
Heng Cao, Jianying Hu, Chen Jiang, Tarun Kumar, Ta-Hsin Li, Yang Liu, Yingdong Lu, Shilpa Mahatma, Aleksandra Mojsilović, Mayank Sharma, Mark S. Squillante, Yichong Yu

Business Analytics and Mathematical Sciences Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598
{jengcao@us.ibm.com, jyhu@us.ibm.com, chjiang@us.ibm.com, ktarun@us.ibm.com, thl@us.ibm.com, jylu@us.ibm.com, yingdong@us.ibm.com, mahatma@us.ibm.com, aleksand@us.ibm.com, mxsharma@us.ibm.com, mss@watson.ibm.com, yichong@us.ibm.com}

In this paper, we present a suite of innovative operations research models and methods called OnTheMark (OTM). This suite supports the effective management of human capital supply chains by addressing distinct features of human talent that cannot be handled via traditional supply chain management. OTM consists of novel solutions for (1) statistical forecasting of demand and human capital requirements, (2) risk-based stochastic human-talent capacity planning, (3) stochastic modeling and optimization (control) of human capital supply evolutionary dynamics over time, (4) optimal multiskill supply-demand matching, and (5) stochastic optimization of business decisions and investments to manage human capital shortages and overages. The OTM suite was developed and deployed as an important part of the human capital management and planning process within IBM, providing support for decision making to drive better business performance. This is achieved through important contributions in the areas of stochastic models and optimization (control), and the innovative application and integration of these models and methods in human capital management applications.

Key words: human capital supply chains; human capital management and planning; demand forecasting; risk-based capacity planning; stochastic supply evolution; multiskill supply-demand matching; stochastic models; stochastic optimization and control.

Over the past 50 years, the services sector has grown to dominate economic activity in most advanced industrialized countries. Recent US Bureau of Labor Statistics data reveal that 75 percent of the labor force is employed in the services industry and that services industry output represents nearly 70 percent of the total industry output (Woods 2009). A critical driver of success for any service-delivery organization is its ability to manage and deploy the skills, knowledge, and competencies of its human capital—“having the right people, with the right skills, in the right place, at the right time.” Successful organizations realize that investments in their people are key drivers behind growth, profitability, and client satisfaction. This is especially true for providers of information technology (IT) services such as IBM, who offer a broad range of service products, each requiring people with specific skills, in markets characterized by highly volatile and uncertain client demands. Hence, forward-thinking businesses are beginning to invest in human capital supply chain (HCSC) methods as a major competitive differentiator.

Over recent decades, organizations have achieved significant gains in effectiveness and efficiency by developing advanced models of traditional manufacturing and logistics systems to optimize their supply chain operations. Although such traditional supply chain models and methods have a rich history within the operations research (OR) community, they cannot be directly applied to related problems in HCSCs. People are quite distinct from and more complex to model than the inanimate machine parts in traditional supply chains; thus, new models and methods are required to capture and represent these distinct
features and complexities. Concepts such as hiring, training, retaining, learning, and acquiring new skills fundamentally influence the optimization problems relevant to human capital management (HCM), making them more difficult to model and solve, but also more realistic; therefore, these concepts are of crucial importance to the successful management and planning of HCSCs in practice.

A prime example of the importance of effective HCM is Integrated Technology Services (ITS), an IBM line of business. ITS is a premier integrator of IT in the services industry; its annual revenues are approximately $4 billion. Through its 10 service product lines, ITS delivers a range of integration and support service products, henceforth called service engagements (e.g., the design and implementation of a data warehouse), for all IBM hardware and software products and beyond. The ITS business focuses on human talent; the majority of its engagement costs consist of human services rather than costs related to hardware and software. Unlike machine parts in traditional supply chains, the people comprising the ITS supply chain are not consumed during the service-delivery process and represent a long-term investment. The productivity and efficiency of people depend upon workload and utilization, whereas these notions for machine parts are less sensitive to workload and utilization. People can evolve in various complex ways, and are capable of deploying more than one skill at a time and doing so across multiple engagements. The service-delivery process is further complicated by uncertainty in resource staffing, simultaneous allocation of multiple resources (because engagements require different skills), and resource sharing (because people time-share their skills across different engagements).

The effective and efficient management and planning of human talent, while addressing the fundamental differences between human capital and traditional supply chains, is pivotal to the ITS delivery of service engagements. Seeking novel solutions, IBM ITS partnered with IBM Research to develop OnTheMark (OTM), an HCM framework comprising innovative OR models and methods to address the challenges and complexities of HCSCs. We develop the capability to forecast engagement demand and human-talent requirements. We derive stochastic modeling and optimization methods to support risk-based capacity planning that determines human-talent levels to maximize business performance. We develop an approach to predict future talent and skill composition via stochastic modeling and optimization (control) of supply evolution. We perform an optimal matching of multiskill talent against demand targets; we then optimize investment decisions to address talent shortages and oversages via our combined modeling and optimization solutions.

