Louvain School of Management

The Panini collector’s problem: Optimal strategy and trading analysis

Auteur : Olivier Delmarcelle
Promoteur : Frédéric Vrins
Année académique 2018-2019.

Acknowledgments

I would first like to thank my thesis advisor, Professor Frédéric Vrins, for his guidance and help on the problems I faced while writing this paper. He always encouraged me to give the best of myself with constructive remarks. Professor Frédéric Vrins also deserves a second thank for being an amazing teacher that made the classes during my master a place of enlightenment.

I would also like to thank all the professors and assistants that dedicated their time to conduct the exceptional courses given during the five years of my studies. Thanks to all of them, I amassed an important quantity of valued knowledge.

Finally, I must express my very profound gratitude to Soobin Cho and Dominique Kocklenberg for their invaluable support during the preparation of this paper and for their thoughtful inputs regarding the writing. This accomplishment would not have been possible without them. Thank you.

Contents

Glossary and notation

During the development of this paper, I realized that some notations and explanations were confusing for a reader not familiar with my research. I was advised to provide a comprehensive list of the terms and the notation used in this paper. Please feel free to reference to this section whenever you need during your reading.

Coupon A coupon is the most common name in the theory for the smallest component of a collection. For a collector of Panini stickers, coupons and stickers are exactly the same.

Collector A single individual aiming to complete his personal collection of stickers at a minimum cost. Depending on the model, he can collect stickers individually or with the help of other collectors.

Strategy A strategy can be used by a collector to decrease his expected cost while completing his collection. The strategy dictates which source of stickers the collector should choose.

Shift point The shift point is the point at which the collector will be indifferent between the two source of stickers. Before that point, *Packets of random stickers* will be preferred and beyond that point, *Stickers ordered from the supplier* will be favored instead.

Sources of stickers Two possible sources of stickers are considered. *Packets of random stickers* represent the packets usually sold in bookstores or supermarkets containing 5 random stickers inside. *Stickers ordered from the supplier* are specific stickers that collectors can order on the supplier's website at a premium price. Each collector can choose which sticker he wants to order and therefore avoid the randomness of the first option. However, usually only a limited number of stickers can be obtained this way.

Number of unique sticker The number of unique stickers is used to track the state of a collector's collection. It represents the number of different unique stickers the collector owns (i.e., the number of stickers he has without taking into account duplicates). Only once the collectors own all of the unique stickers available, will his collection be considered complete.

Waiting time The *Waiting time* is a random variable representing the number of stickers a collectors has to buy before completing the collection (i.e., if a collector had to buy 5000 stickers to obtain all of the unique stickers available, the waiting time is 5000). The *waiting time* also refers to the number of stickers bought to obtain a subset of all of the available unique stickers. A characteristic that will be largely explored here is the *expected waiting time*.

Collection cost The expected cost for a collector to complete his collection. Depending on the strategy used by the collector, the collection cost composed of the waiting time and the number of *Stickers ordered from the supplier* multiplied by their respective price per stickers.

Swap An exchange of duplicates between two collectors. Each collector exchanges one of his duplicates to obtain a sticker he was missing.

$\boldsymbol{Z_{j,s}^{n,m}}$ This notation refers to the *Waiting time* with the usual assumption used in the literature. Each subscript and superscript refers to a parameter characterising the waiting time.

- s represents the collection size, or the number of unique stickers required to complete the collection (for the World Cup 2018, most collections were composed of 682 unique stickers).
- j represents the stopping time, the collection state at the end of the waiting time (i.e., $Z_{50,682}^{n,m}$ represents a waiting time to obtain 50 unique stickers while $Z_{682,682}^{n,m}$ represents the waiting time to complete the collection).
- m represents the number of stickers contained in *Packets of random stickers*. If m is equal to 1, stickers can be bought randomly one by one. If m is greater than one, it is assumed that each packet does not contain any duplicates. This means that larger packet size will always be profitable to the collectors by reducing their number of duplicates.
- n represents the number of collectors cooperating. It is assumed that collectors are cooperating perfectly (i.e., cooperating collectors are buying stickers as a group and each individual collection will be completed at the end of the process when each unique sticker is collected n times.)

$\boldsymbol{C_{j,s}^{n,m,k}}$ A variant of $Z_{j,s}^{n,m}$ under different assumptions for cooperation between collectors. Therefore, $Z_{j,s}^{1,m}$ and $C_{j,s}^{1,m,k}$ are equivalent since there is no cooperation possible. The cooperation in this situation is comparable to a network in the shape of a circle. In this network, each collector swaps stickers with only two other collectors. The parameter k represents the number of stickers bought before collectors start to swap with each other. This model is detailed in section 4.2.1.

$\boldsymbol{S_{j,s}^{n,m,k}}$ A variant of $Z_{j,s}^{n,m}$ under different assumptions for cooperation between collectors. Therefore, $Z_{j,s}^{1,m}$ and $S_{j,s}^{1,m,k}$ are equivalent since there is no cooperation possible. The cooperation in this situation is comparable to a network in the shape of a star. In this network, each collector swaps stickers with all other collectors. The parameter k represents the number of stickers bought before collectors start to swap with each other. This model is detailed in section 4.2.2.

1 Introduction

Every four years, the FIFA World Cup takes place and raises interest in both football matches and the associated stickers collections. The collection of stickers during the FIFA World Cup is a long-standing tradition. Though usually a pastime reserved for children, alongside the collection of cards or other toys, collecting memorabilia in the shape of football players stickers has also garnered interest of the adults as well.

Since 1970, the Panini company has been producing albums and stickers collections for every World Cup. The albums are essentially books with the World Cup theme filled with numbered "holes" waiting to be filled by the corresponding stickers. The stickers represent pictures of teams and players participating in the World Cup. Theses stickers are sold separately inside small packets of five stickers, and their contents are randomized. The goal for anyone possessing an album is to collect every different sticker available to complete the collection. The act of completing the album is usually a lot of fun for football fans.

Thus far, this business model has been very lucrative for Panini, as collectors have to buy between two and ten times more stickers than the size of the collection due to the large number of duplicates obtained. Collectors rarely pay attention to the cost of buying stickers before starting their personal collection, and once a collection is started collectors generally buy stickers until completion regardless of the cost involved. One of the ways to surmount the randomized sticker collection is to order specific stickers from Panini directly. Although Panini restricts the number of these sales, it helps unlucky collectors or collectors unable to trade with others to complete their collections without spending a large amount of money.

The huge popularity of these collections can be explained by three factors. First of all, since collections include a random collection of stickers every collector enjoys the thrill of opening new packs of stickers hoping to obtain a missing sticker. Secondly, trading stickers is the source of numerous social interactions since collectors either meet face-to-face or in meeting events to exchanges duplicates and acquire missing stickers. Thirdly, the act of collecting stickers follows the trend of the World Cup competition and feeds into the game frenzy.

The main objective of this thesis is to critically analyse the potential case of a collector intending to complete his 2018 World Cup collection, containing 682 unique stickers. This includes estimating the expected cost for the completion of the collection and determining an optimal strategy for the collector. Consideration will be paid to both the cases of a single collector and that of a group of collectors. In addition, the interests of the Panini company will also be critically analysed. Included in this analysis will be an examination of the available profit for the supplier related to the size of the collections produced, random packets size, and the number of stickers available for direct order. Consideration will also be given to the profitability of a online trading-platform where collectors may exchange stickers.

In order to reach this objective, this thesis will explore relevant literature such as the *Coupon Collector Problem*, a very similar problem where one collector intends to complete a collection of s coupons by obtaining coupons randomly one at a time. Results have been discovered regarding expectation, variance and distribution of the *Coupon Collector Problem* and its generalizations in the literature and will be used as a basis for the subsequent analysis. As the literature alone is inconclusive for some generalization of the problem, it will be bolstered through the use of Monte Carlo Algorithms. These kinds of algorithms replicate a random event a large number of times (commonly more than 10^5 times) to determine its distribution. A large portion of the analysis of this thesis will rely on Monte Carlo results.

2 Literature review

A collector intending to complete his 2018 World Cup collection will experience a problem similar to the *Coupon Collector Problem* (CCP). Indeed, the completion of a collection of stickers can be exactly described with a generalized form of the CCP. In the most basic formulation of CCP, readily usable formulas have been found to compute the characteristics of a CCP. However, the most basic formulation does not exactly replicate the 2018 World Cup collection and is insufficient to perform proper analyses. It must be extended to take into account *Cooperation* and *Group drawing*. On one hand, the *Cooperation* generalizes the CCP for various number of collectors cooperating in the collection process. More specifically, 2 collectors cooperating would buy stickers together and allocate the stickers among them to complete two collections as fast as possible. On the other hand, the *Group drawing* refers to multiple stickers obtained at once. The classical CCP problem only considers the intake of new stickers/coupons one by one. Obtaining more than one stickers at once alters the probabilities of newly obtained stickers, hence why the CCP was generalised for intakes of more than one sticker. Both extensions are necessary to analyse the 2018 World Cup collection accurately and this will be explained in the subsequent section. In addition, the concept of an *Optimal strategy* will also be introduced at the end of this section.

An assumption that will be maintained throughout this paper is that each coupon has the same chance to be drawn (i.e, coupons are drawn from an uniform distribution). It is worth mentioning that although CCP with unequal probabilities are thoroughly examined within the literature but their mention will be omitted for the purposes of this thesis as it is outside the scope of the problem analysed here.

2.1 Basic coupon collector problem

The Coupon Collector Problem can be historically traced as far back as 1712, when de Moivre computed the odds to obtain a specific set of faces on a dice [Hald, de Moivre, and McClintock, 1984]. One century later, Laplace formulated the problem as a lottery and searched for the probability of obtaining all lottery stickers after a given number of trials [Laplace, 1878, pp. 5–24]. More recently, Erd and Ri presented an asymptotic formula characterising the distribution of the *waiting time* - the number of coupons required to complete the collection [Erd and Ri, 1961].

Formally, the basic form of the CCP can be presented as follows : A set S comprising a number of unique coupons labelled $\{1, 2, 3, ...s\}$ is defined. From this set, one random coupon is drawn with the probability of $\frac{1}{s}$ as all of the coupons have an equal chance to be drawn. The selected coupon is then replaced in the set before another draw occurs. The process ends once every unique coupon has been drawn at least once. Then, the number of draws that was required to reach this point is referred as the *waiting time*.

The *waiting time* will be presented as $Z_{j,s}^{n,m}$ throughout this paper. Parameters n and m are related to more complex version of CCP and are equal to 1 for the most basic CCP.

Different results have been discovered about this basic form. Some are presented here :

1. The probability to obtain a new unique coupon after l unique coupons have already been

drawn, l ranging from 0 to $s-1$.

2. The expected waiting time to obtain a number j of unique coupons, j ranging from 1 to s (i.e. $\mathbb{E}(Z^{1,1}_{50,682})$ when $j=50$ and $s=682$, the number of draws required to obtain 50 unique stickers among the 682 available).

The first result is relatively straightforward since after having drawn l unique coupons, the chance to obtain a new unique coupon is equal to the number of not-drawn-yet unique coupons divided by the number of available unique coupons : $\frac{s-l}{s}$.

The second result has been solved as follows; since the expected waiting time to obtain j unique coupons is the sum of the expected number of draws to obtain each unique coupon one by one up to j. For unique coupon 1, the expected number of draws will be 1 as the probability to obtain a new unique coupon is one. Obtaining a second unique coupon shows a probability of $\frac{s-1}{s}$ per draw. Therefore, the expected number of draws to obtain the second unique coupon is $\frac{s}{s-1}$. For the third, the expected number of draws is $\frac{1}{s-2}$ and for the subsequent coupons up to coupon l : $\frac{s}{s-l}$. When $j=s$, the expected waiting time can be expressed as :

$$\mathbb{E}(Z^{1,1}_{s,s}) = s\left(\frac{1}{s}+\frac{1}{s-1}+...+\frac{1}{1}\right) = sH_s \tag{1}$$

since the series $(\frac{1}{n}+\frac{1}{n-1}+...+\frac{1}{1})$ is also known as n-th Harmonic number H_n, the sum of the reciprocals of the first n natural numbers. H_n can be found either using the recursive relationship $H_{n+1} = H_n + \frac{1}{n+1}$ or with the development

$$H_n = ln(n)+\gamma+\frac{1}{2n}-\sum_{k=1}^{\infty}\frac{B_{2k}}{2kn^{2k}} = ln(n)+\gamma+\frac{1}{2n}-\frac{1}{12n^2}+\frac{1}{120n^4}-\frac{1}{252n^6}+... \tag{2}$$

Where $\gamma =$ is the Euler–Mascheroni constant and B_{2k} are the even Bernouilli numbers.

When s is large, H_n can be approximated with $ln(n)+\gamma+\frac{1}{2n}$ and it is possible to write :

$$sH_s = sln(s)+\gamma s+\frac{1}{2}+O(1/s)$$

The distribution of $Z^{n,1}_{s,s}$ has been asymptotically determined by Erd and Ri [Erd and Ri, 1961], which can be simplified for the basic formulation of CCP as :

$$\lim_{s\to\infty} P\left(\frac{Z^{1,1}_{s,s}-s\,ln(s)}{s}<x\right) = \exp\left(-e^{-x}\right) \tag{3}$$

A non-asymptotic distribution has been provided by Stadjes for $Z^{1;m}_{s,s}$ [Stadje, 1990], which shorten for the basic CCP to :

$$P(Z^{1,1}_{s,s}=k) = \sum_{j=0}^{s-1}(-1)^{s-j+1}\binom{s}{j}\frac{s-j}{s}\left(\frac{j}{s}\right)^{k-1} \tag{4}$$

Since there are 682 unique stickers in the 2018 World Cup collection, and this is a rather large number, the asymptotic distribution exposed by Erd and Ri provides a very accurate result with no

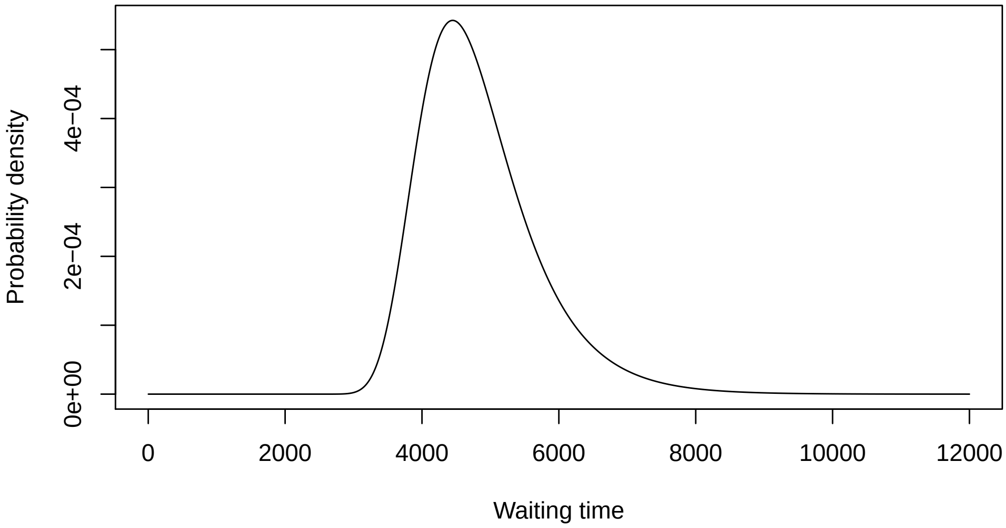

Figure 1: Probability Density Function of $Z^{1,1}_{682,682}$

significant difference from the non-asymptotic distribution developed by Stadjes. Figure 1 shows a probability density function using equation (4).

The variance can be computed from either the asymptotic of the non-asymptotic distribution. Without using the distribution, Doumas formulated a convenient formula to compute the variance directly [Doumas, 2015] :

$$V(Z^{1,1}_{s,s}) = \frac{\pi^2}{6}s^2 - s\,ln(s) - (\gamma+1)s + O\left(\frac{ln(s)}{s}\right) \tag{5}$$

2.2 Cooperation extension

The basic version of CCP only accounts for a single collection, meaning there are no swaps or exchanges considered. However, it is possible for collectors to swap stickers with other collectors, thereby significantly reducing their waiting time. To model the possibility of swaps between collectors, the literature relies on the assumption of a perfect cooperation between collectors :
Collectors tend to cooperate within a group to allocate coupons in an optimal way by sharing costs until every collector finishes his collection.
Using this assumption, computing the waiting time for a group of n collectors simply reduces to finding the number of draws required to obtain n times all available unique coupons. Newman solved this problem for multiple collections and provided a formula for the total expected waiting time [Newman, 1960] :

$$\mathbb{E}(nZ^{n,1}_{s,s}) = s\int_0^\infty \big(1-(1-S_n(t)e^{-t})^n\big)dt \tag{6}$$

Where n represents the number of collectors cooperating in the group and $Z^{n,1}_{s,s}$ represents the

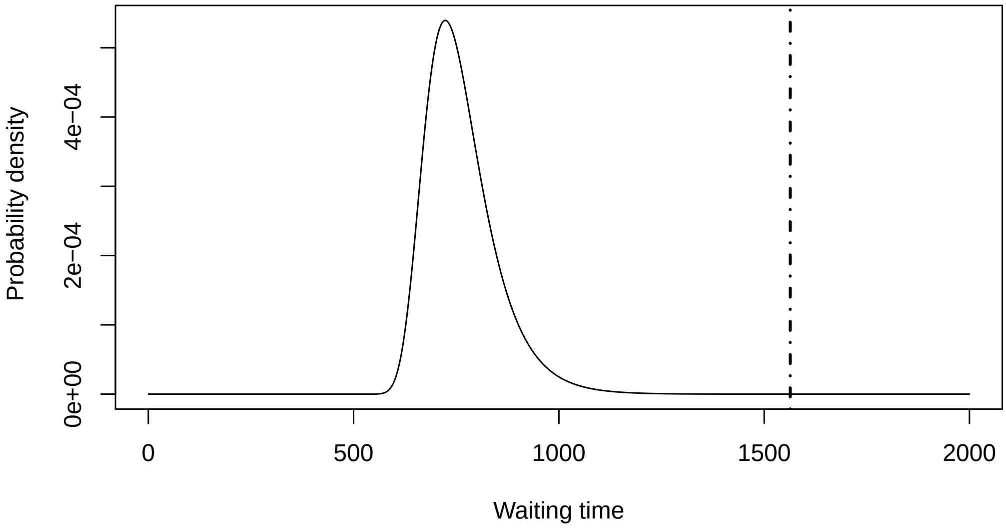

Figure 2: Probability Density Function of $Z^{10,1}_{682,682}$.
The dotted line represents the expected waiting time, computed with equation (6).

waiting time for a single collector within the group. $S_n(t)$ is defined as :

$$S_n(t) = \sum_{j=0}^{n-1} \frac{t^j}{j!}$$

When n is large and for fixed m, the expected waiting time can be approximated to the asymptotic development :

$$\mathbb{E}(nZ^{n,1}_{s,s}) = s\, ln(s) + s(n-1)\, ln(\, ln(s)\,) + s\, C_n + O(n)$$

Where C_n is a constant value depending on n. Erd and Ri found the following expression to obtain its values [Erd and Ri, 1961]: $C_n = \gamma - ln((n-1)!)$

Erd and Ri further determined the asymptotic distribution of $nZ^{n,1}_{s,s}$:

$$\lim_{s\to\infty} P\left(\frac{nZ^n_s}{s} - ln(s) - (n-1)\, ln(\, ln(s)\,) < x\right) = \exp\left(-\frac{e^{-x}}{(n-1)!}\right) \tag{7}$$

from which we derive the probability density function of $Z^{10,1}_{682,682}$ in figure 2.

