Effort Analysis Using Collective Stochastic Model

Vugranam C. Sreedhar

IBM TJ Watson Research Center,
Yorktown Heights, NY, 10598, USA
vugranam@us.ibm.com

Abstract. In this paper we consider the problem of work order (WO) arrivals and time spent on work orders in service delivery to derive the asymptotic behavior of a strategic outsourcing contract. We model both the work order arrivals and time spent on the work orders, also known as effort, as a collective stochastic process. We use the resulting model to derive the probability that a contract will exceed the allocated budget for resolving work orders, and also to calculate the staffing requirement for resolving work orders.

Keywords: Collective Stochastic Model, Poisson Process, Renewal Process, Workload, Effort, Service Delivery.

1 Introduction

Strategic outsourcing (SO) happens when one company outsources part of its business to another company. A service provider and a service consumer negotiate a contract that outlines different kinds of work that needs to be done in terms of managing the consumer's business. A strategic outsourcing company, such as IBM, manages Information Technology (IT) infrastructure and applications for many different companies. A breach of contract happens when services are not delivered as negotiated in the contract. Very often, even when services are delivered that are in par with what is negotiated in the service level agreements (SLAs), a service consumer can quickly become unhappy when things go wrong. There are many reasons why a contract can become troubled or risky, incurring loss to a service provider. A service provider strives very hard to provide services that will increase profitability, customer loyalty and customer value. An SO contract often include SLAs that when violated, the service consumer can impose penalty on the service provider.

A large service provider, such as IBM, have service delivery centers to manage several customers. The management of IT of a customer is broken down into different kinds of work orders (WOs). A work order can be as simple as a request to change someone's password to as complex as migrating 100 physical servers (along with the applications) to a cloud environment. Very often complex WOs are broken down into smaller WOs that are easy to track and manage. Different WOs take different amount of time to resolve. A key question is then to ask is:

X. Franch et al. (Eds.): ICSOC 2014, LNCS 8831, pp. 502–509, 2014.
© Springer-Verlag Berlin Heidelberg 2014

How much time (or effort) is needed, and hence how many full time employees (FTEs) are needed to resolve work orders, say in a month or a year?

In this article we develop a collective stochastic model (CSM) to determine the total time or *effort*, and hence the number of FTEs, needed to resolve work orders over certain time period such as a month or a year. The main contribution of this paper is to apply the well established theory of collective stochastic process model, and in particular ruin theory developed in actuarial science, to model services delivery system [7]. Modeling services delivery system is a non-trivial exercise, and developing mathematical models will allow future researchers to optimize and gain deeper insights into the complex behavior of services delivery system. To the best of our knowledge, ours is the first comprehensive attempt to leverage concepts from actuarial science and ruin theory to model portions of services delivery system, and in particular, to model effort, contract loss probability, and staffing requirements.

2 Collective Poisson Model

Work orders arrive one at a time and each work order is independent of each other. Let $\{N(t), t \geq 0\}$ denote the number of work orders that was processed before time t. We assume that $N(0) = 0$, and $N(t) \geq 0, \forall t \geq 0$. In other words, there are no work orders before $t = 0$, and there cannot be negative number of work orders. Therefore, $N(t)$ is non-decreasing in t. For $s < t$, we also have $N(t) - N(s)$ equals the number of work orders in the time interval $(s, t]$. We can now define the nth work order arrival as $T_n = \inf\{t \geq 0 : N(t) = n\}$ and the inter-arrival time of work order as $A_n = T_n - T_{n-1}$. The model described above captures the basic set of assumptions needed to describe a work order arrivals. It is important to keep in mind that $N(n)$, T_n, and A_n are all random variables and for $n \geq 0$, they form a stochastic process.

A (homogeneous) Poisson process is a very simple stochastic process that has two important properties: independence property and stationary property. The independence property states that for $\forall i, j, k, 0 \leq t_i \leq t_j \leq t_k, N(t_j) - N(t_i)$ is independent of $N(t_k) - N(t_j)$. In other words, the number of events in each disjoint interval are independent of each other. The stationary property states that $\forall s, t, 0 \leq s < t, h > 0$, $N(t) - N(s)$ and $N(t + h) - N(s + t)$ have the same distribution.

A homogeneous Poisson is too restrictive when we include the time it takes to resolve a work order. We next assume that the time it takes to resolve a work order, that is, the *effort*, itself is a random variable. We use Collective Poisson Process to model the aforementioned situation. A stochastic process $\{X(t), t \geq 0\}$ is called a collective Poisson process if it can be represented as follows:

$$S(t) = C_1 + C_2 + \ldots C_{N(t)} = \sum_{i=1}^{N(t)} C_i, t \geq 0 \tag{1}$$

where $\{N(t), t \geq 0\}$ is a Poisson process and $C_1, C_2, \ldots C_{N(t)}$ are iid random variables and are independent of $\{N(t), t \geq 0\}$. Here C_i represents the effort or time spent on a work order. The total effort during the period $(0, t]$ is then given by $S(t)$.