Since 2008, OTM has been deployed as part of the ITS HCM process, helping to improve the management and planning of human talent. Based on our experience with quarterly reviews, the OTM solution demonstrates relative revenue-cost benefits commensurate with 2–4 percent of ITS revenue targets over previous approaches. Beyond this internal application, we also develop prototype implementations of OTM models and methods within IBM platforms to support future IBM solution offerings for a broad collection of clients beyond IT services.

This paper presents an overview of the OTM solution, focusing primarily on mathematical modeling and optimization contributions. We first provide a high-level overview and then present a somewhat more detailed discussion of the OR models and methods. Cao et al. (2010) provide a more detailed technical description. We conclude with a summary of our worldwide OTM deployment within IBM. Throughout the paper, we use the term skill to generally refer to the full extent of human capital abilities, competencies, knowledge, skills, and talents. Examples include database architect, Java programmer, and network specialist.

Solution Overview

To address the challenges and complexities of HCSCs, we develop OR solutions with the goal of determining the talent levels and investments over time that maximize business performance. Although various investments are possible, we focus on hiring, training, promotions, and retention (via incentives) as representative decisions available to ITS. We implement our OR solutions as core capabilities of the OTM suite to support an overall HCSC process. OTM consists of: (I) demand forecasting; (II) risk-based capacity planning; (III) supply evolution and optimization;
Figure 1: The flowchart illustrates the capabilities and interrelationships in the OTM suite of OR models and methods.

(IV) multiskill shortage and overage analysis; (V) skill shortage and overage management (see Figure 1).

Various business and organizational aspects of ITS influenced our partitioning of the OTM solution among its constituent elements, which are depicted in Figure 1 and mapped to corresponding OR models and methods in Table 1. This highlights another important contribution: the integration of such a collection of OR methodologies. The OTM design also supports interactive sessions that allow users to evaluate different service-delivery scenarios under varying assumptions and conditions. This requirement creates additional methodological challenges to compute solutions of complex stochastic modeling and optimization problems in an efficient manner.

Beyond combined usage via the OTM suite, each capability is used independently by different ITS organizations to support various HCSC processes.

Given the distinct perspectives of different ITS users, OTM capabilities operate across a diverse set of planning horizons with time scales ranging from weeks and months to quarters and years. We next present an overview of the primary OTM elements and then describe each element in somewhat more detail in the sections that follow.

**Demand Forecasting**

The first step, demand forecasting (I), statistically characterizes the service engagement demand over the planning horizon in terms of revenue to the provider, number of engagements to be delivered, required time for starting each service delivery, and skills required for delivery. Service engagements are described in terms of revenue, duration, and type, often without linkages to staffing templates. To obtain a more accurate view of demand, we use machine learning methods to estimate the skill staffing requirements of each engagement from historical data. Statistical forecasting techniques are used to estimate the demand for each engagement; together with our staffing templates, they yield a complete characterization of demand. An example output consists of 10 and 5 engagements over the next two quarters, respectively, to implement data warehouses (engagement demand), with start times uniformly distributed over each quarter and each engagement requiring two database architects, four Java programmers, a network specialist, and so on (staffing templates). Such
demand forecasting (I) outputs serve as input to risk-based capacity planning (II) and supply evolution and optimization (III).

**Risk-Based Capacity Planning**

The next OTM capability concerns capacity planning—determining the capacity levels for people skills that satisfy demand and maximize business performance. Given the distinct features of people and HCSCs, OTM’s capacity planning requirements create a need for new stochastic modeling and optimization solutions that capture the business risks, HCSC dynamics, and demand uncertainty. In traditional supply chains, product demand is often converted into machine part demand through bills of materials and then provided as direct input into supply-demand matching. However, it is crucial in HCSCs to first analyze the financial implications and risks associated with distinct features of people and human talents, including longer-term costs and complex dynamics, as a function of engagement demand.

To address these challenges, we develop a risk-based capacity planning solution (II) that models the dynamics, trade-offs, and uncertainties associated with HCSCs and solves a corresponding stochastic optimization problem based on this model. The risk-based capacity planning models consist of multiclass stochastic loss networks with simultaneous resource allocation in which: losses model the risks of lost demand (and associated revenue) due to insufficient human capital at the time of engagement delivery; teams of people are required to jointly deliver service engagements; and multiple classes model different types of service engagements and human talents. Performance is a function of the revenues for service engagement delivery, discounted by probabilities for service engagements at risk of being lost, and the costs for maintaining skill capacity levels. We then solve the corresponding risk-based stochastic optimization problem to determine the skill capacity targets that maximize expected business performance.

OTM risk-based capacity planning addresses the revenue-cost dynamics and trade-offs between the risks of too few people with appropriate skills to deliver an engagement when needed and the risks of too many underutilized people. These revenue-cost dynamics and trade-offs can be evaluated and optimized from diverse perspectives by using risk-based capacity planning in different ways. This includes ideal skill capacity targets (independent of existing human capital) and optimal skill capacity levels and business investment decisions to achieve these levels (subject to existing supply constraints), both to maximize business performance.

