left(1-e^{-a}\cdot\sum_{k=0}^{a}\frac{a^k}{k!}\cdot\left(1+\mathrm{O}\left(\frac{1}{n}\right)\right)\right) = n\left(1-\frac{e_a(a)}{e^a}\left(1+\mathrm{O}\left(\frac{1}{n}\right)\right)\right) =$$

$$n\left(1-\frac{e_a(a)}{e^a}+\mathrm{O}\left(\frac{1}{n}\right)\right) = n\left(1-\frac{e_a(a)}{e^a}\right)+\mathrm{O}\left(1\right) ,$$

so Theorem 7 is proved. □

4 Parallel System of Urns

In the previous section we investigated a serial system of urns. In this section we consider another variant of the system, namely, let $\mathcal{G} = (\{0,\ldots,k+1\}, E)$, where $E = (\{0\}\times\{1,\ldots,k\})\cup(\{1,\ldots,k\}\times\{k+1\})$. We assume, as before, that the capacity of all urns is n and that at the beginning of considered process all balls are white.

At each step we select one path $0 \to i \to (k+1)$, where $i \in \{1,\ldots,k\}$ with the same probability. Next, we choose balls from selected urns, we move the balls according to the arrows and put a black ball in the selected place in the urn U_0. Observe that $p_0 = p_{k+1} = 1$, $p_{0a} = \frac{1}{k}$, $p_{a(k+1)} = 1$ and $p_a = \frac{1}{k}$ for each $a \in \{1,\ldots,k\}$.

Let $W_{k+1}(t)$ denote the number of white balls in the urn U_{k+1} after t^{th} step. From Theorem 1 we know that

$$\mathbf{E}\left[W_{k+1}(t)\right] = a\left(1-\frac{1}{kn}\right)^t + (b+ct)\left(1-\frac{1}{n}\right)^t$$

for some coefficients a, b and c. Moreover, $\mathbf{E}\left[W_{k+1}(0)\right] = \mathbf{E}\left[W_{k+1}(1)\right] = \mathbf{E}\left[W_{k+1}(2)\right] = n$, from which we deduce that $a = \frac{k^2 n}{(k-1)^2}$, $b = \frac{n-2kn}{(k-1)^2}$, $c = -\frac{n}{(k-1)(n-1)}$, so

$$\mathbf{E}\left[W_{k+1}(t)\right] = n\left(\frac{k^2}{(k-1)^2}\left(1-\frac{1}{kn}\right)^t + \left(\frac{1-2k}{(k-1)^2}+\frac{-t}{(k-1)(n-1)}\right)\left(1-\frac{1}{n}\right)^t\right) .$$

Since $B_{k+1}(t)$ is the number of black balls in corresponding urns after t^{th} step, we get:

$$\mathbf{E}\left[B_{k+1}(t)\right] = n\left(1 - \frac{k^2}{(k-1)^2}\left(1-\frac{1}{kn}\right)^t + \left(\frac{2k-1}{(k-1)^2} + \frac{t}{(k-1)(n-1)}\right)\left(1-\frac{1}{n}\right)^t\right) . \quad (12)$$

Remark The parallel system considered in this section consisting of $k+2$ urns is equivalent, from the point of view of the last urn, to the serial system of three urns: the first urn has capacity n, the second urn has capacity $k \cdot n$ and the third one has capacity n.

4.1 Asymptotics

The formula (12) is much easier to analyze than the formula (3). We formulate without proofs four results about the parallel system of urns which corresponds to Theorems 4, 5, 6 and 7:

Theorem 8 *If a is fixed and n grows to infinity then*

1. $\mathbf{E}\left[B_{a+1}\left(a\cdot n\cdot(\ln n + 2\ln(\frac{a}{a-1})\right)\right] = n - 1 + \mathrm{O}\left(\frac{\ln n}{n}\right)$,
2. $\Pr[Z_{a+1}(a\cdot n\cdot(\ln n + 2\ln(\frac{a}{a-1}) + \ln(\ln(n)))) \neq n] = \mathrm{O}\left(\frac{1}{\ln n}\right)$,
3. $\mathbf{E}\left[B_{a+1}\left(\sqrt[3]{6an^2}\right)\right] = 1 + \mathrm{O}\left(\frac{1}{n^{\frac{1}{3}}}\right)$,
4. *if $a \geq 2$ then*
$$\lim_{n\to\infty}\frac{\mathbf{E}\left[B_{a+1}\left(n\cdot a\cdot\ln\frac{2a^2}{(a-1)^2}\right)\right]}{n} = \frac{1}{2} + \epsilon_a\ .$$
where $\epsilon_2 = 0.1112\ldots$ and the sequence $(\epsilon_a)_a$ is decreasing and $\lim_{a\to\infty}\epsilon_a = 0$.