From figure 2, a large difference can be seen between the asymptotic distribution from equation (7) (continuous line) and the expected waiting time computed using the integral in equation (6) (dotted line). This difference is a consequence of the asymptotic nature of the analytical distribution developed by Erd and Ri. In fact, the number of collectors n drastically affects the accuracy of the asymptotic distribution.

Figure 3 represents a comparison between the expected waiting time from the asymptotic equation (7) and the non-asymptotic result coming from equation (6). Continuous lines represent the exact values while dotted lines represent the asymptotic development. Each colour denotes a number of cooperating collectors for which one continuous line and one dotted line have been computed. For $n = 1$, which describes the basic CCP studied previously, continuous and dotted lines are overlapping, indicating that the asymptotic result provides a very accurate approximation. However, the

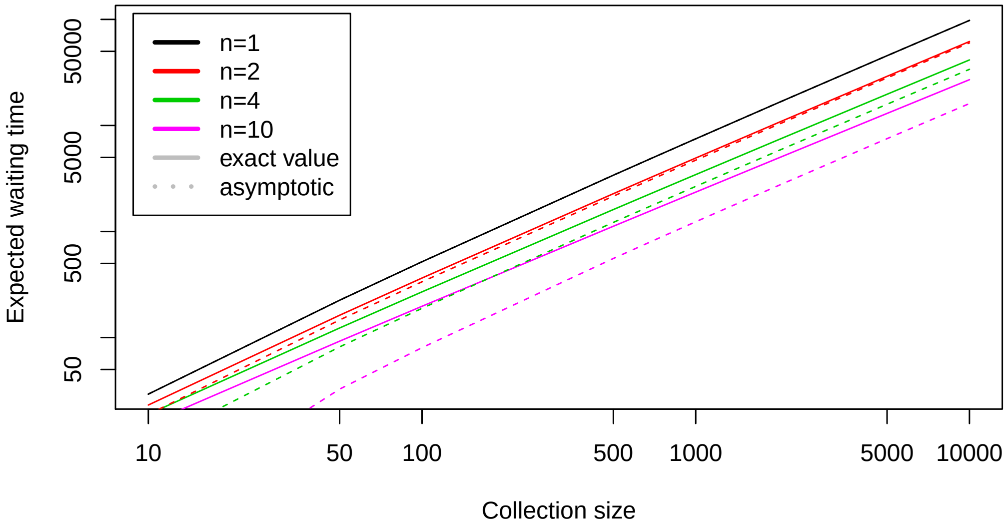

Figure 3: Comparison between the exact value and asymptotic development of the waiting time for various number of collectors.

story changes for values of $n > 1$. When $n = 2$, the asymptotic result starts to underestimate the exact value. Increasing s, the collection size, lowers the error but a large value of s of more than 5000 is required for both results (asymptotic and non-asymptotic) to coincide. For $n = 4, 10$ the gap becomes larger and more significant even at large value of s.

Clearly, $s = 682$ is nowhere large enough to be approximated by the asymptotic results with a value of n greater than 1. Besides asymptotic development, the current literature lacks methods to obtain the distribution of the waiting time for cooperating collectors. The distribution can however be computed numerically, using *Monte Carlo Algorithms*, to be discussed in section 3. Despite this gap in the literature, the results coming from equation (6) remain however correct and will be used to verify the validity of the numerical approximation.

2.3 Group drawing extension

The basic CCP relies on one important assumption regarding draws of coupons. The coupons are drawn one at a time and immediately replaced in the set of coupons after the draw has been recorded. Draws of multiple coupons at once before replacing them in the set are not considered.

The contribution of Stadjes [Stadje, 1990] provided an analysis of the CCP including draws of multiple coupons at once. The group drawing approach is similar to the collections produced by Panini since Panini stickers are usually sold per group of 5.

Stadjes exposed an analytical solution for the distribution of the waiting time as a function of the number of coupons per draw m:

$$P(Z_{s,s}^{1,m} = k) = \sum_{j=0}^{s-1} (-1)^{s-j+1} \binom{s}{j} \frac{\binom{s}{m} - \binom{j}{m}}{\binom{s}{m}} \left(\frac{\binom{j}{m}}{\binom{s}{m}} \right)^{k-1} \tag{8}$$

Where k represents the number of draws performed (or equivalently, the waiting time). Since the shape of the distribution is similar to the basic CCP presented in figure 1, the figure for this extension is omitted.

Stadjes also disclosed a formula to compute the moments of the distribution, from which equations for expected waiting time and variance can be derived:

$$\mathbb{E}(Z_{s,s}^{1,m}) = \binom{s}{m} \sum_{j=0}^{s-1} (-1)^{s-j+1} \frac{\binom{s}{j}}{\binom{s}{m} - \binom{j}{m}} \tag{9}$$

$$V(Z_{s,s}^{1,m}) = \binom{s}{m} \sum_{j=0}^{s-1} (-1)^{s-j+1} \binom{s}{j} \binom{j}{m} \left[\binom{s}{m} - \binom{j}{m} \right]^{-2} \tag{10}$$

The impact of the parameter m on the waiting time remains little for small values of m. This will more deeply analysed in section 3.3.3, where a sensitivity analysis on m will be performed.

2.4 Optimal strategy

Drawing coupons is not the only way to collect coupons. In some cases, it is possible for collectors to purchase specific coupons directly from the supplier. However this comes at a premium price compared to the price of random stickers inside packets. Due to having multiple sources of coupons available, the collector may choose to obtain a certain number of coupons using random draws and then decide to buy coupons directly from the supplier afterwards. This behaviour represents the *Strategy* applied by the collector. The point at which the collector switches from one source of coupons to the other is called *Shift point*. An *Optimal strategy* is the strategy applied at the shift point that minimizes the expected cost for the collector to complete of his collection.

Stadjes discussed the optimal strategy for the CPP including group drawing [Stadje, 1990]. He made use of the price for each source of stickers, using α as the price for a packet of coupons (containing m coupons) and β as the price to buy one specific coupon directly from the supplier. He defines the optimal shift point j^*, corresponding to the number of unique coupons obtained, at which a collector should stop buying packets of coupons and instead buy specific coupons directly from the supplier.

The general expression for the expected cost of a strategy is the following :

$$c(j) = \alpha\mathbb{E}(Z_{j,s}^{1,m}) + \beta(s-j) \qquad \text{with } j = 1, 2, ..., s \tag{11}$$

Where $\mathbb{E}(Z_{j,s}^{1,m})$ represents the expected waiting time to obtain j unique coupons within the number of available unique coupons s. $c(j)$ is then minimized for the optimal value of j, j^*. The optimal strategy minimizing the cost will be denoted as c^*.

When $m = 1$, the problem is reduced to the basic CCP as defined in section 2.1 and j^* can be computed explicitly as $[s+1-\beta^{-1}\alpha s]$, where [.] refers to the integer part function. The minimum cost using the optimal strategy is given by :

$$c^* = \alpha s \sum_{k=0}^{[s-\beta^{-1}\alpha s]} (s-k)^{-1} + \beta(s - [s+1-\beta^{-1}\alpha s])$$

When $m > 1$, j^* has to be found by computing $c(j)$ for all of the values of j and finding the one that minimizes c. Stadjes presented the following development to compute c with the values of m greater than one :

$$c(j) = \alpha \binom{s}{m} \sum_{i=0}^{j-1} (-1)^{j-i+1} \frac{\binom{s}{i}\binom{s-i-1}{s-j}}{\binom{s}{m} - \binom{i}{m}} + \beta(s-j) \qquad \text{with } j = 1, 2, ..., s \tag{12}$$

And once $j*$ is found, the minimum expected cost is equal to :

$$c^* = \alpha \binom{s}{m} \sum_{i=0}^{j^*-1} (-1)^{j^*-i+1} \frac{\binom{s}{i}\binom{s-i-1}{s-j^*}}{\binom{s}{m} - \binom{i}{m}} + \beta(s-j^*)$$

It is also possible to rewrite these equations in order to get rid of the parameter α, the price of one packet of coupons. If we define $\theta = \frac{\beta}{\alpha}$, the premium ratio, and $v(j) = \frac{c(j)}{\alpha}$, the standardised expected cost, we can reformulate equation (11) to :

$$v(j) = \mathbb{E}(Z_{j,s}^{1,m}) + \theta(s-j) \tag{13}$$

While equation (12) can be rewritten as :

$$v(j) = \binom{s}{m} \sum_{i=0}^{j-1} (-1)^{j-i+1} \frac{\binom{s}{i}\binom{s-i-1}{s-j}}{\binom{s}{m} - \binom{i}{m}} + \theta(s-j) \tag{14}$$

From the strategic perspective, equations (13) and (14) are easier to analyse. While $v(j)$ does not present the actual expected cost (although it is easily obtained as $c(j) = \alpha v(j)$), it has the advantage of using the ratio between β and α as unique parameter. Because $v(j)$ and $c(j)$ are both minimized for the same j^*, there is no need to obtain $c(j)$ to find out the optimal strategy. The value of j^* is actually influenced only by the ratio between the price to buy one coupon directly from the supplier and the price to buy one packet of coupons, which is precisely θ. Therefore, finding j^* using equations (13) and (14) results in the optimal switch point for multiples values of β and α - as long as their ratio is equal to θ.

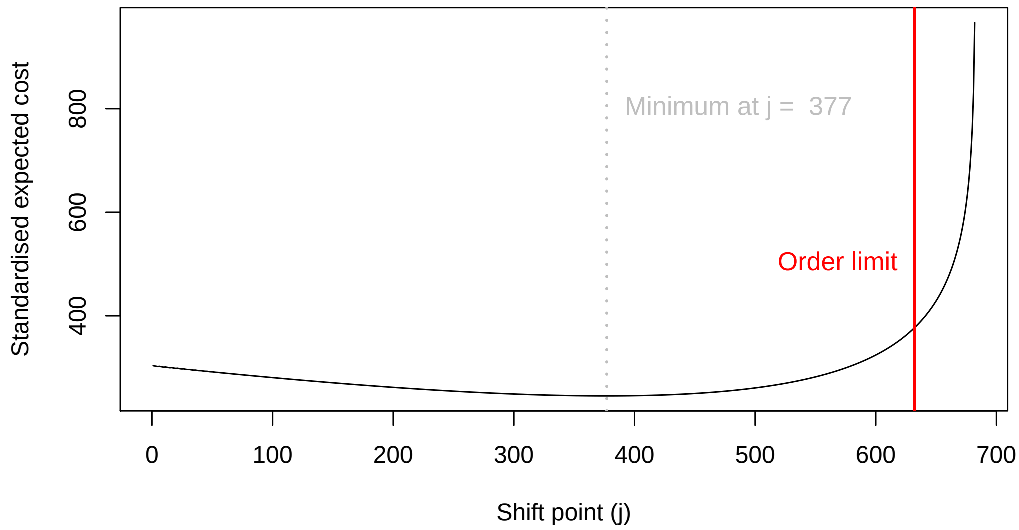

Figure 4: Standardised expected cost from equation (14), with $m = 5$, $s = 682$ and $\theta = \frac{4}{9}$ as parameters.

Figure 4 presents the standardised costs for a collector collecting the 2018 World Cup Panini album as a function of the shift point applied in his strategy. The ratio θ has been computed from the prices exposed on Panini's website for both sources of stickers ($\alpha = 0.9$ and $\beta = 0.4$). It can be seen that a minimum is reached for a shift point of 377, meaning that the optimal strategy for a collector would be to collect 377 stickers randomly from packets, then order the remaining 305 missing stickers from the supplier at the premium price. The resulting standardised expected cost v^* is 245.27 and the corresponding expected cost c^* of 220.75, are significantly less than the standardised expected cost without applying any strategy (shift point equal to ∞) of 966.4, and the corresponding expected cost of 869.76.

It is important to discuss the fact that the number of stickers that one collector can order from the supplier is often limited, and the 2018 World Cup collection case does not make an exception to this rule. A maximum of 50 stickers per collectors may be obtained through this means. In this case, the collector cannot apply the optimal strategy if the *Order limit* (the size of the collection minus the maximal number of stickers a collector may order) is greater than j^*. Then, because $c(j)$ and $v(j)$ are always convex, the first usable j greater than j^* will becomes optimal, leading the *Order limit* turn into the new optimal shift point. Under the *Order limit* constraint, the optimal strategy for a collector of the 2018 Word Cup collection changes the shift point to 632, for a minimum standardised expected cost of 376.73, and an expected cost of 339.06.

3 Numerical approach

As observed in the exploration of the literature, numerous results exist to easily compute interesting characteristics of a Coupon Collector Problem and its extensions. However, some results have been found only asymptotically and have produced a very poor approximation for a normal-sized collection size, thereby limiting their usefulness. Furthermore, results have been discovered for two extension of the CCP, the *Cooperation* and *Group drawing* extensions, but there are no results for a CCP using both extensions at once.

To cope with the limits of the existing literature and the lack of analytic results for a CCP including the two mentioned extensions, one alternative is to use a numerical approach, by the means of Monte Carlo algorithms.

Monte Carlo is a broad term generally used when random distributions are used to solve a problem. Numerous instances of the problem are simulated using samples of a known initial distribution. The initial distribution is then altered by multiple transformations (that are sometimes difficult to analyse analytically) and returns an output distribution. If done properly, the output distribution is a good approximation for the exact distribution resulting for the problem.

Using Monte Carlo algorithms, we will be able replicate the randomisation of the contents of a packet of stickers multiple times in order to obtain the waiting time required to complete a collection. This procedure will be repeated many times to obtain a distribution of the waiting time. The outputs of Monte Carlo algorithms will then be compared to analytical results coming from the literature, then extended to a CCP using both *Cooperation* and *Group drawing* simultaneously. Finally, this most complex version of the CCP will be used to determine and analyse an optimal strategy for cooperating collectors.

3.1 Monte Carlo modelling and convergence

To replicate the process of a collector buying packets of stickers and to obtain the waiting time, we developed an algorithm using the language R and Rcpp (integration of C++ within R). The reasoning behind the use of a C++ integration is to boost the efficiency of the algorithm. The precision of a Monte Carlo algorithm depends foremost on the number of times the algorithm was executed. and the use of a lower level language allows it to be executed a larger number of times than a higher language such as R or Python for the same computing time. For this reason, the majority of the Monte Carlo algorithms used in the paper were coded using Rcpp, improving the algorithm speed by up to a factor of 20.

Strictly speaking, the behaviour of the algorithm is replicating a collector starting with an empty collection and buying packet of m stickers one by one until he obtains n units of each sticker. Each packet is sampled without replacement from the entire pool of stickers $1, ..., s$. The algorithm tracks the number of packets that were bought once the collector completed his collection(s). Each execution of the algorithm, also called simulation, provides a single waiting time, and repeating the process more than ten thousand times starts to depict an accurate distribution for the waiting time.

Similarly to the previous section, the waiting time computed with Monte Carlo will also be presented as $Z^{n,m}_{j,s}$, where n, m, s, j are the parameters affecting the Monte Carlo algorithm. n is the number of collection to be completed (i.e., the number of collectors cooperating), m is the number of stickers in each packet, s is the collection size (i.e., the number of unique stickers available) and j is the stopping time (i.e., the number of unique stickers obtained at which the algorithm stops).

Since the accuracy of a Monte Carlo largely depends on the number of simulations, a strong commitment in the development of this paper was to use as large a number of simulations as possible to ensure proper and readily usable results. Various papers using the Monte Carlo recommend computing the number of simulations required to obtain a desired precision; however, this step has been omitted here as it would have not changed the methodology. As access to very powerful computers was limited while writing this paper, computation using more than 10^7 simulations was impossible since they would take more than days to compute and the time available was limited. Despite this, the accuracy of the results of the algorithm remain satisfactory, especially for the results related to the expected waiting time.

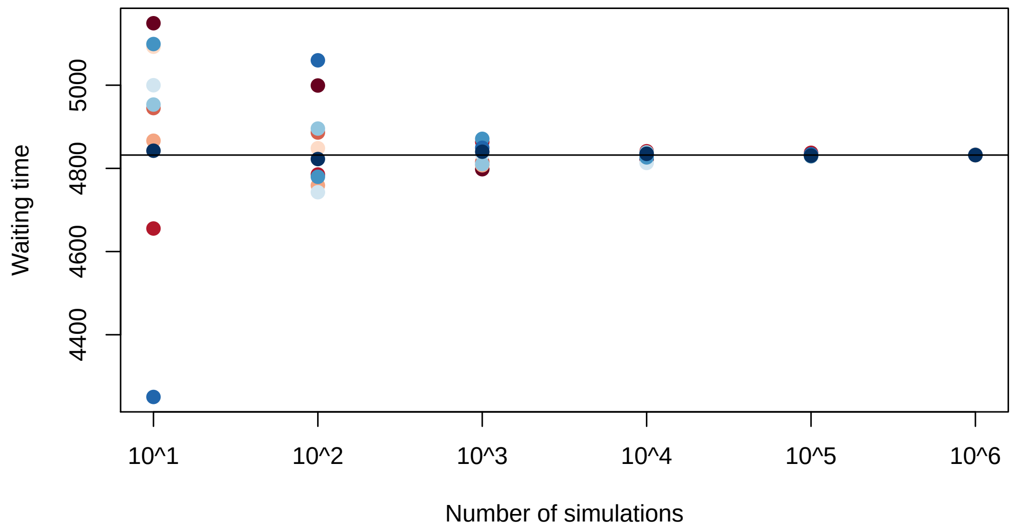

Figure 5: An example of convergence to the analytical value for Monte Carlo algorithms. The horizontal line represents the waiting time computed using analytical results from the literature.