An example output includes expected profit (revenue) being maximized over the next quarter (under a 12 percent gross profit margin constraint) with 200 (250) database architects and 500 (600) Java programmers. The outputs of risk-based capacity planning (II) over each time interval, which include corresponding skill capacity levels, serve as input to each of supply evolution and optimization (III) and multiskill shortage and overage analysis (IV).

**Supply Evolution and Optimization**

Another OTM capability concerns modeling the evolution of human capital and optimizing decisions to control the evolutionary dynamics of human talent. The resource dynamics in traditional supply chains are limited, with decisions typically restricted to ordering. In contrast, HCSCs are often characterized by complex time-varying supply-side dynamics, with many people acquiring skills, gaining efficiencies, and changing roles, some people attriting, and new people being hired. This results in substantial changes in the talent levels and skill composition of human capital over a given planning horizon. The complex time-varying evolutionary dynamics of human capital create a need for stochastic temporal models of supply-side dynamics and stochastic optimization of these human-talent dynamics over time through various investment decisions, including traditional hiring (ordering) options.

To address these challenges, we develop a supply evolution and optimization solution (III) that models the evolutionary dynamics of human talent and skill composition and solves a corresponding stochastic optimization (control) problem based on these models to maximize expected business performance over time. The supply evolution models consist of discrete-time multidimensional stochastic
processes whose states represent people with various combinations of talents and skills. The multi-period supply evolution optimization (control) problem incorporates these discrete-time stochastic processes together with decision variables and associated lead times for each action.

An example output includes evolution without intervention, rendering shortages in database architects and Java programmers that cause lost revenue over the next few quarters, and maximization of expected profit over the next few quarters by hiring two database architects and eight Java programmers, while incentivizing six Java programmers to delay retirement. Such investment decisions over planning horizons comprising stochastic behaviors that vary over time, together with the corresponding talent levels over the planning horizon, serve as input to both risk-based capacity planning (II) and multiskill shortage and overage analysis (IV).

Multiskill Shortages and Overages

The next OTM capability concerns the analysis of skill shortages and overages. In traditional supply chains, the resource shortages and overages are determined directly from demand and supply quantities. In contrast, the ability of people to perform multiple functions, and do so at the same time within one or more engagements, necessitates an advanced form of matching the multiskill supply to the demand.

To address these challenges, we develop a multiskill shortage and overage analysis solution (IV) that: models the multiple skills of people over each interval of the planning horizon involving time-invariant stochastic behaviors; optimizes the matching of such multiskill people against skill capacity targets to minimize a weighted sum of the skill shortages and overages; and computes shortages and overages for each individual skill under this optimal matching. We formulate the optimization problem for each interval of the planning horizon as a linear program; we then solve this optimization problem over the entire planning horizon as a dynamic program.

An example output includes, among a set of 10 people with both database architect and Java programmer skills in the next quarter and 14 such people in the quarter after next, 6 (9) who are optimally matched to deploy database architect skills, with the remaining 4 (5) optimally matched to deploy Java programmer skills in the next quarter (quarter after next). These multiskill shortage and overage analysis (IV) outputs, including optimal matchings of human capital to skill capacity targets and corresponding skill shortages and overages across every time interval comprising the planning horizon, serve as input to risk-based capacity planning (II) and supply evolution and optimization (III).

Skill Shortage and Overage Management

The final OTM capability concerns the optimization of business decisions to manage skill shortages and overages. We develop this solution (V) based on an iterative approach that combines the main OTM optimization capabilities by exploiting their interactions, dependencies, and time scales until convergence to a fixed-point equilibrium (steady state). Specifically, risk-based capacity planning optimization (II) determines the human-talent levels, supply evolution and optimization (III) determines the evolutionary dynamics of future skill composition, and multiskill shortage and overage analysis (IV) determines the matching of human capital against skill capacity targets, all with the goal of maximizing business performance over time. This yields a set of investment decisions for optimally managing skill shortages and overages. An example output includes addressing an overage of one database architect and a shortage of six Java programmers, among others, over the next few quarters by retraining one database architect to acquire another skill, hiring four Java programmers, and incentivizing two Java programmers to delay their retirement.

The dynamics and uncertainties of HCM decisions occur at various time scales, as reflected in our OR models and methods. Risk-based capacity planning operates over stationary intervals (typically on the order of a month or quarter) within the planning horizon; supply evolution and optimization operates across the entire nonstationary planning horizon (typically on the order of a few quarters, a year, or longer); and multiskill shortage and overage analysis operates across both stationary and nonstationary intervals. The notion of nonstationarity (stationarity) refers to stochastic behaviors that vary in distribution (are distributionally invariant to time shifts) over sufficiently long intervals. We consider the ITS HCSC
across relatively long planning horizons that include different forms of dynamics, uncertainties, and investment decisions. These time horizons are modeled as nonstationary intervals that consist of stationary periods, where the different forms of complex dynamics and uncertainties change from one stationary period to the next period according to a general stochastic modulation process.