3 Renewal Process Model

In this section we extend the Poisson process by assuming the inter-arrival times for work order arrival using *Renewal Process* [8]. Let $\{A_n, n > 0\}$ be a sequence of random variable representing the inter-arrival times of work orders, and let $T_{n+1} = T_n + A_n$ be the arrival times of work orders. We define a renewal process for $\{N(t), t \geq 0\}$ so that

$$N(t) = \max\{i \geq 0 : T_i \leq t\} \tag{2}$$
$$= \min\{i \geq 0 : T_{i+1} > t\} \tag{3}$$

To ensure that the work orders do not all collapse at $A_i = 0$, we also assume that $P(A_i = 0) < 1$. Once again we assume both the independence and stationary properties for work order arrivals. It can easily be shown that $N(t)$ as defined by Equation 2 cannot be infinite for some finite time t [8]. In renewal process the inter-arrival times A_n is distributed with a common distribution function F_A, and $F_A(0) = 0$ and $T_n = 0$. The points $\{T_n\}$ are called the renewal times. Notice that the function F_A is a Poisson distribution function for Poisson process. Let us assume that the distribution function F_A has mean μ, one can then show the following result:

$$\lim_{t\to\infty} \frac{N(t)}{t} = \begin{cases} \mu^{-1}, & \text{if } \mu < \infty \\ 0, & \text{if } \mu = \infty \end{cases} \tag{4}$$

Recall that with collective Poisson process it was simple to derive a model for aggregated work order (see Equation 1). On the other hand it is almost impossible to determine the distribution F_A for renewal process $\{N(t), t \geq 0\}$. So we use the *central limit theorem* to get an approximate work order distribution. Let $0 < Var[A_i] < \infty$ and $\mu = E[A_i]$, then $\forall x \in \Re$

$$\lim_{t\to\infty} P\left(\frac{N(t) - t\mu^{-1}}{\sqrt{ct}}\right) = \Phi(x) \tag{5}$$

where $c = \mu^{-3} Var[A_i]$, and $\Phi(x)$ is the standard normal distribution function. The above results allows us to look for $E[N(t)]$ for which we can use *renewal function*. We then define the renewal function as the average number of renewals in the interval $(0, t]$ as $M(t) = E[N(t)] + 1$.

Let $F^{(k)}$ denote the k-fold convolution of F_A, which is the underlying distributions of the renewal process $\{N(t)\}$. Since $\{N(t) \geq k\} = \{A_k \leq t\}$ for

$k = 1, 2, \ldots$ we can derive the following result relating the mean value and the distribution.

$$\begin{aligned} M(t) &= 1 + \sum_{k=1}^{\infty} P(N(t) \geq k) \\ &= 1 + \sum_{k=1}^{\infty} P(A_k \leq t) \\ &= \sum_{k=0}^{\infty} F_A^{(k)}(t) \end{aligned} \tag{6}$$

The mean or the expected number of renewals $M(t)$ is a non-decreasing and continuous on $\Re$ and it uniquely determines the distribution F_A. The renewal function for Poisson process is $\lambda t + 1$. We can now extend the collective Poisson process model (Equation 1) to collective renewal process model by assuming $N(t)$ is a renewal process. In actuarial science, the collective renewal process is often called as the Sparre Anderson Model [9].

4 Effort Size Distribution

In this section we will address the random nature of work order effort size. Recall that when a system administrator (SA) works on a work order, he or she will spend some amount time to resolve the issue related to the work order. The amount of time spent on a work order, called the *effort*, is itself a random variable. The effort size depends on various factors including the complexity of the work order, SA experience, etc. To simplify the presentation we will assume *effort* to include all of these marginal costs, and use the term *effort size* to be the representative random variable.

We will focus on two kinds of distributions for effort size. First one is the Light-Tailed Distribution (LTD) and the second one is the Heavy-Tailed Distribution (HTD). The tail of a distribution $F(x)$ is defined as $\bar{F}(x) = 1 - F(x)$, which is nothing more than the upper part of the distribution. It is the tail of the distributions that dictates that governs both the magnitude and the frequency of extreme events. The light-tail distribution has more "mild" form of extreme events, whereas the heavy-tail distribution has more "heavier" form of extreme events.

A distribution $F(x)$ is called a light-tailed distribution if there exits constants $\lambda > 0$, $a > 0$ so that $\bar{F}(x) \leq ae^{-\lambda x}$. Light-tailed distribution have "nice" properties that do not put service delivery in greater risk of contract loss when claim size exceeds the budgeted. Exponential distribution with $\lambda > 0$, Gamma distribution with $\alpha > 0, \beta > 0$, and Weibull distribution with $\beta > 0, \tau \geq 1$ are some examples of light-tailed distribution [8].