Figure 5 depicts the increase in accuracy as the number of simulations rises. For a varied number of simulations from 10^2 to 10^6, the mean coming out from the Monte Carlo algorithm was computed ten times. Each point on the graph represents one output coming from Monte Carlo. It can be seen that the outputs of Monte Carlo are largely dispersed when using 10^2 simulations. As the number of simulations increases, the dispersion lowers and the points approach the horizontal line (which represents the analytical expected waiting time) much more consistently. With 10^6 simulations, the points overlap on the horizontal line, indicating a very accurate approximation. All Monte Carlo results exposed in this paper have been re-computed multiple times to ensure they would provide an accurate estimation. While some characteristics such as the variance are more difficult to estimate and would require a larger number of computations to be accurate at the unit point, we still believe that the results presented here are sufficient for the purpose of our analysis.

3.2 Outputs from Monte Carlo algorithms

The outputs from Monte Carlo algorithms will be shown in a comparable way to the structure of section 2. For the basic CCP and for both extensions viewed in the literature, a comparison between the analytical results and the Monte Carlo output will be performed. This will test if the Monte Carlo algorithms produce accurate approximations. Thereafter, the case unexplored in the literature of CCP with *Cooperation* and *Group drawing* simultaneously will be investigated. Finally, we will extend the strategy developed in section 2.4 with the *Cooperation* extension and examine other types of strategies.

3.2.1 Basic coupon collector problem

Using our algorithm with parameters $s = 682$ and $m, n = 1$, we obtained the resulting outputs displayed in figure 6 and table 1 and compared them with the analytical results.

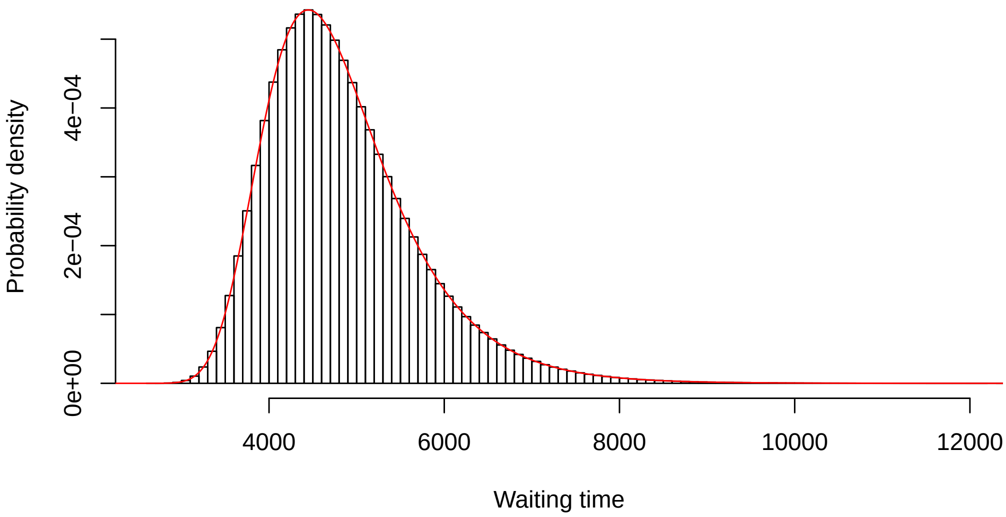

Figure 6: Histogram of simulated waiting times $Z^{1,1}_{682,682}$. Number of simulations : 10^7. The red line represents the analytical density computed with equation (4).

Table 1: Comparison between Analytical and Monte Carlo results for $Z^{1,1}_{682,682}$

	Analytical	Monte Carlo	Difference%
Mean	4844.231	4844.213	0.0004 %
Standard deviation	871.53	871.07	0.05 %

From figure 6 it can be seen that our algorithm replicates the basic CCP correctly. The analytical distribution coming from equation (4), the red line, perfectly fits the shape formed by the histogram of Monte Carlo outputs. From these outputs are directly computed simulated mean and standard deviation and presented in table 1. Simulated mean and variance are compared with the analytical results coming from equations (1) and (5). The column Difference% simply displays the observed difference as a percentage between Monte Carlo and the expected result. It can be observed that the Monte Carlo algorithm successfully approximates the analytical distribution of the waiting time, showing less than 1% difference between analytical and numerical results. It may be noted, however, that the standard deviation appears to converge slower than the mean to the expected values.

The presented distribution of $Z^{1,1}_{682,682}$ appears in a shape similar to a log normal distribution. The left tail of the distribution is small and short, leading to a sharp increase in the probability density as the waiting time approaches 4000. The area between 4000 and 6000 represents the biggest area of the probability density function, accounting for 76% of its surface. Beyond that point is the right tail, where the probability decreases at a slow rate. Though only 1% of the Monte Carlo outputs produced a waiting time above 7500, these outputs can present a waiting time much greater than this point. Consequently, the highest observed waiting time from the Monte Carlo algorithm was 14991.

3.2.2 Cooperation extension

Developing a Monte Carlo algorithm for the *Cooperation* extension only required changing the number of each unique sticker to be collected compared to the basic CCP: a group of n collectors cooperating targets to collect n times each unique sticker available. The case of 10 collectors was analysed here and although $n = 10$ was chosen rather arbitrarily, we believe most group of cooperating collectors in a real-life situation would have a size comparable to it. Since it is complicated to organize and assign stickers within a cooperating group, one would question the stability of large groups of cooperating collectors. Increasing n has a direct impact on the computing time : the end of the algorithm for each simulation will be reached after n collections, unlike the previous basic CCP algorithm only reaching 1 collection.

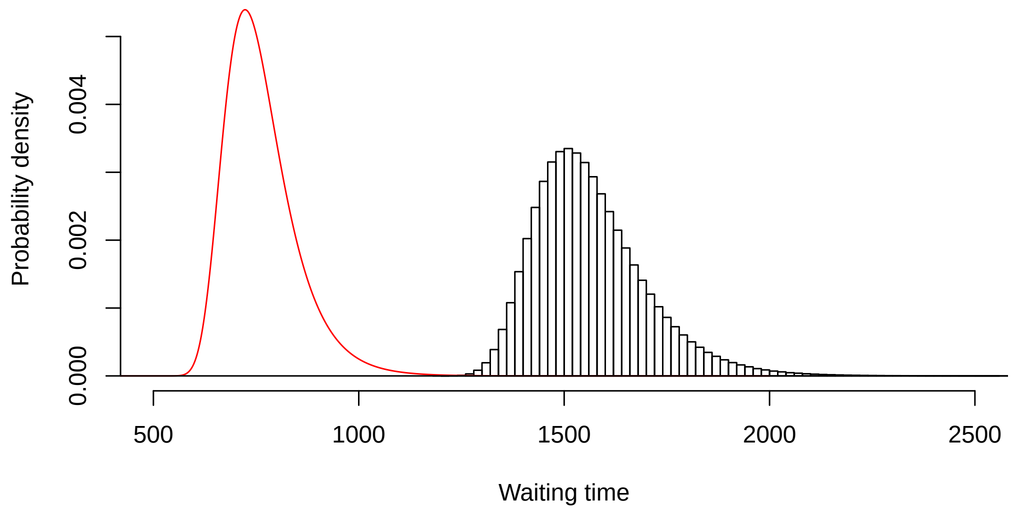

Figure 7: Histogram of simulated waiting times $Z^{10,1}_{682,682}$. Number of simulations : 10^7. The red line represents the asymptotic analytical density computed with equation (7).

Figure 7 and table 2 show the Monte Carlo outputs for $Z^{10,1}_{682,682}$. A large difference can be observed between the asymptotic distribution coming from equation (7) and the Monte Carlo outputs. This is a consequence of the asymptotic nature of the results developed in the literature. As it was noted in section 2.2, the accuracy of the approximation of $Z^{n,1}_{s,s}$ by an asymptotic formula largely depends on the parameter n. Hence, the analytical distribution strongly understates the actual behaviour of $Z^{10,1}_{682,682}$.

Table 2: Comparison between Analytical and Monte Carlo results for $Z^{10,1}_{682,682}$

	Analytical	Monte Carlo	Difference%	$Z^{1,1}_{682,682}$
Mean	1563.37	1563.33	0.002 %	4844.231
Standard deviation		135.54		871.53

As represented on table 2, only the mean of the Monte Carlo outputs could be compared to the analytical expected waiting time computed with equation (6). For the mean at least, the analytical results and Monte Carlo outputs present a very small difference of 0.002 %. Albeit it is not possible to verify the accuracy of the distribution obtained from the Monte Carlo outputs, the good approximation of the mean leads us to think that the Monte Carlo distribution is also accurate.

When it comes to the shape of the distribution, it is similar to the shape of the basic CCP. Changing the n parameter does not affect the general shape of the distribution but it rather alters the width and the position of the distribution. Table 2 also presents a quick comparison of the mean and the standard deviation between $Z^{10,1}_{682,682}$ and $Z^{1,1}_{682,682}$. It can be observed that the increase of n from one to ten drastically impacts both the mean and standard deviation of the waiting time. The expected waiting time is approximately divided by three following the change, while the standard deviation is reduced to only one sixth of the basic CCP value. The effect of *Cooperation* is strongly noticeable, especially for the standard deviation, because splitting the cost allows the cooperating collectors to minimize their exposure to unlucky situations.

3.2.3 Group drawing extension

For the *Group drawing* extension, the Monte Carlo algorithm is similar to the basic CCP; however it includes a change in the process of random drawing. Instead of drawing one sticker from the set of available unique stickers (i.e, equivalent to a sampling with or without replacement), five stickers are drawn at once from the set (this time, equivalent to sampling without replacement). This change has a direct effect on the computing time, since using a sampling algorithm without replacement is usually slower than algorithm with replacement. Conforming to the case of the 2018 World Cup collection, a packet of stickers contains five random stickers, fixing the value of m to 5.

Both figure 8 and table 3 present the outputs of the Monte Carlo algorithm. The distribution appearing in figure 8 is almost identical to the distribution for the basic CCP displayed in figure 6. Once again, the analytical distribution perfectly outlines the Monte Carlo histogram, indicating an accurate approximation. A further comparison is exposed in table 3, where only a little difference can be seen between the mean and the standard deviation of Monte Carlo outputs with the expected waiting time and the standard deviation coming from equations (9) and (10). A comparison with the basic CCP is also provided. Unlike the *Cooperation* extension, the increase of m is almost unnoticeable, resulting in an expected waiting time only about 0.25 % lower compared with the basic CCP.

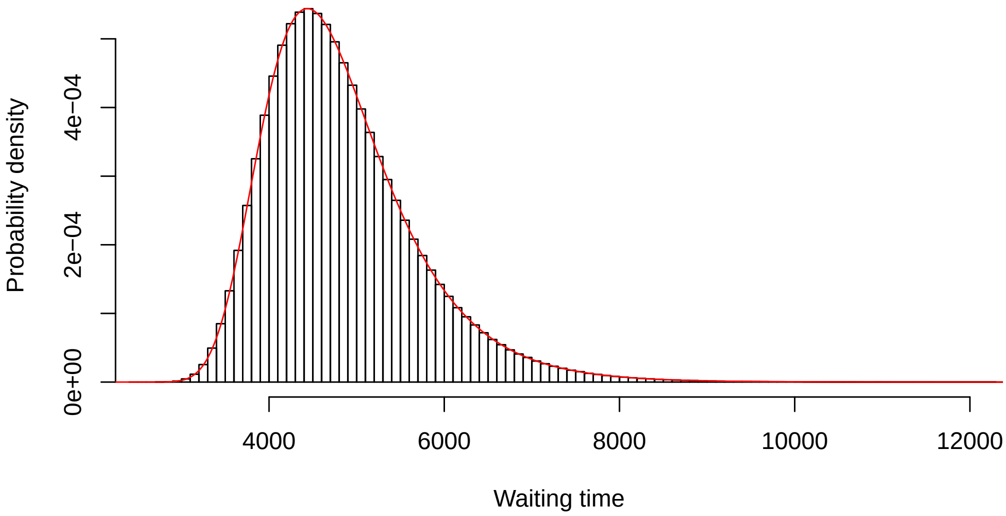

Figure 8: Histogram of simulated waiting times $Z^{1,5}_{682,682}$.Number of simulations : 10^7. The red line represents the analytical density computed with equation (8).

Table 3: Comparison between Analytical and Monte Carlo results for $Z^{1,5}_{682,682}$

	Analytical	Monte Carlo	Difference%	$Z^{1,1}_{682,682}$
Mean	4832.004	4832.506	0.01 %	4844.231
Standard deviation	868.96	868.66	0.04 %	871.53

3.2.4 Cooperation and Group drawing combined

The main objective of the numerical analysis was to obtain the distribution of the waiting time for a CCP characterised by *Cooperation* and *Group drawing* simultaneously. As this case have been unexplored by the literature, so far no analytical results exist. Given that the results for the *Cooperation* extension are mostly asymptotic and unreliable, it is unlikely that this characterisation of the CCP might be easily solved using analytical tools. Therefore, a numerical approach using Monte Carlo algorithm can serve as the most convenient method to obtain result for this type of CCP.

For *Cooperation* and the *Group drawing*, a Monte Carlo algorithm was developed by implementing one modification to take into consideration additional parameters. For the *Cooperation* extension, the algorithm was altered to enable the collection of n times each available unique sticker. For the *Group drawing* extension, the drawing mechanism was enhanced to draw without replacement a various number m of stickers inside the same packet. Those two improvements are not exclusive and can be combined at the same time into a new Monte Carlo algorithm replicating the behaviour of an extended CCP. It has to be noted that this complexity comes at a cost as it lengthens the computing time. While this new algorithm can replicate a different CCP with much broader specificity, the computing time is affected twice as much by the values of n and m. Hence, the computing time for $Z^{10,5}_{682,682}$ is about four times longer than the same number of simulations for $Z^{1,1}_{682,682}$.

Figure 9 and table 4 show the Monte Carlo outputs for $Z^{10,5}_{682,682}$. This time, no analytical results can verify the accuracy of the Monte Carlo algorithm. This algorithm is however based on two previous implementations that were very accurate, leading us to believe the error from this Monte Carlo output would be comparable with the ones previously identified, or at most a factor of these. Since the difference observed for the previous algorithms was at most 0.04 %, even if this error was ten times higher, it would still remain relatively accurate with an error of less than 1 %. Given the scope and the objectives of this paper, it is unlikely that an error of this size would affect the conclusions at the end of the analysis.

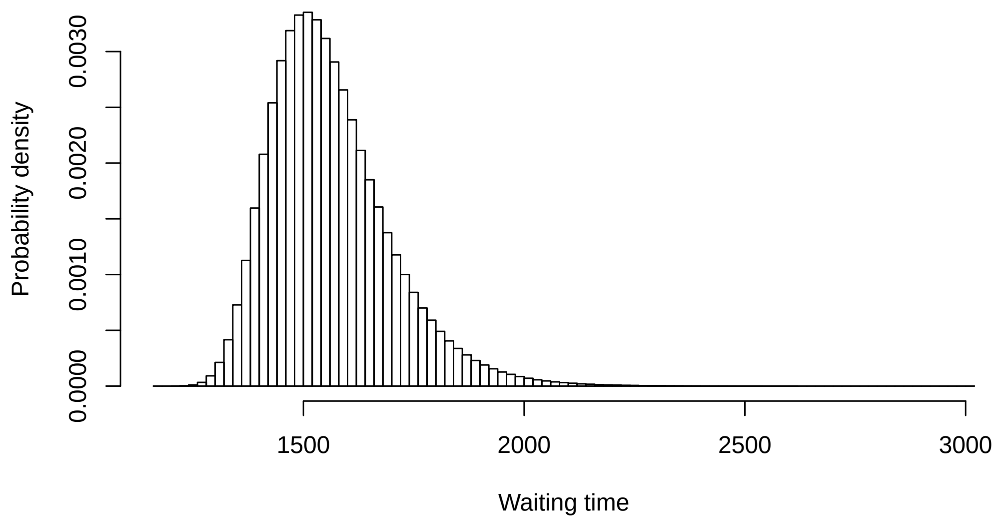

Figure 9: Histogram of simulated waiting times $Z^{10,5}_{682,682}$. Number of simulations : 10^7.

The distribution of the Monte Carlo outputs on figure 9 presents little difference from the shape of the outputs on figure 7. Around 86 % of the outputs are standing between 1400 and 1800. If compared with the previous distribution, the left tail is shorter while the right tail is longer, which slowly decreases with the waiting time. Close to 1 % of the outputs are beyond 2000, while the maximum waiting time observed was 3000. Interestingly, only 44 % of the outputs show a waiting time above the mean waiting time of 1561. This is explained by the long right tail of the distribution, where events with low probability but large waiting time are pushing the mean to a higher value.

Table 4 compares the mean and the standard deviation of $Z^{10,5}_{682,682}$ with those of $Z^{10,1}_{682,682}$ and $Z^{1,1}_{682,682}$. It can be easily observed that the CCP with both extensions is much closer to the CCP with only the *Cooperation* extension. This is probably explained due to the low effect of the *Group drawing* extension with only five stickers in a packet. A comparison between $Z^{10,5}_{682,682}$ and $Z^{10,1}_{682,682}$ demonstrates that *Group drawing* extension continues to improve the waiting time slightly, lowering

Table 4: Comparison of Monte Carlo outputs of $Z^{10,5}_{682,682}$ with other types of CCP.

	$Z^{10,5}_{682,682}$	$Z^{10,1}_{682,682}$	$Z^{1,1}_{682,682}$
Mean	1560.76	1563.37	4844.231
Standard deviation	135.17	135.54	871.53

the mean by .15 % and the standard deviation by .25 %.

The results of this extended CCP combined with both extensions are an exact replication of the behaviour of one or multiple collectors cooperating to complete a 2018 World Cup collection. The presented results are therefore readily applicable to analyse this specific collection. For other collections, one might replicate the methodology used here to obtain an accurate approximation of the distribution of the waiting time based on a different collection size s, particular number m of stickers inside a packet and various number n of cooperating collectors.

3.3 Sensitivity analysis on CCP parameters

Figure 10 represents a quick comparison of the last four Monte Carlo outputs obtained. It can be seen that the number n of collectors cooperating has a very significant impact on the waiting time, whereas changing the number m of stickers inside a packet has little impact. In the upper part of figure 10, it is possible to discern two distribution based only on the parameter n. The impact of modifying m is hardly noticeable that it is necessary to zoom in on the top of the densities to observe any difference between $m = 1$ and $m = 5$.