**Demand Forecasting Models**

We forecast future demand for service engagements from three sources: (1) ongoing engagements being delivered; (2) opportunities of potential deals at different stages in the sales pipeline; and (3) deals anticipated based on market research and experience, but not concrete enough for entry in the pipeline. Service engagements require a collection of different skills, where the staffing templates that link engagement types and expected revenue to skill requirements in HCSCs are considerably more complex than traditional bills of material and labor.

We apply statistical techniques to engagement delivery data to compute probabilistic characterizations for the completion times of ongoing engagements and the corresponding availability of deployed human capital. For pipeline opportunities, we develop statistical models based on logistic regression techniques to predict the probability of each pipeline deal being won (i.e., contract signed) based on attributes such as lapse time and movement in the pipeline, client information, and deal size. These win probabilities are then used to probabilistically characterize the number of engagements of each type from the pipeline and their start times over future periods. For expected deals, we use quarterly revenue targets and typical deal sizes to probabilistically characterize the number of engagements of each type to be started to make up the difference between revenue targets and expected revenues from ongoing engagements and pipeline opportunities.

Given a lack of linkage between service engagements and required skills, we develop models to estimate staffing requirements for each engagement type. Following the statistical clustering methodology of Hu et al. (2007), we generate groups of similarly staffed engagements using historical data on engagement delivery. Such clusters for each engagement type are validated and refined through discussions with subject matter experts. Because these staffing templates evolve and require some degree of customization, we develop an automated methodology for generating and adjusting staffing templates by exploiting the semisupervised clustering and machine learning framework of Hu et al. (2008a, b). The resulting dynamic taxonomy of staffing templates provides important input to ensure accurate forecasting of human-talent requirements from all three sources of engagement demand.

**Risk-Based Capacity Planning**

The service engagement demands and skill requirements from our forecasting models are provided as input to the risk-based capacity planning solution. Our goal is to determine the skill capacity levels that best satisfy demand while addressing the complex trade-offs among revenue, cost, and associated risks. To highlight some of the difficulties involved, consider a scenario demand forecast for 10 service engagements over the next quarter. The best skill capacity levels to deliver this set of engagements depend on the degree of overlap among the corresponding delivery-time processes. At one extreme, if the delivery of all engagements overlaps, then 10 times the required staffing levels of one engagement is required; at the other extreme, if the delivery of these engagements is disjoint, then only the staffing levels of one engagement are required. Because the reality is typically somewhere between these extremes, a model of these complex dynamics and stochastic behaviors is needed to appropriately determine the skill capacity levels that maximize business performance subject to service engagement demand forecasts.

We therefore model the capacity planning problem as a stochastic loss network in which the risk of losing service engagement demand (and associated revenue) due to insufficient capacity in one or more required skills at the time the service engagement must be delivered is represented by the stationary loss probabilities of the stochastic network. It is important to note that for our purposes, we use the general notion of arrivals in the stochastic loss network to model
Interfaces 41(5), pp. 414–435, © 2011 INFORMS

The planning horizon consists of a sequence of intervals, each on the order of a month or quarter and each modeled as a stationary loss network whose fixed parameters change from one interval to the next according to a general stochastic modulation process. The length of each interval is sufficiently long for the multidimensional stochastic process modeling the loss network to reach stationarity. We then formulate and solve a capacity planning optimization problem to determine the skill capacity levels for our stochastic loss network that maximize performance, in terms of expected revenue, cost, and profit, over the entire planning horizon.

Our derivations and theoretical results for the risk-based capacity planning models are based partly on the slice methods developed by Jung et al. (2008) when restricted to Poisson arrival processes. In addition to the innovative application of these results, we extend our risk-based capacity planning models and methods to handle general arrival processes. This, together with our stochastic risk-based optimization results and methods, provides theoretical support for our approach to model and optimize over the complex dynamics and uncertainty that characterize capacity planning trade-offs in HCSCs.

**Stochastic Risk-Based Models**

We model each stationary period of the capacity planning problem as a stochastic loss network consisting of multiple skills and multiple service engagements. The delivery of an engagement requires units of capacity from people with specific skills. Instances of each type of engagement demand must be delivered at time epochs following an independent stochastic arrival process. Such a delivery opportunity is at risk of being lost if available skill capacity is less than that required in one or more skills; otherwise, the engagement is delivered and the capacity is reserved for each required skill throughout the duration of the engagement. The delivery duration times are independent and identically distributed random variables following a general distribution with unit mean. Engagement delivery (arrival) epochs and duration times are mutually independent.

Our stochastic loss network models the following: for each skill, the stationary probability that service-delivery arrivals encounter insufficient capacity (per-skill stationary blocking event probability); for each engagement, the stationary probability that service-delivery arrivals (at the times these engagements need to be delivered) find insufficient capacity for one or more required skills (per-engagement stationary loss-risk probability); and the vector \( \mathbf{n} \) of the number of active service-delivery engagements of each type in equilibrium. Each column of the skill capacity requirements matrix represents the skill staffing template for each type of service engagement from our demand forecasting models. Then, by definition, the state space corresponding to \( \mathbf{n} \) is constrained by the skill capacity vector and given by the polytope (1) of the appendix.