5 Comparison of Serial and Parallel Systems of Urns

In the following table we compare dynamics of the sink node in the parallel system of urns with total $k+2$ urns and the dynamics of the sink node in the serial system of urns also consisting of $k+2$ urns. Let us recall that $k \geq 2$.

Expected number of black balls	Serial structure	Parallel structure
1	$((k+2)!n^{k+1})^{\frac{1}{k+2}}$ (Thm. 6)	$\sqrt[3]{6kn^2}$ (Thm. 8.3)
$\sim \frac{1}{2}n$	$(k+1)n$ (Thm. 7)	$nk\ln\frac{2k^2}{(k-1)^2}$ (Thm. 8.4)
$n-1$	$n(\ln n + (k+1)\ln\ln n - \ln(k+1)!)$ (Thm. 4)	$kn(\ln n + 2\ln\frac{k}{k-1})$ (Thm. 8.1)

This table shows how many black balls have to be put to the system to obtain given expected number of black balls in the sink urn in serial and parallel structures with the same total capacity. Notice that in the serial model the first black ball appears in the sink urn in time $\theta(n^{\frac{k+1}{k+2}})$, while in the parallel - in time $\theta(n^{\frac{2}{3}})$. On the other hand, in the serial model the sink urn is almost full at the time $\sim n\log n$, while in the

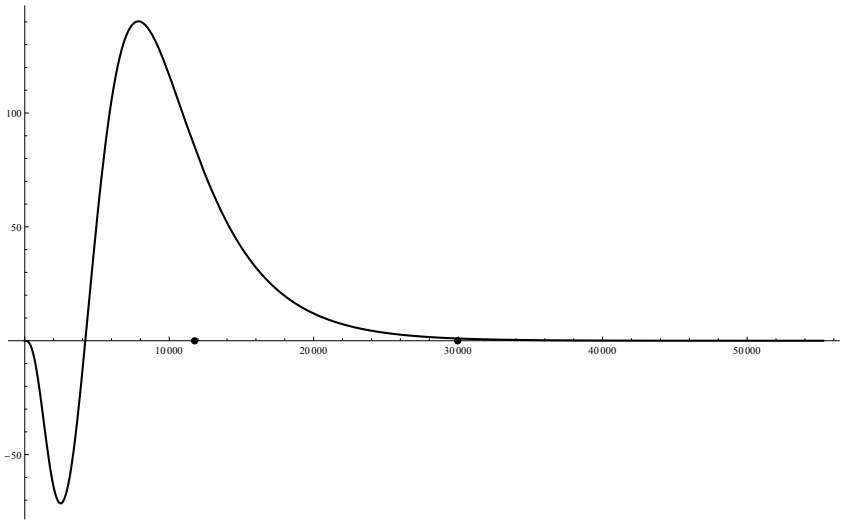

Fig. 2: Graphs of the difference $\mathbf{E}\left[S_4(t)\right] - \mathbf{E}\left[P_4(t)\right]$ for $n = 1000$. $S_4(t)$ is the number of black balls in the serial model and $P_4(t)$ is the number of black balls in the parallel model.

parallel model - in time $\sim kn \log n$. Figure 2 shows a graph of the difference between expected values of occupancy of the sink urns in serial and parallel models for $k = 4$. The first dot represents the moment when the sink urn in the serial model is almost full and the second dot - the moment when the sink urn in the parallel model is almost full.

6 Final Remarks

We presented results for two classes of urn and bin models with applications to the security analysis of MIX-servers. The results from this paper partially describe properties of systems of urns and some interesting theoretical questions are left unanswered. In the next section we formulate some of our additional observations.

6.1 Expected time

Let us consider a serial system of two urns. Let $T_n^{(2)}$ denote the first round when all balls in the second urn are black. We may model this system as a Markov chain with states $M = \{(x, y) : x, y \in \{0, \ldots, n\}\}$, where (x, y) corresponds with the situation when there are x black balls in the first urn and there are y black balls in the second urn. Then, $T_n^{(2)}$ is the hitting time of a set $A = \{0, \ldots, n\} \times \{n\}$ for the Markov chain M starting from the state $(0, 0)$. Let $h_{a,b}$ denote the hitting time of A for Markov chain starting from the state (a, b). Then, for $(a, b) \notin A$ we have

$$h_{a,b} = \frac{1}{1 - \frac{ab}{nn}} \left(\frac{(n-a)(n-b)}{n^2} h_{a+1,b} + \frac{b(n-a)}{n^2} h_{a+1,b-1} + \frac{a(n-b)}{n^2} h_{a,b+1} + 1 \right).$$

The number $\mathbf{E}\left[T_n^{(2)}\right] = h_{0,0}$ can be evaluated numerically for reasonably small values of the parameter n. Our calculus suggests the following **hypothesis**:

$$\mathbf{E}\left[T_n^{(2)}\right] = n(H_n + \ln\ln n + o(1))\ ,$$

where H_n denotes the n^{th} Harmonic number.