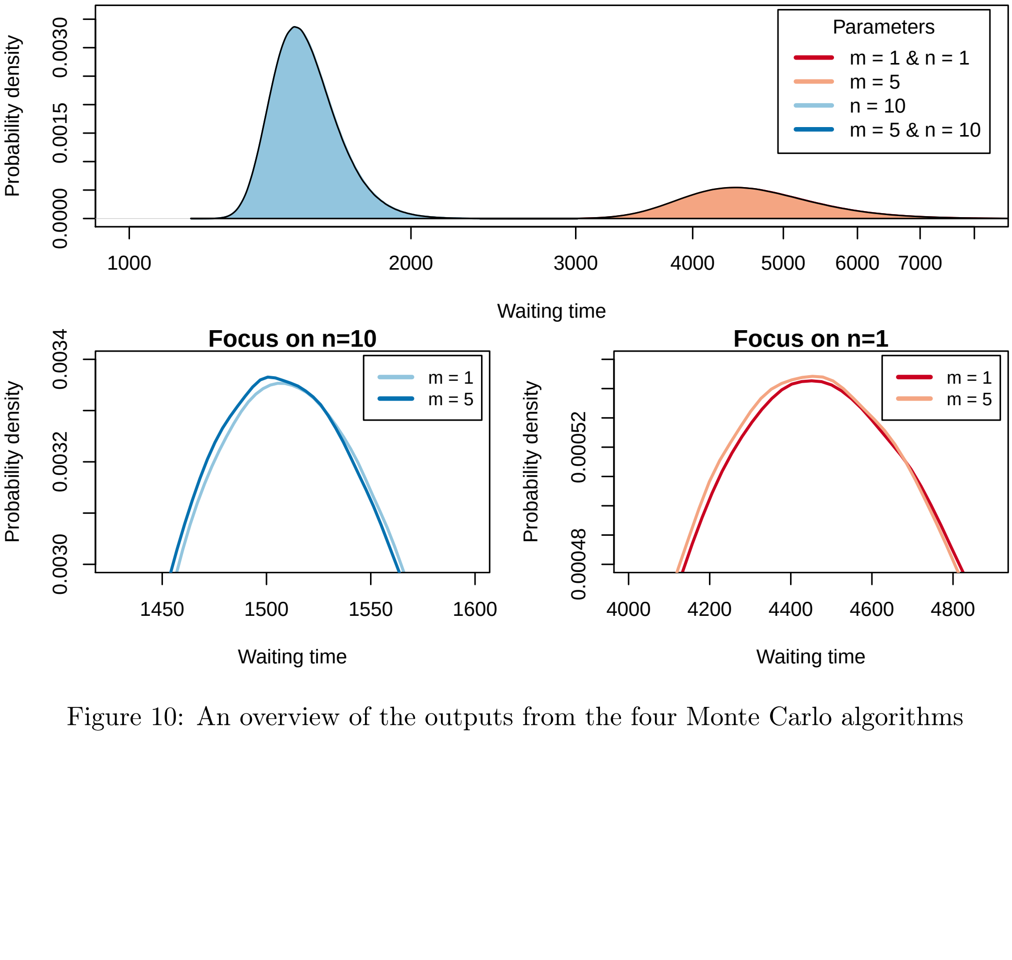

Figure 10: An overview of the outputs from the four Monte Carlo algorithms

The low impact of the parameter m is probably because m is much smaller than the collection size s. Studying the evolution of the waiting time for larger values of m is one of the objectives of this subsection. Additionally, it is also interesting to investigate how the waiting time is altered when changing other parameters of the model, such as s and n. For example, given that the waiting time cannot be inferior to s, regardless of the other parameters, is there a certain value of n that will allow the expected waiting time to be equal to s? This is one type of question that will be answered in this sensitivity analysis.

3.3.1 Number of cooperating collectors

As previously observed, the number of collectors cooperating has a very significant impact on the waiting time for each of the participating collectors. It is expected that the larger the number of collectors, the lower the waiting time, up to an asymptotic limit of s. To perform this analysis, waiting times $Z^{n,5}_{682,682}$ were computed for values of n from 1 to 100. Figure 11 displays the resulting expected waiting time and the standard deviation.

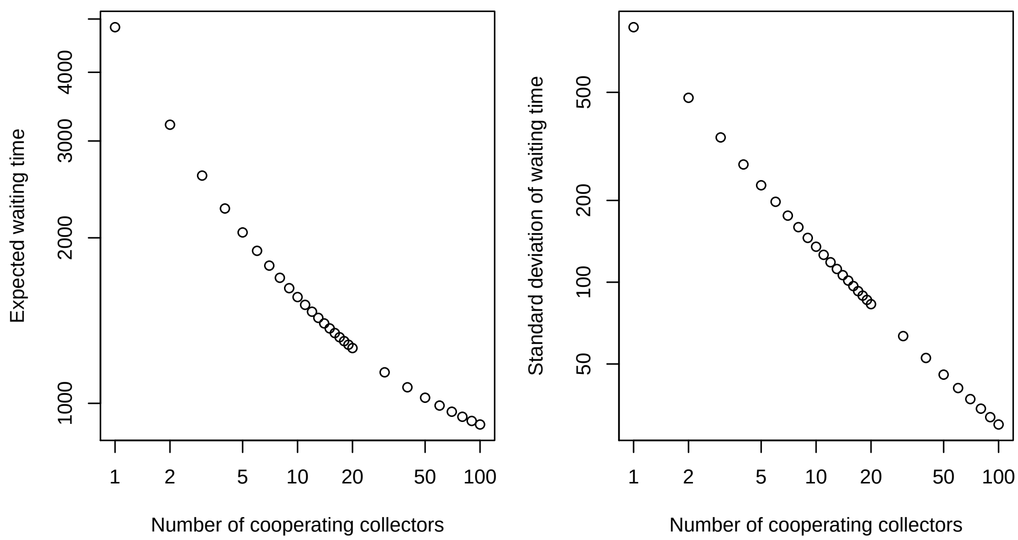

Figure 11: Evolution of the expected waiting time and its standard deviation for $Z^{n,5}_{682,682}$. Number of simulations : 10^6 per point.

It can be seen that each addition of a new collector has a reduced impact compared to the previous one. The maximum expected waiting time is related to a single collector and the expected waiting time drops rapidly as more collectors start to cooperate. The expected waiting time is decreased by half when 4 collectors are cooperating. If 9 collectors are cooperating, this further lowers the expected waiting to one third of its original value. However, reducing the waiting time to one quarter of the initial value requires more than 20 collectors. With 100 collectors, the expected waiting time is 915, close to one fifth of the value for a single collector. Due to the heavy computing time required when a large number of collectors are involved, points beyond 100 collectors were not computed. It is expected, however, that the expected waiting time continues to lower as the number of collectors grows, until an asymptotic limit of around 682 for an infinite number of collectors considered.

The standard deviation follows a similar pattern, although it shows a more pronounced decreased. From the initial value of 869 for a single collector, the standard deviation is rapidly reduced to about a quarter when 5 collectors are cooperating. A group of 19 collectors brings the standard deviation to less than 10 % of the single collector equivalent. With 100 collectors, the standard deviation is almost three times lower, at 30. The important decrease of standard deviation makes sense. The main reason for the large volatility of the waiting time for a single collector comes from unlucky situations where the collector struggles to collect a small number of remaining missing stickers. As the number of collectors cooperating increases, these situations tend to disappear as hardly obtainable missing stickers are swapped with duplicates. Outside the organisational complexity it can create, it is logical to state that it is in the interest of collectors to cooperate in as large groups as possible.

3.3.2 Collection size - number of unique stickers

The collection size is certainly the most important parameter of the collection as it directly affects the waiting time, since with each unique sticker added to the collection size, the expected waiting time will grow by at least one. Studying the evolution of the waiting time for different collection sizes also informs whether the analysis presented in this paper might be relevant to other collections or not. Should the waiting time behave differently for a smaller or bigger collection, that would mean that the conclusions of this paper would not be acceptable. Besides, it is also interesting to explore whether the collection size specified for the collections published by Panini comes from an economic reasoning.

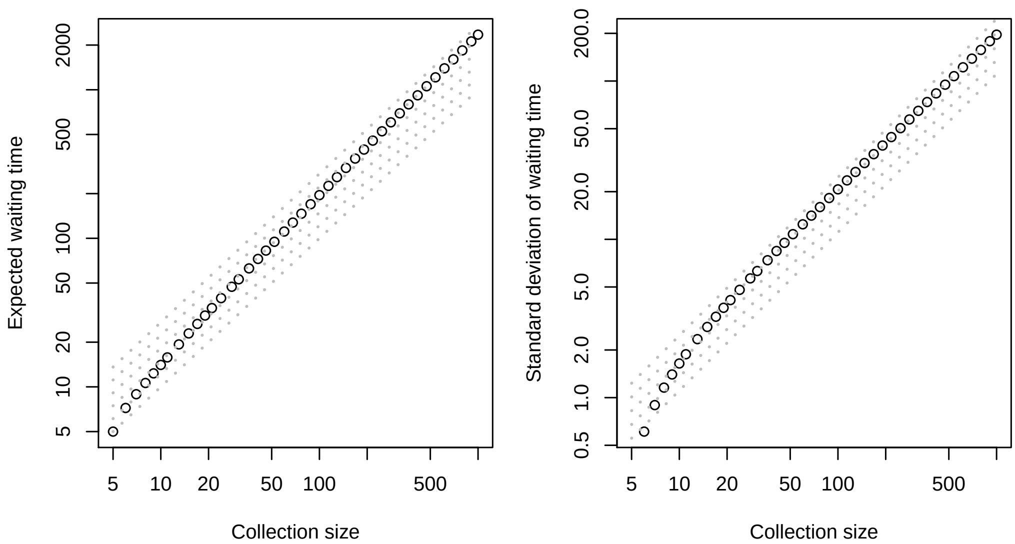

Figure 12: Evolution of the expected waiting time and its standard deviation for $Z_{s,s}^{10,5}$. Number of simulations : 10^6 per point. The grey lines represent linear functions.

Figure 12 exposes the result of this analysis. The mean and the standard deviation were computed for a number of different collections sized between 5 and 1000. To outline the behaviour of the expected waiting time and the standard deviation, grey lines representing linear relationships $y = k{*}x$ were added. In comparison with the grey lines, it appears that neither the expected waiting time nor the standard deviation are linearly related to the collection size. Indeed, the expected waiting time equals 14 with a collection size of 10 stickers, becoming 195 for a collection size of 100 stickers, and finally reaching 2350 at the collection made up of 1000 unique stickers. For those points, the linear coefficient is first 1.4, then 1.95 and finally 2.35. This clearly indicates a non-linear relationship. A similar observation can be made in respect to the standard deviation, especially for the smaller collections. However, when the collections are larger than 20, the standard deviation starts to grow linearly with the collection size. The explanation behind the low expected waiting time and the standard deviation for a smaller collection size is partly due to the *Group drawing* component. This analysis has been performed using $m = 5$, which means that for a collection of 5, no randomness is involved and one packet will be sufficient to complete the collection. This effect rapidly decays for a larger collection size, as the ratio between m and n decreases.

The expected waiting time is not linearly related with the collection size, which demonstrates that it is in Panini's best interest to publish collections of a larger size. There is twice the benefits for raising the number of unique stickers in a collection. First, the increase will allow Panini to sell more stickers to collectors as a consequence of the elevated waiting time. Secondly, the larger the collection size, the more cost-effective the stickers production. Indeed, if the marginal cost of enlarging the collection is constant, the corresponding marginal profit increases as a result of the non-linear relationship.

3.3.3 Stickers packet size

When the *Group drawing* extension was introduced, its impact remained minimal when compared to the influence of the *Cooperation*. One explanation behind this meagre impact could lie in the scale of the parameter m used until now. In comparison with the size of the 2018 World Cup collection and its 682 unique stickers, obtaining 5 unique stickers at once in a packet is only a small fraction of the entire collection. To investigate the behaviour of the relationship between packet size and waiting time, computations of $Z_{682,682}^{10,m}$ were performed for the values of m ranging from 5 to 682. Figure 13 presents the results.

It can be seen that the impact of the packet size remains rather low for small values of m. The effect becomes more noticeable once m reaches a value of 100, lowering the expected waiting time to 1497, 63 units below the expected waiting time of $Z_{682,682}^{10,5}$. The waiting time then continues to lower as the packet size grows at a moderate speed, reaching 1249 for a packet of 410 stickers and falling under 1000 for a packet of 610. In the end, the waiting time drops to 682 once the packet size reaches 682 with a single packet containing all unique stickers. The impact of the packet size is highly related to the size of the collection. Not having any duplicates inside a packet is helpful for collectors who are starting their collections, but this advantage quickly fades away as more packets are bought and duplicates between different packets are found. Compared to a collection size of 682, a packet of 5 is less than 1% of the collection and thus way too small to be noticeable in the long run. Only packets of more than 20 % of the collection size start to have a noticeable impact by reducing the expected waiting time by 5 %.

From this analysis, it appears that changing the size of the packet has close to no effect on the waiting time for smaller sized packets. Therefore, Panini does not really help collectors by ensuring packets contain no duplicates but this possibly contributes to the image of the brand without

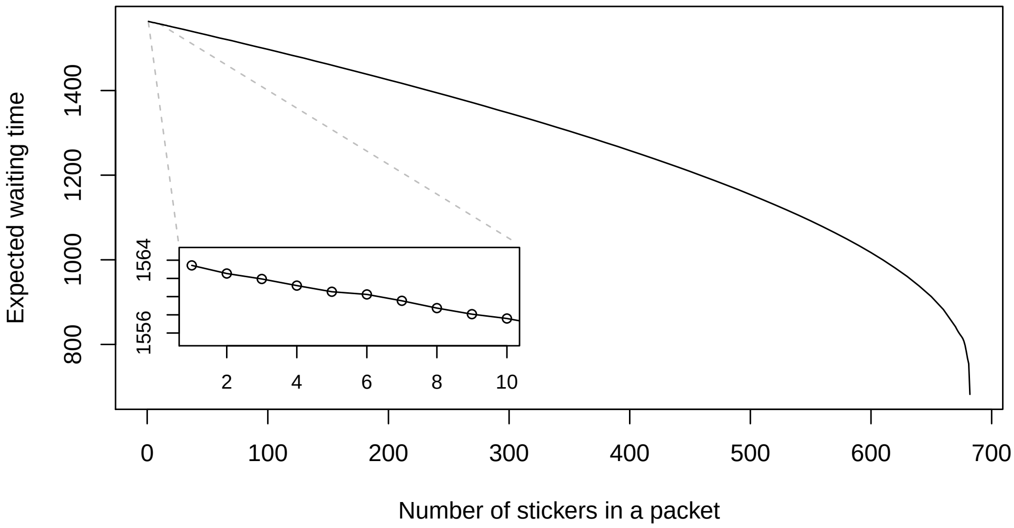

Figure 13: Evolution of the expected waiting time and its standard deviation for $Z^{10,m}_{682,682}$. Number of simulations : 10^6 per point. The smaller window displays a focus on the behaviour of the expected waiting time for $m = 1, ..., 10$.

hurting sales. That being said, the true cost-benefits off this practice are uncertain, as it probably costs money for Panini to organize its supply chain with the goal of not having any duplicates in packets.

One question that can be raised is whether Panini actually pays attention to whether a packet has two identical stickers. Given the large number of unique stickers available, the probability of obtaining two identical stickers in a packet is rather small even if packets are randomly constituted. This probability can be identified using the following formula :

$$P(\text{not having m unique stickers in a packet}) = \frac{s^m - \frac{s!}{(s-m)!}}{s^m}$$

Which gives a probability of 0.0146 or 1.46 %. This low number provokes further questions such as if Panini actually makes efforts to avoid packets with duplicates. For a single collector, finding a packet with a duplicate in every 50 packets or more could simply indicate an error in the factory process. However, it is difficult to investigate this question further without the requisite data: a larger number of packets to track the number of duplicates within them.

3.4 Numerical analysis on optimal strategy

With the new numerical tools developed with the help of Monte Carlo algorithm, it is possible to compute a distribution for any kind of waiting time $Z^{n;m}_{s,s}$. This allows for a deeper analysis of the strategy introduced in the literature review, notably the consideration of groups of collector. This section will make use of the numerical tools to investigate questions previously outside of reach using the analytical approach.

3.4.1 Strategy with cooperation

Section 2.4 presented the strategy for a single collector as defined by Stadjes [Stadje, 1990]. Equation (13) and (14) provide the expected cost for different shift points j at which the collector should stop buying packets of stickers and instead order stickers directly from the supplier at a premium price. These expressions are readily usable for any value of the parameter m, the number of stickers in a packet. However, the number of collectors cooperating was not covered by Stadjes. It is possible to extend equation (13) for any n, the only issue being the need to compute $\mathbb{E}(Z^{n,m}_{j,s})$:

$$v(j) = \mathbb{E}(Z^{n,m}_{j,s}) + \theta(s-j) \qquad \text{with } j = 1, 2, ..., s \tag{15}$$

Luckily, this is exactly what the Monte Carlo algorithm (developed in section 3.2.4) is able to compute. The only difference between this problem and section 3.2.4 is the need to compute waiting times for values of j different from s. So far, only waiting times similar to $Z^{n,m}_{j=s,s}$ have been computed. A value of j different from s implies that the algorithm stops after obtaining j unique stickers, as opposed to obtaining all available unique stickers when $j = s$. While this is straightforward for a single collector, there is additional work required for more than one cooperating collectors. The objective of the algorithm for multiple collectors is to collect n times each available stickers. For a value of $j < s$, this cannot be simply translated into collecting n times j unique stickers since this would be inefficient. Indeed, it is possible to obtain n collections with j unique stickers in each of them without having n units of j stickers in total (i.e., the unique stickers within each collection may be different from one collection to another). Therefore, the end point of the algorithm for $Z^{n,m}_{j,s}$ has to be defined when reaching

$$\sum_{i=0}^{s} \min(n, k_i) = nj,$$

where k_i is the number of units of stickers $i \in s$ when considering all cooperating collections.

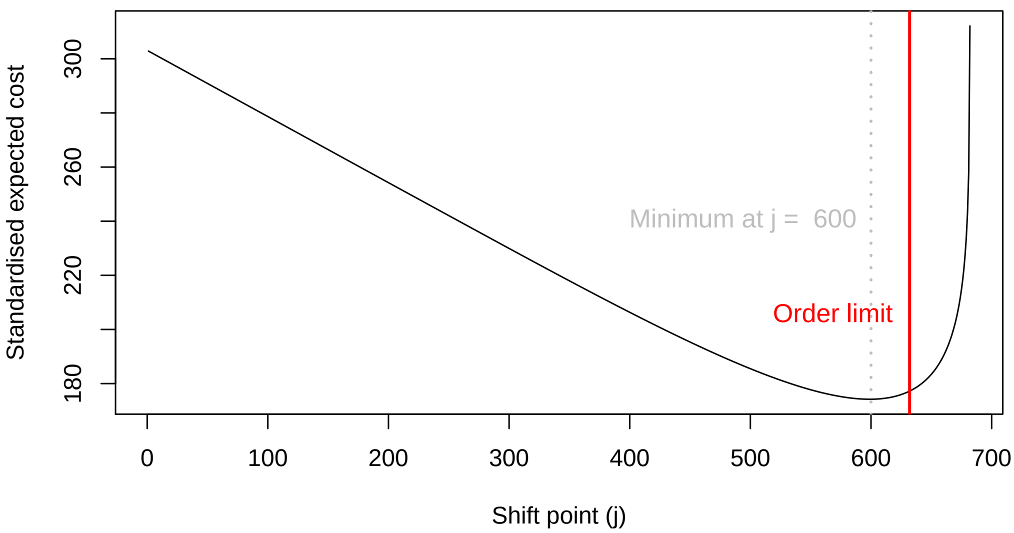

Figure 14: Standardised expected costs for $Z^{10,5}_{j,682}$. Number of simulations : 10^5.