When, as part of satisfying demand, the times at which engagements need to be delivered are modeled as Poisson processes, then the above multidimensional stochastic process reduces to the famous Erlang loss model, which has been studied since the seminal work of Erlang (1917), as Brockmeyer et al. (1948) discuss. In this model, it is well known that the stationary distribution, representing the probability of the system being in state \( \mathbf{n} \) in equilibrium, is unique and exhibits an explicit solution in terms of a multivariate Poisson distribution constrained by the polytope (1) through a normalizing constant that quantifies the total probability mass of this polytope. The stationary loss-risk probability for each engagement type is directly expressed as a function of the
stationary distribution and model parameters. This loss-risk probability vector plays an important role in our stochastic model and optimization as a measure of the service engagements at risk of being lost, together with associated lost revenue estimates.

The complexity of computing the above normalizing constant is known to be \( \sharp P \)-complete (a higher complexity class than NP-complete) in the size of the network (Louth et al. 1994). Hence, an Erlang fixed-point approximation (EFPA) has been long used as an efficient alternative to the exact Erlang loss formula. The EFPA is based on approximating the blocking event probabilities of individual skills by a set of fixed-point equations, and then approximating the loss-risk probabilities for each service engagement type in terms of these skill blocking event probabilities. Whitt (1985) and Kelly (1991) provide additional details on the Erlang loss formula and EFPA.

Although EFPA resolves the prohibitive computational costs of the exact formula, it is well known that EFPA can provide relatively poor estimates for loss-risk probabilities in various model instances, which we found to be the case upon application within the ITS HCSC. To address these challenges, we first observe that by definition the mode of the stationary distribution of \( n \) relates to a solution of the optimization problem (2) in the appendix. Then, for each service engagement type, we derive a convex relaxation of the original optimization problem (2) restricted to each number of active engagements (slice) within the polytope (1). We solve this optimization problem to obtain the corresponding mode of the distribution for each slice.

This yields the approximation in Equations (3)–(5) for the engagement loss-risk probabilities by exploiting the definition (3) and the derived relationship (6). To reduce the computational complexity of this approach to that of EFPA, we also develop a three-point slice method in which the above convex relaxation is solved for the upper and lower limits of the polytope and for the mode of the overall distribution; the mode is then estimated for all other slices between successive pairs of the three computed modes, as Jung et al. (2008) and Anselmi et al. (2009) discuss.

It has been shown by Jung et al. (2008) that these slice-based approximations are asymptotically exact in a large-network limiting regime, as originally studied by Kelly (1991), under which the scaled vectors of skill capacities and engagement arrival rates tend toward infinity in fixed proportion with respect to a scaling parameter. More precisely, the loss-risk probabilities from the slice methods are shown to converge to the exact loss-risk probabilities as this scaling parameter tends to infinity. We further prove strict dominance of the accuracy of the slice methods over previous Erlang approximations by deriving and comparing large deviations results for the exact Erlang formula, the EFPA in the large-network limiting regime, and our slice-based methods.

We also develop accurate solutions for stochastic loss networks under general renewal arrival processes. Specifically, we derive a Gaussian fixed-point approximation based on the underlying multidimensional Gaussian process in which the engagement loss-risk probabilities are given by Equations (7)–(10). We further establish the asymptotic exactness of our Gaussian fixed-point approximation under the above large-network scaling.

**Stochastic Risk-Based Optimization**

Although stochastic loss networks model the dynamics and complexities of risk-based capacity planning, our objective is to determine the skill capacity levels that maximize business performance over the planning horizon. Focusing on a stationary interval of the planning horizon to simplify the presentation, our stochastic capacity planning optimization is based on revenues gained for delivering engagements at the time of their required delivery and on costs incurred for deploying skill capacity levels. The former expected realized revenues represent the difference between expected demand revenues and expected revenues at risk of being lost. More precisely, the formulation is given by the optimization problem (11) with our loss-risk probability approximations as constraints. Additional constraints can be included to guarantee desired serviceability levels or market share. We also consider formulations that maximize revenue or minimize cost subject to constraints on gross profit margins or cost and revenue targets. In each formulation, the stochastic loss network captures various sources of uncertainty and fundamental relationships among demand, supply, and engagement loss-risk probabilities.
These formulations render a nonlinear program, and our solution exploits properties of the underlying stochastic loss network with either Poisson or general renewal arrival processes. This includes establishing that, for a stochastic loss network in the above large-network limiting regime, the solution of the stochastic optimization problem converges asymptotically, in the limit as the scaling parameter tends to infinity, to the optimal solution under the exact engagement loss probabilities. Our experience in practice indicates that the scale of the risk-based capacity planning optimization is sufficiently large to yield accurate solutions.