6.2 System with different capacities

Let us consider a serial system consisting of two urns, where the first urn has capacity n, the second urn has capacity m and $m \neq n$. We call such structure the (n, m) system. Let $B_{n,m}(t)$ and $Z_{n,m}(t)$ denote the number of white and black balls in the second urn after t^{th} round. From Theorem 1 we deduce that

$$\mathbf{E}\left[W_{n,m}(t)\right] = a\left(1 - \frac{1}{n}\right)^t + b\left(1 - \frac{1}{m}\right)^t$$

for some constants a and b. Taking into account that $\mathbf{E}\left[W_{n,m}(0)\right] = \mathbf{E}\left[W_{n,m}(1)\right] = m$, we get $a = \frac{mn}{n-m}$, $b = \frac{m^2}{m-n}$, so finally we get

$$\mathbf{E}\left[B_{n,m}(t)\right] = m\left(1 - \frac{n}{n-m}\left(1 - \frac{1}{n}\right)^t + \frac{m}{n-m}\left(1 - \frac{1}{m}\right)^t\right)\ .$$

Notice that $\frac{\mathbf{E}[B_{n,m}(t)]}{m} = \frac{\mathbf{E}[B_{m,m}(t)]}{n}$. Therefore, if $n = a \cdot m$ and $\mathbf{E}\left[B_{n,m}(t)\right] = m - 1$, then $\mathbf{E}\left[B_{m,n}(t)\right] = n - a$. Hence, if $a \gg 1$, then the filling time for the $(m, a \cdot m)$ system is essentially longer than the time required for filling the $(a \cdot m, m)$ system.

7 Application to Mix Networks

As mentioned in the introduction, the primary motivation for our research comes from security analysis of mix networks. Notice that the idea of mix networks is the essential building block for all anonymity preserving methods used in practice in the network communication (including TOR protocol (Dingledine et al. (2004))).

Mix network We consider a system with senders, receivers and a special-purpose unit (called the MIX-server). Messages are not sent directly from senders to receivers but every message is sent to the MIX-server as a ciphertext. Several messages entering MIX-server are collected, cryptographically recoded, randomly permuted and forwarded to respective receivers. Thanks to it, messages leaving the MIX-server become (from the external observer's perspective) indistinguishable. Thus, the sender of a message cannot be linked with the recipient of this message and this should guarantee anonymity (so-called *unlinkability* of senders and receivers). Substantially different variants of the basic protocol adjusted to particular conditions (acceptable latency, volume of the traffic etc.) appeared in a well-developed body of literature devoted to anonymous communication (see e.g. Danezis and Diaz (2008)).

MIX cascade Let us notice that the MIX-server knows the correspondence between senders and receivers. That is, if the single MIX-server is corrupted, then the adversary can link senders with receivers. The simplest remedy that improves the security is to use several MIX-servers instead of a single one. The basic realization of this idea called MIX-*cascade* has already been suggested in the seminal paper (Chaum (1981)). In this protocol, a message, to be delivered to the proper receiver, must be processed (recoded) by a fixed number of consecutive MIX-servers.

In some variants of the protocol working in practice, the new message submitted to the first MIX-server is put in the buffer. Then, if there are more than n messages in the buffer, the new message replaces one of the messages randomly chosen from the buffer. The pushed-out message is submitted to the 2^{nd} MIX-server. The processing in consecutive MIX-servers is exactly the same. Finally, a single message leaves the last server and is removed from the system and delivered to the receiver.

Other configurations of MIX-servers are considered in literature (eg. *parallel MIX-cascade* in Klonowski and Kutylowski (2005); Golle and Juels (2004)). It is a matter of investigation how to connect existing MIX-servers to obtain the best security/functionality properties.