A distribution $F(x)$ is called a heavy-tailed distribution if there exits constants $\lambda > 0$, $a > 0$ so that $\bar{F}(x) > ae^{-\lambda x}$. We can also express heavy-tailed (and hence light-tailed) distributions using properties of moment generating functions.

A distribution function $F(x)$ is a heavy-tailed distribution if its moment generating function $M_x(t) = E[e^{tx}]$ is infinite $\forall t > 0$. Pareto distribution with $\alpha > 0, \lambda > 0$ and Weibull distribution with $\beta > 0, 0 < \tau 1$ are examples of heavy-tailed distribution. Even though claim size of work orders cannot be infinite, it is possible for claim sizes to exceed the budget size, which can eventually lead to troubled contracts.

5 Contract Loss Probabilities

In the previous two sections we developed models for WO arrivals and WO effort. In this section we will combine the two models to calculate *the probability that a contract will exceed the allocated budget for resolving work orders.*[1] The Contract Loss Probability (CLP) gives a good indication of the health of a contract. This quantity can be used for staffing decision, resource allocation, staff training, and work order dispatch optimization.

In a typical SO contract during engagement phase, the customer environment is "discovered" and "analyzed" for sizing the cost of the contract. Various factors, such as the number of servers, types of servers, number of historical tickets that were generated and resolved, management process, etc., are used to determine the cost of the contract. A typical cost model include unit price such as cost per server per month. The way these unit prices are computed is more of an art than science. Productivity factors, market competition, economy of scale and other external factors are also incorporated into the pricing or cost model. Once a contract is signed, service provider allocate quarterly or monthly budget for different services of the contract and when operational cost exceeds the allocated budget, the contract is considered to be "troubled" and management systems are put in place to track the services.

Let us assume that each client account has a periodic (say, quarterly) budget $q(t) = rt$, which is the budget rate, and so $q(t)$ is deterministic. We can then define the following *contract loss process*: $Z(t) = a + rt - S(t), t \geq 0$, where a is some initial base budget allocated for resolving work orders. We can see that if $Z(t) < 0$ for some $t \geq 0$, then we have a contract loss for that time period, that is, effort spent exceeds the allocated budget for resolving work orders. Assuming collective Poisson process, a minimum requirement in determining the contract budget rate r is then given by $r > \lambda E[S]$, where λ is the Poisson WO arrival rate. The above condition is called the net profit condition. A safer condition would be to include a safety factor ρ, so that $c > (1 + \rho)\lambda E[S]$.

We can define the contract loss time as $\tau_0 = inf\{t \geq 0 : S(t) > 0\}$, and the contract loss probability as $\phi(z) = P(\tau_0 < \infty | S(0) = z) = P_z(\tau_0 < \infty)$. If we assume that $X(t)$ is a collective Poisson process, we can then calculate the contract loss probability $\phi(z)$ as a closed form solution by focusing on the tail

[1] It is important to keep in mind that a contract will allocate budget for different activities, and resolving work order is one of the major activities of a contract. In this article we will just focus on budget for resolving work orders.

end of the claim size distribution. Let $\psi(t) = 1 - \phi(t)$ denote the tail of the contract loss probability, then

$$\psi(t) = \frac{\theta}{1+\theta} \sum_{n}^{\infty} \frac{1}{(1+\theta)^n} F^{*(n)}(t), t \geq 0 \tag{7}$$

where $F^{*(n)}$ is the n-fold convolution of the distribution function $F(x)$, and $\theta = (\frac{r}{\lambda\mu} - 1)$, $\mu = E(C_i)$, r is the budget rate, and λ is the Poisson arrival rate of the work orders. Now when the effort sizes are (light-tailed) exponentially distributed $P(C_i > c) = e^{-c/\mu}$, we can derive the following contract loss probability:

$$\psi(t) = \frac{1}{1+\theta} exp\left(-\frac{\theta}{(1+\theta)\mu} t\right), t \geq 0 \tag{8}$$

Notice that we made two assumptions when deriving the above contract loss probability: (1) work order arrivals follows a Poisson process, and (2) effort or time spent on work orders follows (light-tailed) exponential distribution.

6 Pricing and Staffing Requirements

A key problem in service delivery is determining the staffing requirement for handling work orders. We make a simplifying assumption that a staff or a system administrator can work one work order at a time, with no multi-tasking or context switching. Let $\Pi(S) \in \Re$ denote the budget, and hence staffing requirement, to handle work order effort S. We can then identify the following properties for calculating the staffing budget for an account:

1. $\Pi(S) \geq E[S]$. In this case we have nonnegative effort loading.
2. If S_1 and S_2 are independent, then $\Pi(S_1 + S_2) = \Pi(S_1) + \Pi(S_2)$
3. $\Pi(aS) = a\Pi(S)$, and $\Pi(S + a) = \Pi(S) + a$.
4. Let M be the finite maximum effort, then $\Pi(S) \leq M$.