We also determine a region of skill capacities that ensures the arrivals will be served with loss-risk probabilities no more than a given set of bounds. This region characterizes fundamental properties between the capacity vector and the loss probability vector relative to the arrival rate vector, which can be exploited to efficiently search the feasibility space of our risk-based capacity planning problem. Specifically, for a stochastic loss network in the large-network limiting regime, we establish a system of polynomial equations and inequalities that define a region within which the skill capacity vector must fall to achieve a given bound on the engagement loss-risk probability vector. Upon incorporating these polynomial equations and inequalities in our optimization problem to characterize its feasible region, adding corresponding constraints on the loss-risk probability vector, and exploiting theoretical results established by Lasserre (2001), we obtain a near-optimal solution with polynomial computational complexity and probabilistic accuracy guarantees.

The preceding formulations consider idealized optimal skill capacity levels under only costs for deploying skill capacity levels without addressing the costs of adjusting these capacity levels. The general formulation of risk-based capacity planning optimization incorporates all sources of human-talent capacity and their associated costs. In particular, our general formulation is given by the optimization problem (11) and (12), which includes the costs for hiring additional (retaining existing) skill capacity and training existing capacity to acquire skills, where existing capacity takes into account expected attrition and internal transitions from supply evolution and optimization. The skill capacity vector is augmented to reflect the combination of existing and retained skill capacity and new skill capacity from hiring and training actions. The constraints can also include a budget for hiring, training, and retaining costs, either jointly or separately, which is especially useful in practice when various business investment decisions are managed by different organizations, each with its own budget.

Finally, Bhadra et al. (2007) consider the structural properties of a multiperiod stochastic loss network under Poisson arrivals whose solution is obtained via the EFPA. However, the resulting properties do not generally apply to our multiperiod stochastic loss network under renewal arrival processes, nor to our solution via the slice methods.

**Illustrative Examples**

To illustrate the complexities of risk-based capacity planning and the benefits of our approach, consider a simple scenario of an application development (AD) engagement that requires one Java programmer and one engagement manager, and a data administration (DA) engagement that requires one database architect and one engagement manager. The arrival epochs for when AD and DA engagements need to be delivered follow independent Poisson processes with rates 20 and 10, respectively.

Generally speaking, the higher the skill capacities, the lower the engagement loss-risk probabilities. However, the sharing of managers among both engagement types can create complex interactions among the skill capacities. In two instances of this example, which shed light on the complex dynamics in general, we observe that simply incrementing the engagement manager capacity by one can result in dramatically different loss-risk probabilities for a given demand, depending on the various parameter regions in which the system may operate. Figure 2 displays the corresponding results for both cases.

In Case 1, fix the Java programmer and database architect capacity to 10, and vary the engagement manager capacity from 16 to 26. We observe that, although AD and DA loss-risk probabilities both decrease monotonically, the rates of decrease for the two engagement types differ considerably. We also see that for the same type of engagement, the rate can vary significantly for different manager capacities.
16
17
18
19
20
21
22
23
24
25
26
0.2
0.3
0.4
0.5
0.6
0.7
Loss-risk probability
Loss risk probabilities and expected profit as a function of engagement manager capacity
Eng. I
Eng. II
(a) Case 1
16
17
18
19
20
21
22
23
24
25
26
14
12
10
8
Expected profit
Capacity
Expected profit as a function of Java programmer capacity
Eng. I
Eng. II
(b) Case 2
9
10
11
12
13
14
15
16
17
18
19
8
10
12
14
Loss-risk probability

Figure 2: The graphs show an example of risk-based capacity planning.

In Case 2, fix the engagement manager and database architect capacities to 20 and 10, and vary the Java programmer capacity from 9 to 19. This illustrates that incrementing Java programmer capacity by one can have a very different impact on the loss-risk probabilities of both engagements types. Specifically, incrementing one unit of Java programmer capacity yields a decrease in the AD loss-risk probability, while increasing the DA loss-risk probability. Intuitively, this is because under the parameter settings, Java programmers are a bottleneck for AD engagements (i.e., an AD engagement is at risk of being lost most likely because of a lack of Java programmers). By increasing Java programmer capacity, we allow more AD engagements to be accepted; however, this causes greater utilization of the common (manager) skill by AD engagements, and the DA loss-risk probability increases. Furthermore, these complexities significantly impact the question of optimal skill capacity levels, as we can observe from the expected profit curves for both cases (see Figure 2).

More generally, in large real-world applications of our risk-based capacity planning model and optimization, it is quite difficult to distinguish the many degrees of sharing among skills, given the complex interactions and dependencies among the service engagements and their skill requirements. Hence, the effects of changing skill capacity levels on engagement loss-risk probabilities are often very complicated. These effects are further compounded by the complex dynamics and interactions underlying the optimization of skill capacity levels.