Using that trick, $v(j)$ was computed for all possible values of j, as presented in figure 14. The specification of the CCP remained with $n = 10$ and $m = 5$, in accordance with the specification of section 3.2.4. The minimum of $v(j)$ is found at the optimal shift point $j^* = 600$, much later than for a single collector. The corresponding standardised expected cost of v^* is 174.22. Unfortunately, there is still an *Order limit* imposed by Panini, which limits the number of stickers ordered from the supplier up to 50. This translates into a new optimal shift point $j^* = 632$ and a standardised expected cost of 177.22. It can be seen that the change in the optimal shift point due to the *Order limit* is much less significant than for the single collector case. This is explained by the minimum of the strategy $v(j)$ being already very close to the *Order limit* coordinates, as visible in figure 14.

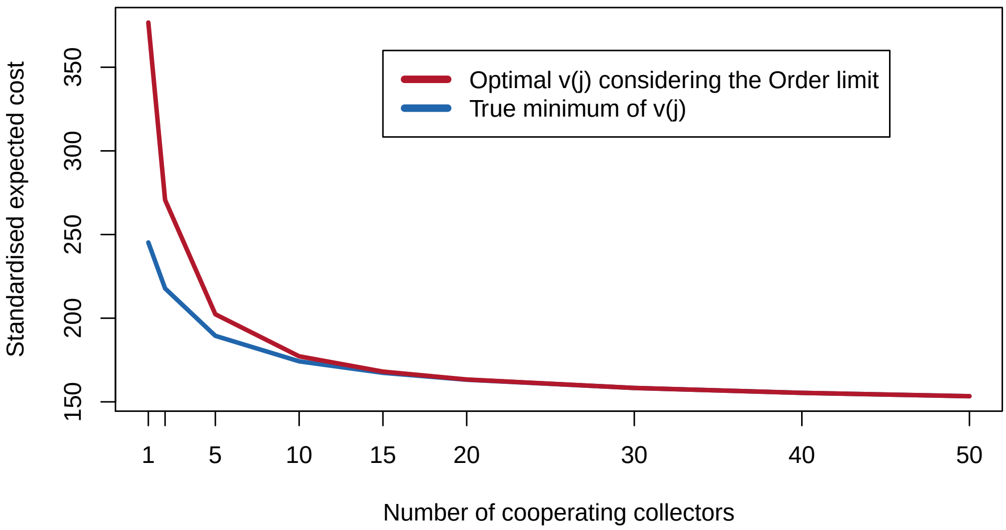

Figure 15: Evolution of the standardised expected costs for $Z^{n,5}_{j,682}$. Number of simulations : 10^5.

To truly analyse how the number of collectors affects the optimal strategy, the same computation was performed by varying n from 1 to 50. For each computed n, the minimum of the standardised expected cost v^* was recorded, so was the optimal standardised expected cost considering the *Order limit* of 632, i.e. $v(\max(j^*, 632))$. Results are presented in figure 15. It appears that the cooperation of collectors remains highly beneficial to lower the waiting time, although less than when no strategy is applied (see sensitivity analysis in section 3.3.1). The behaviour of the *Order limit* was investigated as well, with the blue line representing the minimum of the strategy function $v(j)$ while the red line also accounting for the *Order limit*. It can be observed that the *Order limit* significantly increases the cost for a single collector or for a cooperating group of less than five collectors. This effect however decays rapidly as the number of cooperating collectors grows, becoming insignificant at around 10 cooperating collectors. The constraint of the *Order limit* even disappears for larger groups of more than 30 collectors, where the optimal shift point j^* is greater than the *Order limit* of 632.

3.4.2 Premium ratio pricing and Order limit sensitivity

It is interesting to explore why Panini decided to price the ordered stickers from its website at $\beta = 0.4$. For groups of collectors of size inferior to 30, the *Order limit* continues to constrain their optimal strategy. It has the direct consequence that collectors will usually order at the *Order* *limit* since they are not able to apply the optimal strategy. Interestingly, this represents a missed opportunity for Panini, as Panini could raise the price of ordered stickers to the point at which the optimal shift point coincides with the *Order limit*. After this change, the point at which collectors order stickers remains the same (since they apply the optimal strategy) but the expected cost rises, generating an increase in profits.

Figure 16 explores this problem in detail. Although this figure is quite complex, it is the best way to analyse the pricing of ordered stickers at a premium and the relation with the *Order limit*. The x axis represents variations of *Order limit*, for which the baseline value of 632 is indicated by the dotted line. The y axis represents the optimal standardised cost accounting for the *Order limit*, meaning it is the minimum between $v(j^*)$ and $v(Order\ limit)$. Each line then represents the optimal strategy as a function of the *Order limit*, given a premium ratio θ. The line computed with $\theta = 0.44$ represents the baseline premium pricing of ordered stickers. Finally, the inflexion points are representing the breakpoints where $v(j^*)$ and $v(Order\ limit)$ are equal, meaning the points beyond which the decrease in *Order limit* does not affect the optimal strategy.

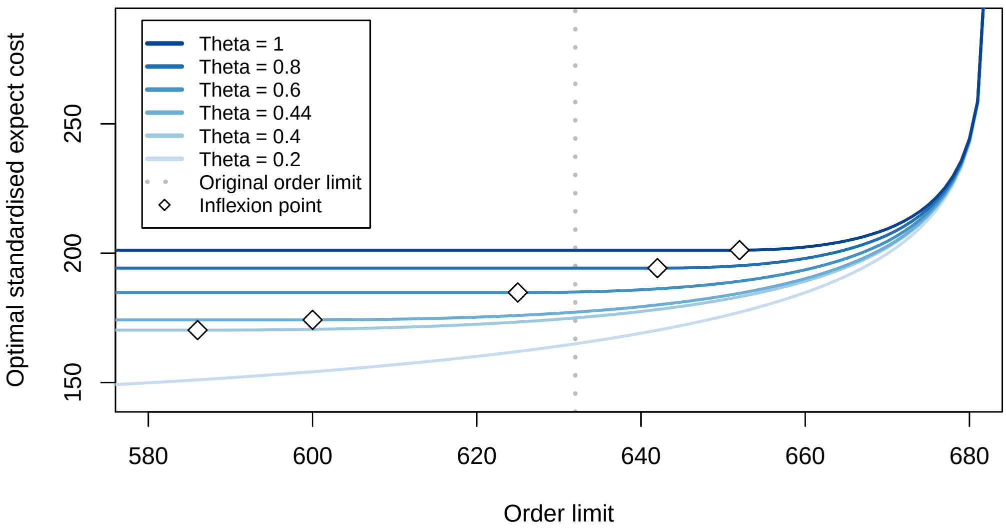

Figure 16: Optimal strategy costs for $Z^{10,5}_{j,682}$, varying the *Order limit* and the premium ratio θ. The dotted line represents the original *Order limit* of 632 as established by Panini. The inflexion points are the points at which the cost function is constant before and increasing beyond that point.

Two main points are observable from the analysis of figure 16. First, it can be seen that modifying the *Order limit* offers little gains or losses for Panini, unless Panini lowers the number of sticker which each collector can order to less than 20 (at which point the standard expected cost for collectors raise above 192, or a 8 % increase). This change could however be poorly perceived by the collectors, for which the number of stickers they can order is reduced by more than half.

The second observation is related to the premium ratio θ, which has a significant impact on the expected cost of the original *Order limit* of 632 and lower. By looking at the intersection between they grey dotted line and the strategy lines for different premium ratios, it can be seen that increasing the premium ratio would directly increase the expected cost for collectors and generate an increase in profit for Panini. For example, when the original *Order limit* is 632, if the premium ratio is raised to 0.6, the standardised expected cost jumps to 185 representing an increase of 7 %. In this case, collectors would still order at the *Order limit* since it remains the optimal shift point. However, it has to be noted that if the premium ratio is increased further, it would change the optimal strategy for collectors. Diamonds in figure 16 essentially represent inflexion points, or the points beyond which the minimum $v(j^*)$ is no longer constrained by the *Order limit*. For example, if the premium price is raised to 0.8, the standardised expected cost increases to 194 but optimal shift point has now changed. The new optimal shift point is beyond the *Order limit*, at $j^* = 642$. This optimal shift point is indicated by the diamond on the figure, and an *Order limit* lower than it will not affect the strategy of collectors.

By exploring multiple values of θ, it can be found that $\theta = \frac{2}{3}$ offers a comfortable increase in Panini's profits by increasing the minimum standardised expected cost to 188.33 while altering the optimal shift point to $j^* = 633$, a value much closer to the *Order limit*. It can also be noted that this value of θ brings the optimal shift point closer to the *Order limit* only for the specific number of cooperating collector analysed. To perform an exhaustive analysis of the pricing of the premium ration, it would be necessary to analyse multiple sizes of groups of collectors. As this is however not the main subject of this paper, the study of other group size of collectors will remain neglected.

3.4.3 Risk analysis of strategies

Thus far this thesis has analysed strategies in terms of minimizing the expected cost for collectors. While determining the optimal strategy to minimize this expected cost is useful (especially since it is what most collectors are usually looking for), it is also possible to formulate the collectors' objective in another form. For example, one collector could indeed be interested in spending at most V to complete his collection, for a given probability p. The higher the probability p, the more limited the risk for the collector.

The first step needed to analyse the strategy for a collector interested in limiting his risk is to look at the distribution of the standardised cost when applying the optimal strategy. In this case, a group of collectors of size $n = 10$ will be considered, applying the optimal strategy under the *Order limit* constraint of 632, resulting in the optimal shift point $j^* = 632$, as determined in the previous sections. Likewise, the premium ratio also retains its original value $\theta = \frac{4}{9}$. Therefore, the standardised cost of the optimal strategy will have the following form:

$$Z^{10,5}_{632,682} + \frac{4}{9}(682 - 632) \approx Z^{10,5}_{632,682} + 22.22$$

The waiting time $Z^{10,5}_{632,682}$ represents the only random component of the strategy. Computing the distribution of $Z^{10,5}_{632,682}$ and adding the fixed cost of 22.22 effectively provides a distribution of the optimal strategy. This distribution was computed and the results are found in figure 17, this time presenting the cumulative distribution function of the standardised cost.

Looking at figure 17, it is easy to determine a probability p and to obtain the corresponding standardised cost. For example, with a probability of $p = 0.9$, the corresponding standardised cost is 179.22. This means that with this strategy, there is a probability of 0.9 that the standardised cost does not exceed 179.22. While this result is informative for the collector, it does not indicate

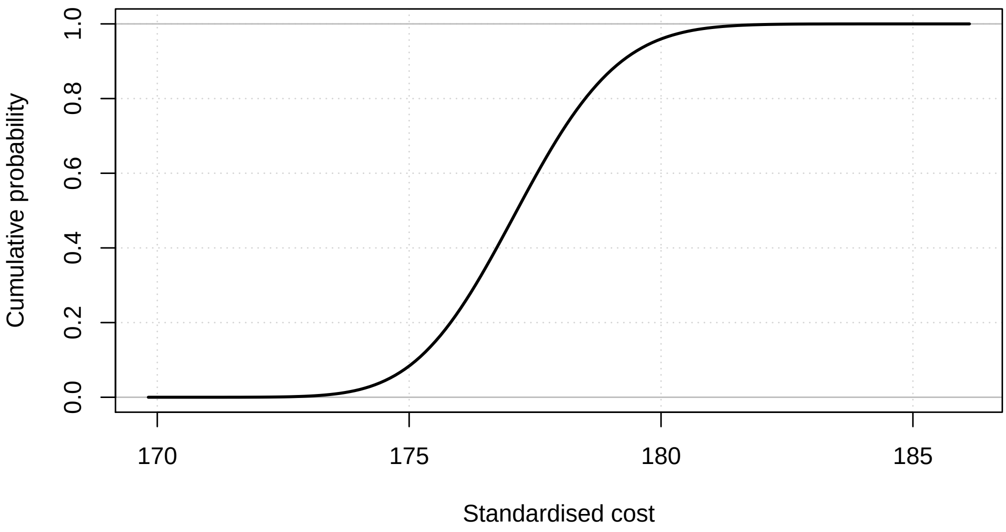

Figure 17: Cumulative distribution function of the costs for the strategy using the shift point $j = 632$. Number of simulations used : 10^7.

whether he should retain this strategy or change to another. Since the optimal strategy only minimizes the expected cost, the cost for a probability of $p = 0.9$ may not be minimized.

Obtaining the minimal cost for a probability of $p = 0.9$ requires computing the distributions of the waiting time for $Z^{10,5}_{j,682}$ for a range of j around the optimal shift point. For each distribution of $Z^{10,5}_{j,682}$, the fixed cost $\theta(682 - s)$ is added to obtain the distribution of the standardised cost. From this distribution is extracted the standardised cost for the cumulative probability of $p = 0.9$.

Figure 18 represents the result of this procedure, revealing the minimum for the shift point $j = 594$ and a corresponding standardised cost of 175.61. This point is unfortunately lower than the *Order limit* and is therefore unusable, leaving the previous result determined in figure 17 as the optimal shift point. Should one collector get rid of the *Order limit* constraint, the minimum at $j = 594$ offers little improvement over the original optimal shift point $j^* = 600$. The limited cost at a probability of $p = 0.9$ can only be lowered from 175.64 to 175.61 by deviating from the optimal shift point.

An interesting characteristic of the detailed part of figure 17 is the saw-tooth behaviour around the minimum. This is explained by the fact that the standardised cost uses two components; the first component coming from the cost of ordering $s - j$ stickers at premium and the second component coming from the waiting time to collect j stickers. The first component decays linearly as j increases while the second behaves as a step function and jumps to higher values at various intervals, more regularly as j increases. Around the minimum, the jumps of the second component are becoming more regular, stabilizing the function and presenting the saw-tooth behaviour.

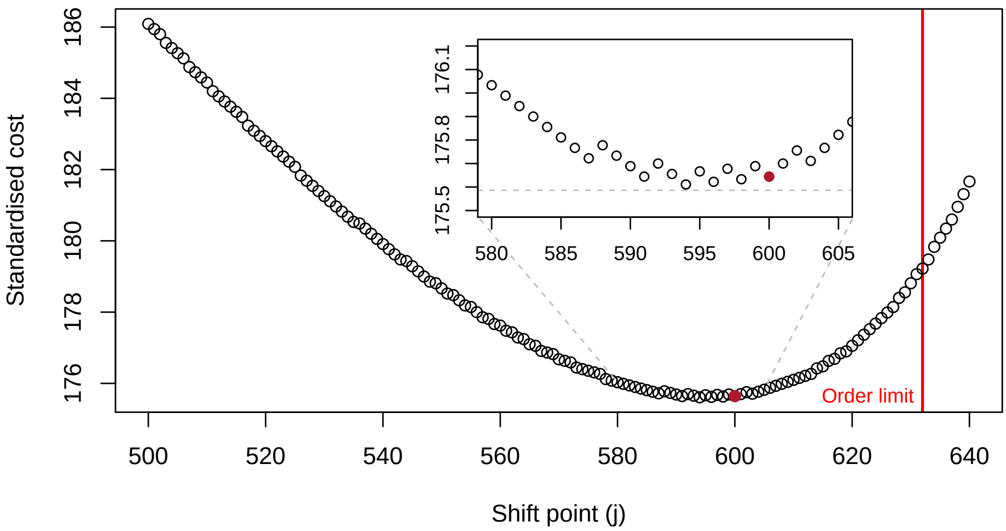

Figure 18: Standardised cost corresponding to a probability $p = 0.9$ computing waiting times $Z^{10,5}_{j,682}$, varying the shift point j. Number of simulations : 10^5 per point. The red dot represents the corresponding cost for the optimal shift point j^*. The smaller window displays a focus on the behaviour of the standardised expected cost for $j = 580, ..., 605$.

3.4.4 Limits of the strategy analyses

So far, the analysis of strategies has only covered strategies starting from scratch, i.e., when no sticker has been collected yet. It would be convenient to have a dynamic strategy, determining the optimal procedure based on the state of a current collection. This is however a tedious challenge when multiple collectors are considered, since the same collection state (indicating the number of stickers collected) may refers to multiple different situations. Indeed, as introduced in section 3.4.1, an optimized allocation of stickers to account for n collections containing j stickers each in the following form:

$$\sum_{i=0}^{s} \min(n, k_i) = nj,$$

However, this allows numerous allocations to reach the value nj. For example, considering $s, n = 3$, $j = 2$ and $m = 1$, there are 10 different allocations leading to $nj = 6$. Among them, 3 specific cases are outlined:

- Three times two stickers, i.e. (3,3,0)
- Three times one sticker, two and one, i.e. (3,2,1)
- Two times three stickers, i.e. (2,2,2)

Each of these cases implies different expectations when obtaining a new sticker. The case (2,2,2) will certainly draw a new useful sticker while the case (3,3,0) has only a probability of $\frac{1}{3}$ of obtaining a needed sticker.

Creating a dynamic strategy would require the analysis of the probabilities related to each possible case. While it is possible for the problem exposed above now, the number of cases to consider drastically increases when using values of $s = 682$, $n = 10$ as parameters. This laborious work has not been realized for this paper and probably presents little interest as the *Order limit* constraint hinders the number of applicable strategies. Without the constraint, a dynamic strategy would however allow collectors to adapt their strategies based on the progression of their collection and the current cost, leading to new ways to maintain the cost below a threshold.

4 Swapping modelling

Throughout this paper, the notion of collectors cooperating has been widely used. However, within this notion is an inherent assumption that every group of cooperating collectors follows a comparable cooperation. The *Cooperation* as defined in the literature is indeed presented as a perfect type of cooperation. The real-life application of such cooperation is questionable. This section intends to review the unrealistic elements in the perfect cooperation case and to provide a more realistic representation of the cooperation between collectors.

4.1 A criticism over the general swapping assumption

In his paper about the double-dixie cup problem [Newman, 1960], Newman considers the completion of n collection by obtaining n units of each sticker, without any regard paid to the number of collectors involved. In the previous sections, we assumed that this result could be applied to a cooperation of n collectors, each of whom completes his own collection. This statement is indeed equivalent to the Newman's formulation of the problem if these collectors are acting as a buying-group, committed to finance from the first to the last coupon needed to complete n collections. This means that collected stickers are not attributed to any collectors before the n collections are complete, even if it is already possible to complete multiple collections.