There are several methods for calculating the staffing budget. The Expected Value principle can be stated as follows [6]: $\Pi(S) = (1 + a)E[S]$, where a is a safety loading factor. The expected value budget is very simple, but it does not take into account the variability in the effort. We can extend this model to include variability as follows: $\Pi(S) = E[S] + aVar[S]$

One issue with the above Variance principle is that different delivery center may have custom staffing budget, depending on local labor policy, pay scale, monetary values, etc. To handle such changes to loading factor, we can use the following modified Variance principle: $\Pi = E[S] + a\frac{Var[S]}{E[S]}$

7 Discussion and Related Work

Our focus in this paper is not to develop a new compound stochastic process model, but to apply concepts from ruin theory in actuarial science for modeling IT service delivery system, and in particular to model "effort" needed to manage a customer IT environment, and to understand under what condition a contract can become troubled. To the best of our knowledge, ours is the first work that models IT service delivery leveraging ruin theory from actuarial science. A lot more work is needed to fully model IT services delivery system. Please refer to the technical report that explains in details on modeling effort, contract loss probability, and staffing requirements, beyond what is explained in the current article [10].

IT service delivery is a complex process with many intricate processes, management systems, people's behavior, and tool sets. Diao et al. proposed a modeling framework for analyzing interactions among key factors that contribute to the decision making of staffing skill level requirements [3,4]. The authors develop a simulation approach based on constructed and real data taking into consideration factors such as scheduling constraints, service level constraints, and available skill sets. The area of optimal staffing with skill based routing is a mature area. Analytical methods are typically complex and do not capture full generality of real IT service delivery systems. The main focus of our paper is not to model the full generality of IT service delivery system. We focus on developing a compound stochastic process model to model effort needed to handle service requests. We focus on understanding the underlying stochastic model for when a contract can become "troubled".

Staffing problem based on queuing theory is old problem and several solutions have been proposed to model in the past. The staffing problem can be simply stated as the number of staff members or agents required to handle work orders, such as calls in a call center, as a function of time. Skill based routing problem is an extension of staffing problem where skills set are incorporated to determine which staff skill is needed as a function of time [5]. Staffing problem are typically modeled a queuing problem rather than as a compound stochastic process. Coban models staffing problem in a service center as a multi-server queuing problem with preemptive-resume priority service discipline and uses Markov chain to model [2].

Buco et al describe a method where in they instrument a management system to capture time and effort when SAs work on work orders [1]. They collect this information from multiple SAs working on different kinds of WOs. The collected data is a sample of the universe of IT service environment. One can use the sampled data to estimate the staffing requirement of a contract.

8 Conclusion

IT services delivery system is a complex system. There has been very little work done to model such a system, mostly due to lack of mathematical maturity in

this field. Fortunately, actuarial science and ruin theory provides a foundational mathematics that can be applied to modeling IT services delivery system. We have made several simplifying assumptions such as WOs are independent of each others, all WOs are the same, etc. We are currently refining the mathematics to relax some of these simplifying assumptions. The resulting analytical model will become even more complex, and so can use a combination of estimators and Monte Carlo simulation for understanding the asymptotic behavior of a contract.

References

1. Buco, M., Rosu, D., Meliksetian, D., Wu, F., Anerousis, N.: Effort instrumentation and management in service delivery environments. In: International Conference on Network and Service Management, pp. 257–260 (2012)
2. Coban, E.: Deterministic and Stochastic Models for Practical Scheduling Problems. Ph.D. thesis, Carnegie Mellon University (2012)
3. Diao, Y., Heching, A., Northcutt, D., Stark, G.: Modeling a complex global service delivery system. In: Winter Simulation Conference, pp. 690–702 (2011)
4. Diao, Y., Lam, L., Shwartz, L., Northcutt, D.: Sla impact modeling for service engagement. In: International Conference on Network and Service Management, pp. 185–188 (2013)
5. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review and research prospects. Manufacturing and Service Operations Management 5(2), 79–141 (2013)
6. Geiss, C.: Non-life insurance mathematics (2010), `http://users.jyu.fi/~geiss/insu-w09/insurance.pdf`
7. Rolski, T., Schmidli, H., Schmidt, V., Teugels., J.: Stochastic Processes for Insurance and Finance. Wiley (1999)
8. Ross, S.: A First Course in Probability. Pearson Prentice Hall (2006)
9. Sparre, A.: On the collective theory of risk in case of contagion between claims. Transactions of the XVth International Congress of Actuaries 2(6) (1957)
10. Sreedhar, V.: Effort analysis using collective stochastic model. Tech. rep., IBM Technical Report (2014)