Finally, we consider a representative application of risk-based capacity planning using ITS data. Figure 3 plots expected profits, revenues, and costs on the y-axis as a function of risk tolerance constraints along the x-axis. The leftmost set of results represents the solution that maximizes expected profit. The risk tolerance constraints become more tight as we move to the right along the x-axis, where the rightmost set of results represents when almost all of the demand is being satisfied. We observe that although expected revenues increase as more demand is satisfied, expected costs also increase and tend to do so at a faster rate. This application illustrates how executives can use risk-based capacity planning to determine the best way to operate their business with respect to expected revenues, costs, and profits, revenue loss-risk tolerances, and other business and economic concerns.

Supply Evolution and Optimization
The management of human capital supply over the planning horizon is another fundamental aspect of HCSCs. Our goal is to understand the evolution of human-talent levels over time and determine the best
actions to adjust these evolutionary dynamics. To highlight some difficulties involved, consider a common scenario in the military in which a long period is required for one to reach the highest ranks (e.g., General). Steps must be taken at all levels of this HCSC (e.g., recruiting, training, experience, and promotions) to ensure that enough talent is available at all levels, with lower-level ranks feeding higher-level ranks over time. A model of these complex dynamics and stochastic behaviors is needed to determine appropriate talent levels that maximize expected performance over time, subject to supply, demand, and organizational constraints. This challenge involves stochastic modeling and analysis of the evolutionary dynamics and flexibility of human capital, including future internal transitions (e.g., promotions, certification, and training) and future external transitions (e.g., hiring, attrition, and acquisitions). The time scale of these dynamics is typically on the order of days, weeks, or months, whereas the overall planning horizon is often on the order of months, quarters, or years.

Therefore, we model the temporal evolution of human talent as a discrete-time, multidimensional stochastic process in which the dynamics and uncertainties vary over time. For this purpose, we leverage historical data on human capital dynamics and business and economic conditions, and input from subject matter experts. The dynamic evolution of human talent also can be influenced in desired directions by various investments and policies in response to uncertain and changing business and economic conditions. Hence, we formulate and solve a multiperiod stochastic optimization (control) problem based on our supply evolution model to determine the investment decisions and resulting evolution of talent levels over time that maximize business performance. Such decisions in each period include hiring, training, promotions, other internal talent transitions, and incentivizing to reduce attrition, together with the lead times associated with each action and policy.

To model and optimize the HCSC evolutionary dynamics at temporal granularities dictated by business operations, we use a multiperiod discrete-time stochastic process defined over a planning horizon comprising $T + 1$ periods. Although our OTM approach is completely general, the time granularity most often used is monthly and quarterly periods.

**Stochastic Evolution Models**

Our stochastic supply evolution models are based on aggregating people into human capital groups (HCGs) according to attributes of interest. Examples include all talent and skills relevant to the business,
and other factors such as levels of competency, productivity, proficiency, and certification. Because each person comprising the supply side is capable of attaining and employing a collection of attributes, our supply evolution models are more precisely based on aggregating people into HCGs according to various combinations of attributes. Within each period of the planning horizon, individuals make dynamic transitions between these HCGs. Such dynamic transitions include both internal and external transitions that lead to a complex stochastic network topology, both within each period and from one period to the next across the planning horizon.

Consider a model comprising five HCGs: C programmers, Java programmers, SQL programmers, programmers proficient in C and Java, and programmers proficient in Java and SQL. Now consider the evolution of individuals among these HCGs over a single period. People are hired into each group or attritted out of each group with certain probabilities, such as the probability that some number of C programmers are hired (attritted) within the current period. Similarly, probabilities are associated with internal transitions among HCGs, such as the probability with which some number of C programmers become equally proficient in Java programming. The corresponding sets of probabilities depend upon the period in which the dynamic transitions occur and the number of C programmers at the start of this period. We exploit properties of this complex network topology to obtain scalable solutions for our stochastic supply evolution models; one example is sparse matrix methods, given that transitions tend to be localized.

Our stochastic supply evolution models consist of a discrete-time, multidimensional stochastic process that records the expected number of people in each HCG. Now consider a portion of this stochastic process corresponding to the population of a single HCG (e.g., Java programmers). At the start of the planning horizon, we will know with certainty that the organization has a specific number of Java programmers, thus yielding a single state for each HCG at time 0. At the next period, the state space of the stochastic process then needs to allow for all possible populations within the HCG of Java programmers, governed by the corresponding set of transition probabilities. For example, assuming a current count of Java programmers at time 0, the number of Java programmers will increase or decrease to a specific number of Java programmers at time 1 with a corresponding set of probabilities for hires and internal transitions into this HCG and for attrition and internal transitions out from this HCG. Upon considering this portion of the stochastic process for all possible populations and expanding this view to the entire stochastic process across the large number of HCGs, we clearly have a prohibitive curse of dimensionality problem as the state space blows up.