In a real-life situation, collectors of Panini stickers hardly behave in this way. For most of them, their collection is personal. The progress of their collection is something they wish to follow, which is part of their fun. It seems unlikely that collectors would be willing to lose this aspect of the collection process. Obtaining new stickers is also something personal and related to the progress of the collection. Collectors tend to open packets on their own, in the hopes of obtaining new stickers for their own collection. That is to say, they often finance their collections on their own.

Moreover, interactions between collectors are complex and rarely relate to a group behaviour. Two different collectors may have common connections, but also connections exclusive to each of them. In addition to that, each of them may also maintain a different number of connections overall. Since relations between collectors are not equal, just as human relations are not equal, there are some inherent ties that are closer and others which are less so. For example, a collector is more likely to swap his stickers first with a close friend, then only later with an acquaintance.

Additionally, it can be noted that the different pace at which collectors collect and swap stickers is not considered. The flow of stickers is different for each collector and collectors do not necessarily look for swaps after every inflow of duplicates. Some collectors may be more likely to buy stickers and swap very quickly, leaving them with almost no-one to trade with since they outpace the rest of the collectors. Likewise, a collector starting his collection very late might struggle to find partner to swap with if most of the other collectors have already finished their collection.

All of these differences have an impact on the collection of each collector. Despite the fact that the real-world situations may differ from the assumptions in general, assumptions are often necessary to obtain results and it would be impossible to obtain a closed form solution similar to the one developed by Newman without this assumption. However, we believe that the impact is significant enough to create a large divergence between results under the perfect cooperation assumption and the real-life problem.

This is why this case uses models, unlike that of Newman, in order to better represent the reality of swaps by collectors. Fortunately, Monte Carlo methods are applicable to virtually any problem. Even without providing a closed form solution, it is possible to compute distributions for more complex models and to compare them to the classical models coming from the literature and the numerical analysis of section 3.

4.2 Network representation

Relations between individuals are often depicted under the form of a network. Someone's network is constituted of all kinds of the relationships of an individual, whether on a personal or professional level.

This general definition of a network has to be adapted for an application to the CCP. Since networks can be defined as considerably complex, for practical reasons only simple networks will be explored in this paper. In particular, this paper will explore two following network models : a network model focusing on low interconnections and a network model focusing on high interconnections.

4.2.1 Circle networks

This first model is shaped in the form of a circle, as represented by figure 19. In a circle style of network, the number of links between collectors is minimized, regardless of the number of collectors. Each collector is only linked to two other collectors, in a way to create a chain of connections that closes on itself, forming a circle.

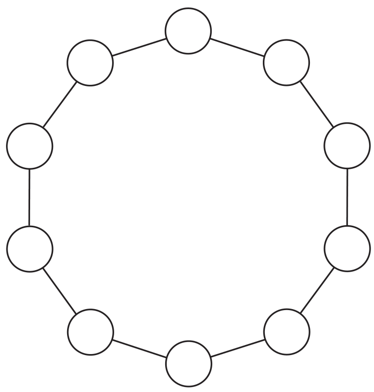

Figure 19: Circle Network

For the modelling of this network, groups of more than two collectors were considered. Collectors inside the network alternate between two phases. The first phase is the supplying of stickers, where each collector buys a certain number k of packets and records the new unique stickers and duplicates obtained. In the second phase, each collector swaps with his two links and performs all the possible exchanges of stickers. Once all possible swaps have been made, phase one starts again, which consequently leads to another second phase. This continues in a cycle until all collectors have completed their collections. Formally, the algorithm representing this model processes the following operations:

1. Each collector draws k times m stickers without replacement from the set of available unique stickers. If a collector completes his collection during this process, he becomes inactive and stops buying stickers.

2. Randomly order the collectors in a list.

3. The first collector of the list is to perform all possible swaps between himself and the collector linked *"at his right"*, based on the circle representation.

4. Repeating the previous operation for all collectors in the list, sequentially.

5. Collectors that did not complete their collection repeat from step 1. Collectors who have finished their collections become inactive.

The swaps realized in this model have a precise definition. A collector participating in a swap aims to exchange one of his duplicates with a sticker he is missing. This means the swap requires both collectors to have a duplicate sticker that interests the other collector for the swap to take place. Therefore, collectors with a complete collection will stop swapping since they have nothing to gain.

By repeating the algorithm multiple times, it is possible to obtain numerical results in the form of Monte Carlo outputs, similarly to the process performed in section 3. While the main interest remains to study the waiting time, a different notation has to be used since the *Circle network* is a completely different assumption. The waiting time corresponding to a *Circle network* will instead be noted $C_{j,s}^{n,m,k}$. The parameters n, m, j, s have the same meaning as in the standard CCP. k represents the delay between each phase of swaps. Unless specified at very high values, k does not strongly influence the behaviour of C. However, since k is linked to the number of times the swap phase happens, it significantly affects the computing time. To limit the computing time, k is then set at a value of 10 for the analysis developed here.

Figure 20 and table 5 both illustrate the resulting outputs for $C_{682,682}^{10,5,10}$, presented directly as a probability density function. To provide a comparison with the standard CCP studied before, the distribution of $Z_{682,682}^{3,5}$ was added to the graph. The choice of the comparison with $Z_{682,682}^{3,5}$ is not arbitrary, since collectors inside the *Circle network* are only related to two other people, each collector has access to swap with only two other collections. This is similar to the case where 3 collectors are cooperating perfectly. Another way to explain the behaviour of collectors in the *Circle network* would be that each collector in the circle experiences a *"non-closed"* three people cooperation. Comparing the *Circle network* to a perfectly cooperating group of the same size is not relevant due to the huge difference in interconnections.

The analysis of figure 20 reveals a distribution with a much larger right tail compared to $Z_{682,682}^{3,5}$. This result is explained by the structure of the network; i.e, when one collector completes his collection, he stops trading with other collectors. This leaves some collectors isolated when their two links have completed their collections. Those left isolated are not able to swap with anyone and must rely on buying packets of stickers to find the remaining unique stickers needed to complete their collection.

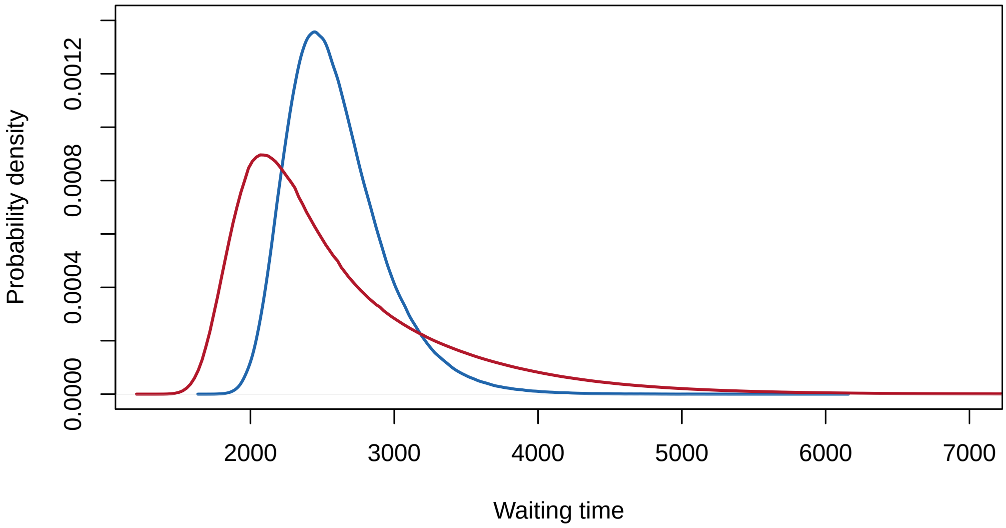

Figure 20: Distribution of the simulated waiting time for $C^{10,5,10}_{682,682}$, in red. Number of simulations : 10^7. The blue line represents the distribution of the simulated waiting time for $Z^{3,5}_{682,682}$, computed with 10^6 simulations.

Additionally, the left tail and peak of the distribution shift to the left compared to the standard CCP: the left tail of $C^{10,5,10}_{682,682}$ presents a significant offset from the left tail of $Z^{3,5}_{682,682}$, indicating that there are more favourable cases of low waiting time in using a *Circle network*. This is confirmed by at the medians of the two distributions presented in table 5. There is indeed a probability of 0.5 of having a waiting time inferior or equal to 2390 using the *Circle network*, compared to 2543 for the standard CCP. The explanation for the offset of the left tail of $C^{10,5,10}_{682,682}$ could originate from the fact that each collector finds himself in relation with different collectors. In the case of $Z^{3,5}_{682,682}$, if one collector does not have access to a specific sticker, then this also limits the two other collectors in the group. In contrast, collectors within the circle network may not share the same issue : even if Collector n°5 cannot get a specific sticker from his neighbours, Collectors n°4 and n°6, those neighbours might have access to this specific sticker from Collectors n°3 and n°7.

Table 5: Comparison of Monte Carlo outputs $C^{10,5,10}_{682,682}$ with a standard version of CCP.

Characteristic	$C^{10,5,10}_{682,682}$	$Z^{3,5}_{682,682}$
Mean	2602.343	2592.705
Standard deviation	757.6434	342.4827
Median	2390	2543.33

Furthermore, looking at table 5, it can be seen that the means of both distributions are very similar. This supports the relevance of the choice of $Z^{3,5}_{682,682}$ for this comparison. The standard deviation coming from the *Circle network* is more than twice larger. This is explained by the overall flatter aspect of the distribution of $C^{10,5,10}_{682,682}$: both tails are more pronounced as collectors can either enjoy a lower waiting time or a much larger one compared to $Z^{3,5}_{682,682}$, leading to a large increase in the variance of the distribution.

4.2.2 Star networks

This second network is fundamentally opposed to the *Circle network*. In this model, interconnections among collectors are maximized. All collectors are connected to all the other collectors either directly or through a central system, leading the network to be shaped like a star, which is illustrated by figure 21.

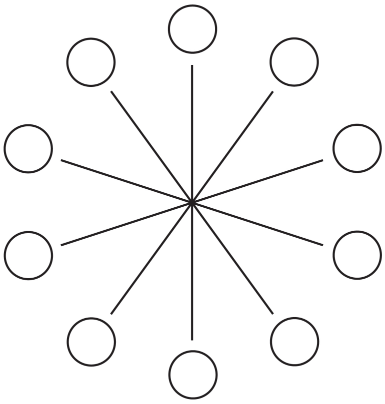

Figure 21: Star Network

The *Star network* behaves in a similar way to the model of the *Circle network*: Collectors also start by buying a number k of packets of stickers in the first phase and record all new stickers and duplicates obtained. The second phase is where comes the main difference, as each collector now swaps with all other collectors. Once every possible swap is performed, the process starts again until all collectors complete their collections. The corresponding algorithm can be formally represented as :

1. Each collector draws k times m stickers without replacement from the set of available unique stickers. If a collector completes his collection during this process, he becomes inactive and stops buying stickers.
2. Randomly order the collectors in a list.
3. The first collector in the list, swaps one time with another random collector. If the swap is not possible with this randomly selected collector, then the first collector selects another random collector until a swap is performed or all other collectors are selected. This process stops immediately after one swap is performed.
4. Repeat the previous operation for all collectors in the list, sequentially.
5. Repeat step 2 to 4 until all possible swaps are performed.
6. Collectors that did not complete their collection repeat from step 1. Collectors who finished their collection become inactive.

Swaps are made one by one to ensure that no collector is at an advantage over the others. However, this has significant consequences on the computing time since the algorithm has to perform more operations compared to the case where two collectors perform all the possible swaps between them at once. Furthermore, the algorithm often has to check if a trade is possible with another collector $n-1$ times for each collector, unlike only once in the case of the *Circle network*. This leads to an increase in the computing time of about n^2 compared to the previous network. Hence, it should be noted that the number of simulations for the Monte Carlo outputs will be substantially lower for the analysis of the *Star network* in this research in order to not engage computations that would last for weeks.

Since the *Star network* introduces a very different assumption from the standard CCP and the *Circle network*, a new notation has to be defined. The waiting time obtained with the *Star network* model will be noted $S_{j,s}^{n,m,k}$. The parameters n, m, k, j, s have the same meaning as for the *Circle network*.

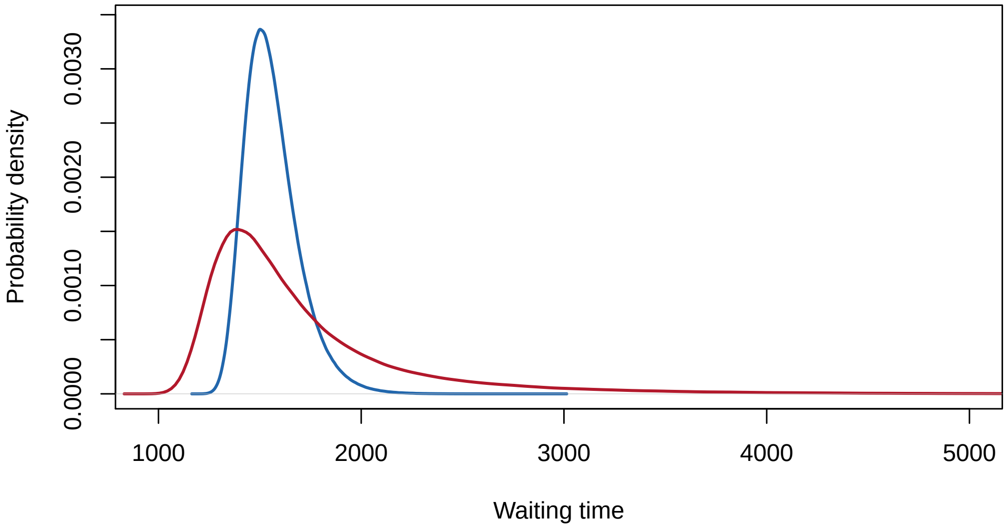

Figure 22: Distribution of the simulated waiting time for $S_{682,682}^{10,5,10}$, in red. Number of simulations : 10^5. The blue line represents the distribution of the simulated waiting time for $Z_{682,682}^{10,5}$, computed with 10^7 simulations.

The results obtained are presented in figure 22 and table 6. This time, the *Star network* is compared to a standard CCP with a group composed of 10 perfectly cooperating collectors. Since the interconnectivity is maximized in the *Star network* (each collector can swap with the other $n-1$ collectors) the comparison with $Z_{682,682}^{10,5}$ is the the most relevant. Observations are roughly similar to the ones pointed out for the *Circle network* analysis. The distribution of $S_{682,682}^{10,5,10}$ presents a very long right tail, indicating that collectors can also find themselves isolated once other collectors in the network have completed their collections. For these isolated collectors, swaps are not possible which leads to a significantly larger waiting time. On the other hand, the left tail is displaced to the left, showing an important probability of cases more favourable than those under the standard CPP assumption. This is explained by the fact that collectors may exit the *Star network* as soon as they complete their collections, while those collectors inside a perfectly cooperating group must wait until all collections are completed before completing their own.

From table 6, it can be seen that the mean for the *Star network* is significantly higher compared to the distribution of $Z_{682,682}^{10,5}$. This behaviour can be explained by collectors exiting the *Star network*

Table 6: Comparison of Monte Carlo outputs $S^{10,5,10}_{682,682}$ with a standard version of CCP.

Characteristic	$S^{10,5,10}_{682,682}$	$Z^{10,5}_{682,682}$
Mean	1708.28	1560.758
Standard deviation	539.7282	135.1746
Median	1550	1541

early, leaving the remaining collectors fewer partners to swap with. Interestingly, the medians are similar for both distributions. The distribution of $S^{10,5,10}_{682,682}$ shows a large difference between the mean and the median, similar to the distribution of the *Circle network*. This is explained by a shape that is more skewed towards favourable cases for both networks. The increase in variance compared to the standard CCP is more pronounced for this network, as is the standard deviation under the *Star network* which is close to four times bigger than that of $Z^{10,5}_{682,682}$. The main reason for the increase in the variance is the very long right tail of the distribution.

4.2.3 Discussion on the limits of the network concept

Both networks presented in this section offer simple alternatives to the perfect cooperation assumed in the CCP, but they are not an actual depiction of the reality. Indeed, it is unlikely to find an organisation of collectors where each of them has exactly two swapping partners, or where all collectors are able to reach all other collectors of the network. Though the number of interconnections usually varies from one collector to another, this was not considered here.

This is only one of the flaws that characterise the two simple models presented in this paper. Ideally the model should account for many other parameters, such as collectors starting their collections at a different times, collectors buying packets of stickers at a different pace, collectors swapping at different times, and collectors swapping with friends in prority, etc.

It is virtually possible to create an infinite amount of different network models to analyse the relations between collectors. Determining which model would best represent the relationships between collectors will not be answered in this paper, as it would require the design of multiple models and study on the effects of each parameter added.

Although simple and imperfect, the *Circle network* and the *Star network* are definitively a better representation of the reality than the perfect cooperation considered in the standard CPP model. Notably, it reveals that the perfect cooperation strongly understates the volatility of the waiting time of collectors by assuming collectors cannot exit the cooperation. From the game theory, it is known that if an individual can obtain a direct gain from not cooperating, the cooperation is then unstable and likely to break. In this regard, the network modelling developed in this paper provides an approach better suited to the reality.

5 Brokering stickers

From the previous analysis, it appears that swaps or exchanges are a great way to lower the expected cost for collectors. In fact, the act of swapping with other collectors has been a tradition in existence since the very first World Cup album was published by Panini in 1970. Historically, swaps happened between friends or at special events where hundreds of collectors met to exchange stickers. More recently, with the surge in the popularity of social networks over the last decade, collectors turn to digital communications to swap with other collectors. Collectors can join specialized groups on Facebook and enjoy looking at the multiple ads for both "duplicates" and "wanted" stickers to look for what suits their needs. More specialized websites also exist, such as *Stickermanager* [Stickermanager, 2019]. These are platforms that allow collectors to virtually track the stickers they need and the duplicates they own and automatically find another user to swap with them. Once a swap is planned, the next step for both users is to simply send the stickers they agreed to exchange by post to the counter-party's address. Social media and internet platforms mean that in-person meetings are no longer mandatory to perform the swap. As these platforms allow a large number of users and swaps, they greatly benefit each collector by significantly increasing their chances of finding collectors to trade stickers with. This has the effect of lowering the waiting time for each collector on the platform.

Although it requires a few assumptions, it is possible to analyse the swaps taking place on a platform similar to *Stickermanager* using the *Star network*, detailed in the previous section. Indeed, the behaviour of collectors on the platform is essentially similar as they look for possible matches for a swap among the entire population of users. However, it is important to note that the *Star network* would imply that every user collects stickers at the same pace, which is assuredly not the case. Nonetheless, the *Star network* is a better approximation than the perfect cooperation model assumed in the standard CCP. Using this approximation, the following section will attempt to study the economic viability of platforms similar to *Stickermanager*.