Blending attacks *Blending attacks* are based on the following trick: the adversary submits to the MIX-server a number of *fake* messages. Such fake messages (using a special encoding) can be easily recognized by the adversary in the bunch of messages sent by legitimate users. If all but one messages are fake, the adversary can easily trace its route. This form of the attack is called an $(n-1)$-*attack* (see Kesdogan and Pimenidis (2004)). Generally, to make the anonymity weaker, the adversary may submit a fraction of all messages and blend fake messages with real ones (i.e. submitted by regular users) to be traced. Depending on particular protocols and implementations of MIX-es, blending attacks may be conducted in different ways. However, the core of the idea is to *flush* as many real messages as possible to isolate only a small number of target messages and then trace them.

One can easily see that such an attack can be precisely described by the model discussed in our paper. Indeed, MIX-servers are represented by urns filled with white balls (representing messages from legitimate senders). The security problem is how many messages the adversary has to submit to the system to remove significant number of legitimate messages and leave only adversarial messages (represented by black balls).

Up to now, blending type attacks for similar models have been investigated for very important practical settings in several papers, e.g. O'Connor (2005); Serjantov et al. (2002); Dingledine et al. (2006). However, to the best of our knowledge, none of the previous analyses gives as precise and general results as our paper.

Practical consequences of obtained results The presented results convince us that the way the MIX-servers are connected has a significant influence on the immunity against flooding-type attacks. In particular, if we can construct a mix network of MIX-servers with some overall capacities, the parallel structure gives a better immunity than the serial one. That is, the expected time necessary to flush all white messages from the last MIX-server is greater in the parallel structure. This follows directly from theorems 4 and 8. We find this fact quite surprising and counter-intuitive. On the other hand, one can see that the parallel structure offers inferior security against cryptographic attacks. Indeed, each message has to be processed only by a single server instead of some $k \geq 2$.

Another indication for the design of mix network is that if we have a series (cascade) of MIX-servers, they should not be of equal capacity. For example, placing a MIX-server of a bigger capacity at the end

of the cascade can improve the immunity against flooding attacks (see section 6.2).

Acknowledgements

This paper was supported by Polish National Science Center – decision No. DEC-2013/09/B/ST6/02251. The authors would like to thank the anonymous Reviewer for all valuable comments and suggestions to improve the quality of our paper.

References

M. Barni, J. Herrera-Joancomartí, S. Katzenbeisser, and F. Pérez-González, editors. *Information Hiding, 7th International Workshop, IH 2005, Barcelona, Spain, June 6-8, 2005, Revised Selected Papers*, volume 3727 of *Lecture Notes in Computer Science*, 2005. Springer. ISBN 3-540-29039-7.

D. L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun. ACM*, 24(2):84–88, 1981.

G. Danezis and C. Diaz. A survey of anonymous communication channels, 2008. URL `ftp://ftp.research.microsoft.com/pub/tr/TR-2008-35.pdf`.

R. Dingledine, N. Mathewson, and P. F. Syverson. Tor: The second-generation onion router. In *USENIX Security Symposium*, pages 303–320. USENIX, 2004.

R. Dingledine, A. Serjantov, and P. F. Syverson. Blending different latency traffic with alpha-mixing. In G. Danezis and P. Golle, editors, *Privacy Enhancing Technologies*, volume 4258 of *Lecture Notes in Computer Science*, pages 245–257. Springer, 2006. ISBN 3-540-68790-4.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley and Sons Inc., New York, 1965.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

P. Golle and A. Juels. Parallel mixing. In V. Atluri, B. Pfitzmann, and P. D. McDaniel, editors, *ACM Conference on Computer and Communications Security*, pages 220–226. ACM, 2004. ISBN 1-58113-961-6.

D. Kesdogan and L. Pimenidis. The hitting set attack on anonymity protocols. In J. J. Fridrich, editor, *Information Hiding*, volume 3200 of *Lecture Notes in Computer Science*, pages 326–339. Springer, 2004. ISBN 3-540-24207-4.

M. Klonowski and M. Kutylowski. Provable anonymity for networks of mixes. In Barni et al. (2005), pages 26–38. ISBN 3-540-29039-7.

NIST. Digital Library of Mathematical Functions, 2013. URL `http://dlmf.nist.gov/8.17`.

L. O'Connor. On blending attacks for mixes with memory. In Barni et al. (2005), pages 39–52. ISBN 3-540-29039-7.

A. Serjantov, R. Dingledine, and P. F. Syverson. From a trickle to a flood: Active attacks on several mix types. In F. A. P. Petitcolas, editor, *Information Hiding*, volume 2578 of *Lecture Notes in Computer Science*, pages 36–52. Springer, 2002. ISBN 3-540-00421-1.

E. W. Weisstein. BetaRegularized. From mathworld – A Wolfram Web Resource, 2013. URL `http://functions.wolfram.com/GammaBetaErf/BetaRegularized/17/01/01/`.