5.1 Platform performance and profit

Assuming that a sticker-swapping platform behaves similarly to the *Star network*, it is possible to measure what is gained by collectors by comparing their expected cost using the platform to their cost without the platform. Another interesting measure is the number of swaps performed per user on the platform. Table 7 presents these measurements for platforms of various sizes.

Table 7: Platform performance approximated with $S^{n,5,10}_{682,682}$

Platform size (Number of users)	10	100	500	1000
Expected waiting time per collector	1708	968	810	778
Gain per collector (Compared with the single collector case)	3123	3864	4022	4054
Standard deviation of waiting time	540.22	236.52	135.25	68.31
Average number stickers swapped per collector	145	161	163	163

It can be seen in the platforms that collectors immediately experience a large gain by joining a platform, even if the size of the platform is rather small. The gains, however, increase only a little beyond a platform of 100 users. A comparable observation can be made on the average number of swaps per collectors, which remains almost unchanged beyond a platform size of 100. The most significant improvement is the volatility that steadily decreases as the number of users grows.

While gains for collectors are clearly evident, it is less straightforward what a platform gains by providing the service. The platforms need to generate revenue to cover the fixed costs required to maintain the platform online and accessible to collectors. There are two possible ways to generate revenue; either by obtaining money directly from the collectors or from a third party as advertising revenue.

When considering both options, the first option is less likely to be profitable since it (i.e., membership fee, commission on swaps, etc) can be seen as an entry cost for collectors to access the platform and collectors would lose interest or consider alternatives such as social media groups or events. The remaining collectors who would still be willing to pay to use it would most likely not be enough to ensure the platform's survival.

This leaves advertising as the main way to generate revenue for the platform, which would be related to the number of visits to the website. Let us assume here that collectors visit the platform on average a constant number of times, regardless of the size of the platform (this assumption is realistic since the average number of swaps does not heavily vary with the platform size). It then means that the number of visits is always a multiple of the number of users, which makes the number of users the more interesting metric. The website *Developereconomics* analysed that the average revenue per user through advertising for the common smartphone app amounts to 0.04$ per user [Wilcox, 2013]. Let us imagine now that a platform could comparably generate a revenue of 0.04€ (approximated to Euro for practical purposes) per collector using the platform.

The cost of setting up the platform and maintaining it online should also be considered. Assuming that the development costs required to design the platform are zero, then only remains the monthly cost of hosting the website on a server. The company *bluehost* provides hosting services for dedicated servers (dedicated servers are required for a large number of users and to ensure the stability of the platform) for a price of 72 € per month [Bluehost, 2019]. Monthly profits of the platform can be derived from calculating advertising revenue minus hosting costs as demonstrated in table 8.

Table 8: Platform monthly profit

Platform size (Number of users)	10	100	500	1000	2000	5000	10000
Advertising revenue (€)	0.4	4	20	40	80	200	400
Hosting costs (€)	72	72	72	72	72	72	72
Profit (€)	-71.6	-68	-52	-32	8	128	328

Note that this is only an estimation. Both advertising revenue and hosting costs may vary from this analysis. Specifically, numerous different pricing schemes exist for the hosting cost and strongly depend on the behaviour of users (i.e., do all users use the website at the same time?) and on the popularity of the platform. Therefore, finding the appropriate hosting services requires more research. The cost presented here is a large upper-estimate. However, using this large upper-estimate, it is already feasible to say that a low number of users on the platform (<500) would mean that the website would be unlikely to be profitable. The break-even point for profitability is around 2000 collectors on the platform. Still, 2000 users is not a realistic goal to achieve since the profit would be too small to compensate for the time and effort necessary to maintain and promote the platform. A realistic goal would lie between 5000 and 10000 users, where profit would start to sufficiently reward the investment.

To give an example, the platform *Stickermanager* accounts for about 230,000 users, probably thanks to which it achieves a decent level of profit. Most users of *Stickermanager* are located in Germany (around 126,000) and only a small fraction of the website user-base is located in Belgium, France and Netherlands (respectively around 2900, 1800 and 1300). This is probably explained by the design of the website, since it is only accessible in English or German. French alternative platforms exist, such as *LastSticker*. *LastSticker* is also available in many other languages, but currently only has around 60,000 users in Europe and less than 1500 in Belgium [LastSticker, 2019]. Therefore, a Belgian-launched platform could perhaps make a profit by capitalizing on the large number of French and Dutch speaking collectors available.

5.2 Stickers' value

The platforms like *Stickermanager* [Stickermanager, 2019] often provide metrics about the current supply and demand of each sticker in the collection, based on the information provided by the collectors using them. These metrics are very useful for collectors to identify if some stickers are rarer than others. Although all stickers have an equal probability to be picked, it is still possible that one sticker appears fewer than the others within a group of collectors. Using these metrics, collectors can choose to value and swap stickers based on their rarity (i.e, if a stickers is about 2 times rarer than the others, then it can be two time more valuable and it can be traded against two common stickers).

To study how value is distributed within a group of collectors, a value measurement has to be defined for each sticker. A sticker's value is determined based on supply and demand within the group of collectors:

$$\text{Sticker's value} = \frac{\text{Number of collectors needing this sticker} * 100}{\text{Number of times this sticker is willing to be exchanged as duplicate}}$$

For example, if two duplicates of a specific sticker are available and two collectors are in need of it, the value of the sticker will be 100. If only one collector needs this sticker, the value will instead becomes 50. Oppositely, if only one duplicate of this sticker is available and two collectors want it, the value will amount to 200. If a stickers is needed but no duplicate willing to be exchanged is available, the resulting value will be ∞.

Now that the value of a sticker is properly defined, the evolution over time of the stickers' values on a platform will be analysed. For this purpose, we will once again rely on the *Star network* Monte Carlo algorithm to obtain results. The algorithm was slightly altered to record the supply and the demand of each of the stickers before each second phase (as defined in section 4.2.2, the first phase consists in opening packets of stickers and the second phase consists in collectors swapping stickers) in order to compute the value relative to each sticker. The delay between phase 1 and phase 2 k has also been changed, from 10 to 15, mostly in order to reduce the computing time. This change should have little impact on the results.

In addition, this analysis will also study the relationship between stickers' values and the number of collectors using the platform by computing the values evolution for various number of users. Since obtaining the values within a large group of collectors is computationally intensive, and available resources are limited, only values for groups of size equal or below 100 were determined. Furthermore, the number of simulations to obtain the following results are limited to 10^2 per group size.

Figure 23 displays the results of the analysis. The figure compares at different *ages* of the collection process (represented by the number of packets each collected has opened) the values for three different sizes of platforms. It can be seen that the values heavily depends on the age of the collection since the majority of stickers have a value above 1000 or are not available after every collector opened 45 packets but the majority shift to either a value below 100 or even zero after each collector opened 150 packets. This observation remains the same for other platform sizes.

The main change brought by a larger number of collectors compared to a small group is regarding the values' dispersion which lowers as the number of collectors increases. Focusing on the group of 5 collectors, it can be seen that, regardless of the age of the collection, most stickers are either not available or worthless. Only a small fraction of the stickers have a value different of zero or ∞. On the contrary, the values determined within a group of 100 collectors show the least amount of not available stickers, and the least amount of worthless stickers at any time of the collection process.

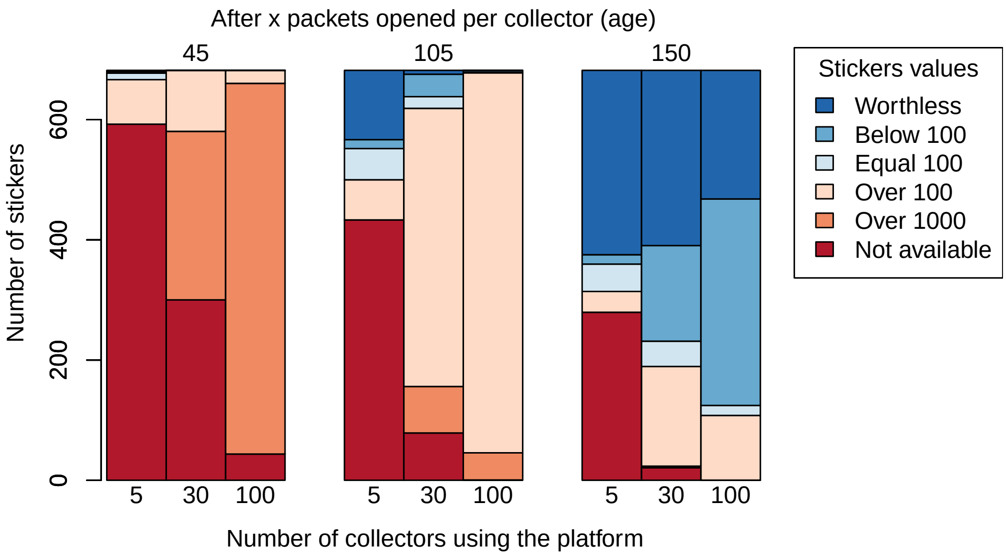

Figure 23: Stickers' values evolution. Each group of bars represents the values after $45, 105, 150$ packets have been opened per collector. Within a group of bars, each bar represents the values for a group of $n = 5, 30, 100$ collectors using the platform. Each colour is assigned to a range of sticker' values, labelled on the legend. "Worthless" indicates a value of 0. "Not available" indicates a value of ∞, since the sticker is not available as duplicate willing to be exchanged.

Since figure 23 represents the absolute stickers values, it does not inform about the relative values of theses stickers against each other. Indeed, if every sticker has a high value since collectors just started their collection, the value of each sticker might remain on par with the average stickers' value. To investigate precisely the dispersion of stickers' values relative to each other, they have to be compared to the average sticker value at the current time. Figure 24 improves the initial approach by comparing relative values of stickers. For example, if the average value of all stickers is 700, a sticker with a value of 1500 would fall in the *>200%* category, meaning it is two times more valuable than the average.

The inspection of figure 24 reveals how the number of collectors affects the values' stability. The brightest colour represents the number of stickers with a comparable value, which could be swapped in a 1 to 1 ratio. Violet coloured bars represent the stickers with a limited supply that have a superior value and are likely to be traded in a 2:3,1:2 or even 1:3 ratio. Inversely, green colours represent stickers with abundant supply and are more likely to be exchanged in 3:2, 2:1 or even 3:1 ratio. It can be seen that the number of stickers with a comparable value always increases as the number of collectors increases. Thereby, after each collector open 45 packets, the group of 30 collectors experience close to 40 % of the stickers being around the average value. For the group of 100 collectors, this amounts to 42 %. On the contrary, only 56 stickers, or 8 % show a comparable value in the group of 5 collectors.

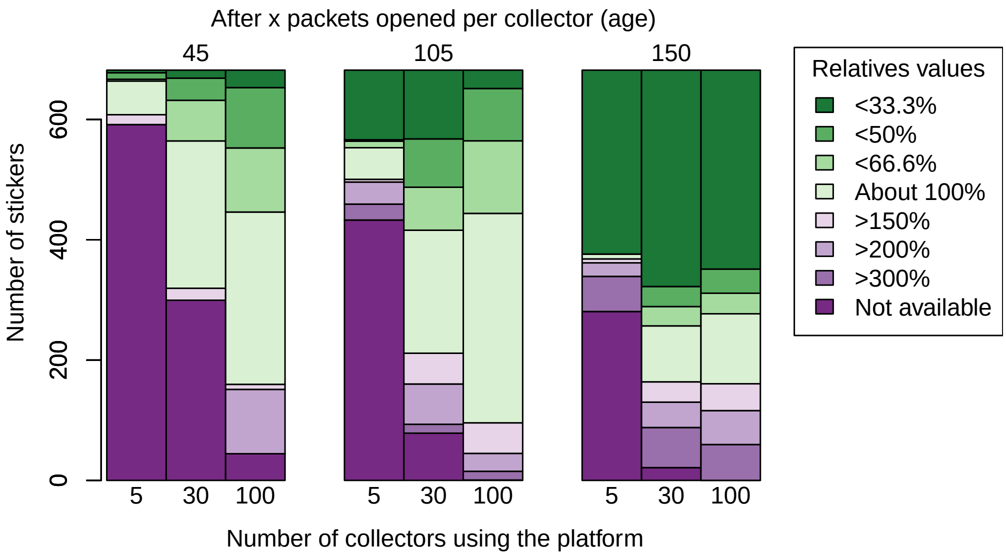

Figure 24: Stickers' relative values evolution compared to the average value. Each group of bars represents the values after 45, 105, 150 packets have been opened per collector. Within a group of bars, each bar represents the values for a group of $n = 5, 30, 100$ collectors. Each colour is assigned to a range of relative values, labelled on the legend. "About 100%" indicates a value between 66.6% and 150%. "Not available" indicates a relative value of ∞, since the sticker is not available as duplicate willing to be exchanged.

A similar observation can be made after every collector opens 105 packets, when the fraction of stickers having a comparable value even reaches 51 % for the platform of 100 collectors, but this lowers to 30 % for the group of 30 collectors. This decrease is explained by the increasing number of stickers that have been collected by most of the collectors of the group, leading to very small value for those stickers; hence the increasing size of the "<33.3%" band for the group of 30 collectors.

This last effect becomes more significant after each collector opened 150 packets, where stickers having a value below 33.3 % of the average are the most represented, regardless of the platform size. The platform of 100 collectors, however, retains the largest number of stickers comparably valued. If the number of stickers close to the average value is an indicator of the value stability within the platform, then it can be clearly stated that the larger the platform, the more stable are the stickers' values. Although impossible to further investigate due to technical constraints, it is expected that platforms of more than 100 collectors would present results in continuity with what was observed for smaller platforms : more stickers with a value close to the average. It is even predicted that for an infinite number of collectors, the values of all the stickers would be equal. This actually makes a lot of sense, since the stickers' values would correspond to their probability of being drawn : evenly distributed.

5.3 A note on the actual rarity of stickers

In the previous section, we just investigated how stickers would be valued on a platform. In fact, this analysis also pointed out how some stickers could be perceived as rarer than others. Even though results from the previous section were produced using an even probability of obtaining each sticker, figure 24 showed that a significant fraction of stickers could display an increased rarity (through a limited supply) even for large groups of collectors. Indeed, the group of 100 collectors always experienced at minimum 7 % of stickers with a value above 200 % or unavailable. This means that even with evenly distributed stickers, a significant portion of them can be perceived as much rarer.

This explains why a common discussion among collectors is about the rarity of each sticker. Although Panini asserts that the chance of drawing each sticker is equal, collectors often experience looking desperately for few missing stickers when they have multiple duplicates of others, leading them to question this affirmation.

A small analysis was performed to observe the probability to obtain a few occurrences of some stickers within a group of 10 collectors. A simple way to find the probability of not drawing a specific sticker after a certain number of packet opened k is using the formula $P_k = (\binom{681}{5}/\binom{682}{5})^k$. In order to find the probability of a sticker appearing exactly once time within a group of 10 collectors, reliance on the Monte Carlo algorithm is required. Monte Carlo can be used to simulate the opening of k packets, and then the counting the number of stickers that have been drawn exactly once, twice, etc. The results displayed in table 9 were obtained by repeating this process 10^5 times.

Table 9: Probability for one sticker to appear rarely

	0 time	1 time	2 times	3 times	4 times	5 times	≤ 5 times
780 packets	0.32 %	1.85 %	5.33 %	10.21 %	14.64 %	16.79 %	49.15%
1170 packets	0.02 %	0.16 %	0.68 %	1.96 %	4.21 %	7.25 %	14.28 %
1560 packets	0.001%	0.011 %	0.069 %	0.26 %	0.75 %	1.74	2.83 %
1940 packets	10^{-4} %	10^{-3} %	0.01 %	0.03 %	0.1 %	0.3 %	0.44%
2340 packets	10^{-6} %	10^{-5} %	10^{-4} %	10^{-3} %	0.01 %	0.04 %	0.06%
3120 packets	0 %	0 %	10^{-6} %	10^{-5} %	10^{-4} %	10^{-4} %	10^{-3} %

In table 9, the columns represent the probabilities for one sticker to appear exactly a certain number of times. The rows relate these probabilities to a number of packets opened. As an illustration, the probability of obtaining a sticker exactly two times after opening 780 packets is 5.33%. The probabilities exactly equal to 0 are assumed, since the Monte Carlo algorithm was not able to identify a single case of a sticker appearing only one or two times. The column on the right end represents the probabilities that a sticker appears 5 times or less. From the table, it can be seen that it is likely for a group of collectors to observe a sticker appearing 5 times or less early in the collection process (note that 3120 packets is the average waiting time for a group of 10 collectors, since the expected waiting time for one collector within the group is $1561/5 \approx 312$ packets, as exposed in section 3.2.4). However, the probability quickly drops as more packets are opened, and fall to 2.83% mid-way to the expected number of packets for completion. Beyond that point, it becomes highly unlikely that a sticker would appear 5 times or less within the group.

These probabilities are only associated to one particular sticker. A rough approximation to obtain the probability of two stickers appearing less than 6 times would be obtained by taking the squared probability of obtaining one sticker 5 times or less, i.e. $0.4915^2 = 0.2415$ after 780 packets. This approximation can also be used for $i = 3, 4, ..$ stickers at once by placing i at the exponent of the probability. This makes the chance of observing multiple stickers appearing rarely within the group the much less likely and would probably be directly related to the stickers' distribution.

This is an interesting result as some specific stickers have been rumoured to be rarer than others for the 2018 Word Cup collection. More specifically, the 50 *shiny* stickers have been reported to appear significantly fewer than expected. Considering the previous results, it would appear extremely unlikely that all the 50 *shiny* stickers appear rarer if they were evenly distributed as other stickers. The YouTube channel TopTradingCards [TopTradingCards, 2018] and Paulden [Paulden, 2018] have investigated this issue with a large sample of stickers. TopTradingCards opened 1000 packets and therefore obtained 5000 stickers, among which only 150 were *shiny*. This value is far from 366, the expected number of shiny stickers to be found if the stickers were evenly distributed. Paulden performed an extensive analysis on his sample of 11,000 stickers and also found out that *shiny* stickers were abnormally less represented than the others. He concluded that *shiny* stickers were close to two times rarer than the other stickers.

These proofs seem to indicate that the stickers of the 2018 World Cup collection are not evenly distributed. While this could invalidate most results presented in this paper, Paulden [Paulden, 2018] pointed out the limited impact on the collection cost when considering swaps between collectors and the ordering of stickers directly from the supplier, altering the resulting cost by about 3%. Thereby the results presented here remain useful for the 2018 World Cup collection but with possibly reduced accuracy. It has to be noted that the non-uniform distribution of stickers is a specific characteristic of the 2018 World Cup collection : Paulden cited a paper from the 2010 World Cup era concluding that their sample of 6,000 stickers was uniformly distributed [Sardy and Velenik, 2010]. Although Panini never mentioned anything about the *shiny* rarity for the 2018 World Cup collection, it was hopefully a single incident.

6 Log-normal approximation of the waiting time

In 1998, Read proposed a log-normal approximation of the waiting time of the basic CCP [Read, 1998] (i.e., as presented in section 2.1, without considering *Cooperation* and *Group drawing* extensions). In this paper, Read describes a methodology to approximate the distribution of the waiting time with a log-normal distribution by equating the first three moments of both distributions. The results of Read's work show that it is possible to obtain a very accurate approximation, with an error below 0.3 % for a CCP of 100 or more coupons.

The advantage of the log-normal approximation is that it has a simple continuous distribution that is readily usable and easy to compute. This is especially interesting considering the importance of computing time faced when simulating Monte Carlo outputs in sections 3 and 4. By storing the first three moments from of Monte Carlo outputs distributions, it is easily possible to re-create an approximated distribution without having to compute the Monte Carlo outputs again.

Read only proposed the log-normal approximation for the basic form of the CCP. Since the results of the CCP have been altered by the two extensions, it is necessary to verify the validity of the log-normal approximation for this extended case. In addition, Read's method will also be applied to the outputs of the *Circle network* and the *Star network*. Read's methodology was applied as follows (for the formulas used, please refer to the original paper [Read, 1998]):

1. Compute the mean, variance and skewness of the outputs.
2. Compute the corresponding mean μ and variance σ used to generate the log-normal, as well as a location-shift parameter c.
3. Generate the approximation, of which cumulative distribution function is $\Phi[(ln(r + .5 - c) - \mu/)\sigma]$.
4. Compare the original distribution to the approximation.

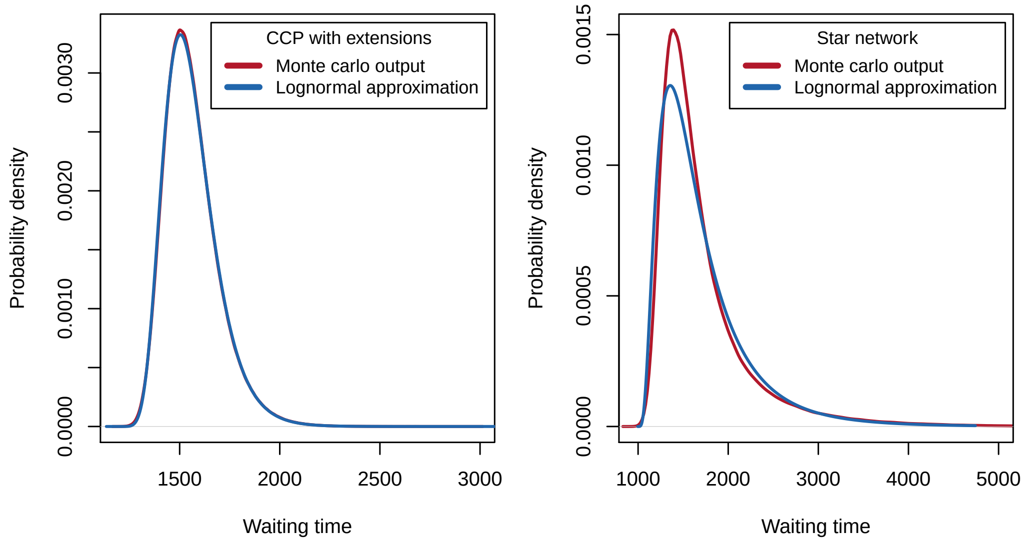

Figure 25: Approximation of the distributions of the CCP including *Cooperation* and *Group drawing* and the *Star network* by a log-normal distribution. Red lines represent the densities of the Monte Carlo outputs and the blue lines represent the corresponding approximation.

The results of the procedure for the CCP with *Cooperation* and *Group drawing* (outputs exposed in section 3.2.4) and for the *Star network* (outputs exposed in section 4.2.2) are presented in figure 25. Since the application of the procedure on both networks results in similar outcomes, the presentation of the *Circle network* results will be omitted.

The graph on the right of figure 25 shows a significant difference between the real distribution of the *Star network* and its log-normal approximation. The approximation is not able to replicate precisely the shape of the original distribution, which is especially noticeable on the peak part. In contrast, the approximation of the CCP almost perfectly coincides with the shape of the original distribution. Further comparison is presented in table 10 where quantiles between the original distribution and the approximation are compared.

Table 10: Accuracy of the log-normal approximation

	Quantiles	.05	.25	.5	.75	.95
CCP with extensions	original distribution	1378	1465	1541	1634	1811
	approximation	1379	1464	1541	1635	1811
	error (rounded)	1	-1	0	1	0
Star network	original distribution	1210	1380	1550	1850	2760
	approximation	1183	1356	1564	1891	2705
	error (rounded)	-27	-24	14	41	-55

The approximation of the extended CCP is relatively accurate, as the only errors line-up with those reported by Read. While absolute errors do reach one unit, this is a considerably small figure in regards to the actual quantile, showing a relative error of less than 0.1 %. On the other hand, the large errors reported for the *Star network* confirm the inadequacy of the log-normal approximation for this model. Although the approximated quantiles remain within 3 % of the original values, the difference is significant enough to justify new computations to obtain the real distribution.

While using the log-normal approximation dispenses the need to compute Monte Carlo outputs to obtain a distribution, Monte Carlo is still required to obtain the moments used in the approximation. Only analytical results for the first three moments could allow the production of an approximated distribution without the use of Monte Carlo. In the absence of analytical results, an attempt was made to fit a non-linear regression that could approximate the moments of the CCP when varying the parameters of the model. Appendix A presents the first three moments of computed Monte Carlo for various numbers of cooperating collectors and collection sizes. Unfortunately a function that would correctly approximate any of the moments was not found at this time.

7 Conclusion

The analysis conducted in this thesis has covered a broad range of topics, from the validation of the literature results to the profitability of setting up a sticker swapping platform. Numerous findings have been identified, and this conclusion outlines the most relevant results from the perspective of a collector, the supplier (Panini), and a swapping platform.

The interesting results of this paper for a collector are mostly related to the collection cost. The cost of completing a collection is heavily influenced by two parameters: the number of collectors to swap with and the number of stickers directly ordered from the supplier. Swapping or trading with other collectors is extremely beneficial to each collector cooperating: it was identified that a single collector would on average need to obtain 4832 stickers, or 966 packets before completing his collection, while a group of 10 collectors cooperating perfectly would lower these numbers to 1561 stickers, or 312 packets. Even when relaxing the perfect cooperation hypothesis and using the proposed model of *Star Network*, considering 10 collectors as independent entities, the expected number of stickers would only rise to 1708, or 342 packets, remaining close to one-third of what is required for an isolated collector.

The impact of the stickers directly ordered from the supplier is also significant. The concept of a collector's strategy was presented that once reaching on the shift point at which a collector should stop buying packets of stickers and instead order stickers directly from the supplier at a premium price. It was found that an optimal shift point exists, minimizing the expected cost for the collectors. Unfortunately, Panini only allows a limit of 50 for the number of ordered stickers, which most of the times prevents collectors from applying their optimal strategy. Nonetheless, ordering the 50 last stickers still reduces the standardised expected cost (cost is standardised when the price of each packet is 1) of a single collector to 377, less than half of the cost of 966 for a single collector not ordering any stickers. Ordering stickers in a group of 10 cooperating collectors further lowers the standardised cost to 177.

A strategy aiming to reduce the risk that the collection's cost goes beyond a certain threshold has also been developed. Although it is possible to optimise the risk by slightly altering the optimal strategy of the collectors, this is often inapplicable because of the limit of 50 ordered stickers imposed by Panini.

From the point of view of the supplier (Panini), interesting results have been displayed about the characteristics of the collection. The supplier has control over the collection size, the number of stickers inside a packet, the pricing of stickers and the number of stickers each collector can order. Varying these parameters allows Panini to change the expected cost of collectors and thereby to increase or decrease profit of the company. It was seen that an increased collection size is always beneficial to Panini since the marginal profit per sticker added to the collection always increases. Although Panini affirms that packets cannot contain duplicates, this has no significant impact on the collection cost of collectors, indicating this could possibly come from their marketing strategy. Furthermore, the chance of obtaining duplicates naturally within a packet of 5 stickers is only of 1.46%, which could question whether Panini actually ensures that no duplicates end up in the same packet.

It also came out that there is little change in the profit when varying the limit imposed by Panini for ordered stickers, unless the limit is drastically reduced or suppressed, which would, however, raise the anger among collectors. Another way for Panini to generate additional profit more discretely would be by increasing the price at which ordered stickers are sold. Indeed, if Panini raised this price to 60% of one packet of stickers (from the current price of 44.4%), collectors would apply an identical strategy but their standardised expected cost rises to 185 compared to 177 for the group of 10 collectors. This change would directly increase the revenue of Panini by 7% while remaining unnoticed by collectors.

The last perspective investigated was the economic interest of setting an online sticker-swapping platform. It was estimated that a platform could generate a lot of wealth for the participating collectors, by allowing them to reach a large number of others collectors to swap with. In exchange, the platform can generate profit through advertising thanks to a large number of users visiting the website. Using rough estimates for the revenues and cost of the platform, a break-even point for a platform registering more than 2000 users was determined. The actual number of users required to make the platform generate enough money to be interesting is between 5000 and 10000.

Platforms can also provide interesting metrics that can help collectors to value the stickers accurately. The absolute value of stickers, determined by the supply and the demand on the platform, tends to lower as collectors accumulate more stickers, while the dispersion of the stickers' values decreases as more collectors are present on the platform. This indicates that for a platform with a very large number of collectors, the value of each sticker would be close to equal.

In addition to these three perspectives, this thesis provided numerous results that enhance the existing literature. Thereby, the inaccuracy of the asymptotic developments presented by Erd and Ri [Erd and Ri, 1961] for the completion of n collections was pointed out. The formulas developed by Stadjes [Stadje, 1990] were accurately reproduced with the use of Monte Carlo algorithms. The interesting case of a collection with a number m of stickers inside a packet and a number n of cooperating collectors is unexplored in the literature. For this case, this thesis provides substantial numerical approximations and extends the concept of strategy from Stadjes.

Furthermore, the assumption of perfect cooperation was compared to the real interactions between collectors, and it was concluded that this assumption was far from reality. Network models were proposed to replace the cooperating assumption. In these models, each collector is considered independent and leaves the network as soon as he finishes his collection. The outputs of Monte Carlo algorithm produced a distribution significantly different from what was studied using the perfect cooperation assumption, notably by a left tail shifted to the left and a longer and fatter right tail. The *Star network* disclosed that the cooperation assumption slightly underestimates the cost for collectors and heavily underestimates the volatility of the cost, which is about 4 times larger in the *Star network* configuration.

Finally, the log-normal approximation first presented by Read [Read, 1998] was applied to the waiting time of the completion of the 2018 World Cup collection, using both the standard model and the *Star network*. It was shown that the results of the *Star model* are not able to be accurately approximated by a log-normal distribution, while the approximation of the standard model had a relative error of less than 0.1 %.

Concerning the limits of the analysis presented, it can be pointed out the limited number of simulations used in the Monte Carlo algorithms. Ideally, the number of simulations required would be determined in accordance with a confidence interval around the estimated value. This step was skipped in this analysis since there is little room for adjusting the number of simulations due to technical limitations.

A second limit would be the lack of results for a dynamic strategy. A dynamic strategy would consider the current state of the collection for one or multiple collectors and provide an expected cost accounting for what has already been spent. This strategy could allow collectors to efficiently react whenever they absolutely want to keep the cost of their collection under a certain threshold. Future research could focus on the development of a dynamic strategy by introducing new metrics to assess the collection state, which is the limiting factor when more than one collections are considered.

Third, the networks exposed in section 4.2 were chosen arbitrarily and are not the best representation possible of real exchanges between collectors. Further work could be interesting in studying other types of networks and how they affect the distribution of the waiting time. Moreover, the modelling of collectors joining, leaving the network and being replaced would be invaluable by creating models closer to the real application.

Finally, the most impacting limit of this paper is regarding the distribution of the stickers in the 2018 World Cup. When the preparation of this thesis started, stickers were assumed to be uniformly distributed (i.e., there is an equal chance to draw any of them). Although the entire analysis conducted here followed this assumption, some references have proved that some stickers are rarer than others and thus that the distribution of stickers is non-uniform [TopTradingCards, 2018],[Paulden, 2018]. Without this assumption, most of the results presented in this work are not accurate for the 2018 World Cup collection, showing a little difference due to the wrong stickers' distribution. The methodology presented remains however appropriate, and the inclusion of a different stickers' distribution in Monte Carlo algorithms can be easily performed for subsequent research.

Hopefully this thesis provides an useful introduction to the numerical analysis as an answer to the complex forms of the *Coupon Collector's Problem* and offers meaningful insights for enthusiastic collectors, Panini, and swapping platforms.

Bibliography

Bluehost. (2019). Dedicated web server hosting services - dedicated solutions - bluehost. Retrieved August 7, 2019, from https://www.bluehost.com/hosting/dedicated

Doumas, A. (2015, January 1). How many trials does it take to collect all different types of a population with probability p ? *Journal of Applied Mathematics & Bioinformatics*, *5*, 1–14.

Erd, P. & Ri, A. (1961). On a classical problem of probability theory, 6. Retrieved August 1, 2018, from https://users.renyi.hu/~p_erdos/1961-09.pdf

Ferrante, M. & Saltalamacchia, M. (2014, May 25). The coupon collector's problem. *MATerials MATemàtics*, *2014*, 35. Retrieved August 1, 2018, from http://mat.uab.cat/matmat/PDFv2014/v2014n02.pdf

Flajolet, P., Gardy, D., & Thimonier, L. (1992, November). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, *39*(3), 207–229. doi:10.1016/0166-218X(92)90177-C

Hald, A., de Moivre, A., & McClintock, B. (1984). A. de moivre: 'de mensura sortis' or 'on the measurement of chance'. *International Statistical Review / Revue Internationale de Statistique*, *52*(3), 229–262. doi:10.2307/1403045

Laplace, P.-S. d. (.-1. A. d. t. (1878). *Oeuvres complètes de laplace. tome 8 / publiées sous les auspices de l'académie des sciences, par MM. les secrétaires perpétuels*. Retrieved May 14, 2019, from https://gallica.bnf.fr/ark:/12148/bpt6k77596b

LastSticker. (2019). Echanger des autocollants panini, système d'échange gratuit d'autocollants et de cartes à collectionner pour albums panini, topps, merlin. Retrieved August 7, 2019, from https://www.laststicker.com/fr/

Newman, D. J. (1960, January). The double dixie cup problem. *The American Mathematical Monthly*, *67*(1), 58. doi:10.2307/2308930

Paulden, T. (2018, July 19). ShinyGate: New data shows shiny world cup 2018 stickers really are rarer. Retrieved from https://atass-sports.co.uk/images/uploads/case-study/ShinyGate-Final.pdf

Read, K. L. Q. (1998). A lognormal approximation for the collector's problem. *The American Statistician*, *52*(2), 175–180. doi:10.2307/2685477

Sardy, S. & Velenik, Y. (2010). Paninimania: Sticker rarity and cost-effective strategy, 5. Retrieved August 1, 2018, from https://www.unige.ch/math/folks/velenik/Vulg/Paninimania.pdf

Stadje, W. (1990). The collector's problem with group drawings. *Advances in Applied Probability*, *22*(4), 866–882. doi:10.2307/1427566

Stickermanager. (2019). Stickermanager. Retrieved August 1, 2019, from https://www.stickermanager.com/en

TopTradingCards. (2018, May 20). 1000 PACKS (5000 STICKERS) panini world cup 2018 GLITZER STICKER CHECK [YouTube]. Retrieved August 1, 2019, from https://www.youtube.com/watch?v=XNVNb69Q0Cg

Wilcox, M. (2013, April 1). How much is an active user worth? [DeveloperEconomics]. Retrieved August 7, 2019, from https://www.developereconomics.com/how-much-is-an-active-user-worth

Appendix A : First three moments of $Z_{s,s}^{n,5}$

n		$s = 10$	$s = 50$	$s = 100$	$s = 500$	$s = 682$	$s = 1000$	$s = 2000$
1	Mean	24.63	217.75	509.95	3384.75	4828.95	7476.69	16350.21
	Sd	8.38	59.44	122.53	634.14	870.01	1280.33	2558.11
	Skewness	1.406	1.203	1.150	1.126	1.145	1.156	1.147
2	Mean	19.95	157.95	359.46	2266.27	3209.88	4919.37	10619.03
	Sd	5.02	33.86	69.89	350.59	476.69	697.48	1392.24
	Skewness	1.256	1.112	1.147	1.107	1.115	1.103	1.101
4	Mean	16.73	120.15	266.51	1607.42	2259.74	3437.19	7312.20
	Sd	3.04	19.95	40.33	199.92	271.02	396.32	781.56
	Skewness	1.123	1.044	1.032	1.065	1.060	1.061	1.060
6	Mean	15.38	104.96	229.57	1351.31	1893.66	2867.27	6051.72
	Sd	2.32	14.80	29.86	144.63	197.28	286.95	567.52
	Skewness	1.087	1.006	0.989	0.978	1.015	0.998	1.032
8	Mean	14.59	96.42	208.86	1210.83	1691.36	2552.74	5361.70
	Sd	1.90	12.10	24.25	118.17	159.39	230.31	454.94
	Skewness	1.030	0.947	0.938	1.002	0.998	1.004	1.009
10	Mean	14.05	90.78	195.35	1118.83	1561.14	2349.94	4917.82
	Sd	1.64	10.37	20.70	99.55	135.00	196.11	384.47
	Skewness	0.987	0.934	0.934	0.954	0.969	0.979	1.000
12	Mean	13.67	86.69	185.73	1053.79	1467.95	2206.30	4604.03
	Sd	1.46	9.12	18.38	87.22	118.89	171.58	337.80
	Skewness	0.963	0.895	0.922	0.934	0.989	0.947	1.020
14	Mean	13.38	83.62	178.38	1004.13	1398.09	2098.44	4367.90
	Sd	1.33	8.30	16.53	78.42	106.75	154.90	300.88
	Skewness	0.951	0.922	0.912	0.954	0.971	0.981	0.958
16	Mean	13.14	81.11	172.54	965.89	1342.66	2012.71	4182.14
	Sd	1.22	7.53	15.12	71.42	96.95	139.14	274.97
	Skewness	0.966	0.872	0.923	0.926	0.944	0.945	0.975
18	Mean	12.95	79.17	167.75	934.78	1297.52	1944.03	4030.92
	Sd	1.13	7.02	13.87	66.23	89.69	129.57	252.33
	Skewness	0.929	0.873	0.886	0.938	0.958	0.962	0.956
20	Mean	12.79	77.45	163.75	908.75	1260.22	1886.78	3906.41
	Sd	1.06	6.51	12.97	61.50	82.83	120.56	233.91
	Skewness	0.923	0.859	0.884	0.914	0.931	0.972	0.987