To address this curse of dimensionality, we consider a decomposition of the discrete-time stochastic process that records the expected number of people in each HCG. This renders a single state for each HCG and an analogous set of probabilities that governs transitions from one state to another within each period and across periods that comprise the planning horizon. More formally, we consider the family of subsets of the set of skills that people possess, in which case the types of people who comprise the supply consist of elements of this family. The state vector $y$ models the expected number of people of each type (i.e., the expected number of people in each state) over a period. We then express in expectation the number of people in each state at time $t+1$ equal to the number of people in the state at time $t$ plus the number of people hired into the state plus the number of people transitioning into the state minus the number of people transitioning out of the state, either to another state or through attrition, each over the period $t$. This expression for the net dynamics of human capital evolution in each period can be written in matrix form with respect to a set of one-step transition probability matrices. From these evolutionary dynamics and direct calculations, we derive the skill composition at a given time in matrix form yielding Equation (13) in the appendix. By having the model variables depend on the period, our stochastic models support time-varying behaviors of various forms (e.g., seasonal effects) for the evolution of HCSC dynamics.

We investigated various approaches to parameterize our stochastic temporal models, obtaining accurate predictions through the inference of base-model parameters from historical HCSC data. For example, the elements of the one-step transition probability
matrices can be calculated from statistical data on the number of transitions from one state to another state over each period. Other model parameters can be calculated in an analogous manner. We needed to adjust such historical data in a few circumstances primarily because of events that were not representative of actual business trends (e.g., one-time acquisitions) or a lack of statistical confidence on the samples for some states. In the first case, we worked with subject matter experts to filter out the impact of these events from the data to ensure that the statistics properly represent the number of transitions with respect to future business trends. We took an analogous approach for other events that might bias the statistics to deviate from actual business trends, for example, the use of rules to factor out errors resulting from data entry mistakes. In the second case, we leverage standard statistical tests for each state to verify a sufficiently large sample size to guarantee a desired margin of error. Whenever the statistical test fails, we aggregate similar skills into a combined state to obtain statistics whose sample size is within the desired margin of error; upon solving the resulting stochastic temporal model, we probabilistically distribute the predicted human capital for this combined state into its constituent states.

Stochastic Evolution Optimization

Although discrete-time, multidimensional stochastic processes model the dynamics and complexities of human capital evolution, our goal is to determine the investment decisions and policies that influence HCSC dynamics in desired directions over the planning horizon. The driving objective is to maximize expected profit across all periods, where revenues depend upon demand and costs include maintaining and achieving the skill capacity levels, noting that maximizing expected revenue or minimizing expected cost subject to constraints on gross profit margin or on cost and revenue targets is included within our solution framework and methodology.

To this end, we associate costs and rewards with each state of our stochastic supply evolution models. Namely, the costs and rewards for each state are captured as functionals of the number of people in the state and the amount of time they spend in the state. For example, the per-period costs for C programmers at a specific experience and proficiency level include the corresponding salary and benefits (e.g., medical, pension) as a function of the state population, with the analogous per-period rewards including revenues driven through service delivery as a function of the population and demand for this state. Cost and reward functions are also associated with transitions between states, where the costs and rewards for transitioning from one state to another is a function of the states involved and the number of people making the transition. For example, the per-period costs and rewards for a set of C programmers becoming equally proficient in Java programming include training costs and related service-delivery revenues. Our stochastic supply evolution optimization models associate lead times for any available actions taken with respect to each state transition (e.g., hiring). These lead times capture the delays between the time the action is initiated (e.g., hiring starts) and the time the result of the action is actually realized (e.g., new employee comes on board). Note that adjusting transition probabilities between states within the corresponding stochastic decision process represents investments, policies, and actions such as training and promotion.

Then, in addition to skill composition trajectory, our stochastic optimization models characterize the evolution of expected cost, revenue, and related financial metrics of the HCSC over time as functionals of the discrete-time stochastic process. Specifically, under appropriate independence assumptions, the expected cumulative costs of human capital over the planning horizon are given by the sum of dot products of the expected cost vector and state vector over all periods. The expected cumulative revenues are obtained in an analogous manner where the expected revenue for each period \(t\) is a function of \(y(t)\), subject to the componentwise demand constraints. Expected profit over the planning horizon is computed as a function of these revenue and cost metrics, including factors such as productivity and efficiency levels that depend upon human capital type, workload, and utilization.

Investment decisions and policies can be exploited to impact HCSC financial metrics; therefore, we solve a corresponding stochastic optimization (control) problem based on the above decision processes to determine the optimal human capital evolution over time. To this end, we express HCSC dynamics
as the discrete-time linear dynamical system in Equation (14) in the appendix. Our goal is to determine the vectors of decision variables for interstate transitional actions and for hiring and retention actions that maximize expected profit over the planning horizon. More precisely, we consider a formulation of the optimization problem (15) (see the appendix) in terms of expected profits as a function of the system state and expected decision costs. Because expected revenue for each period is linear in the state populations up to the forecasted demand for skills within the period, the objective function is (piecewise) linear in the state and decision vectors. Hence, we stack the summation in the objective function (15) (see the appendix) and combine this with the linear dynamical system constraints, expressed in a corresponding matrix form, into a (piecewise) linear program with respect to the combined decision and cost vectors. We additionally have, in a corresponding matrix form, constraints on balancing total flow, limiting total outflow, and limiting total attrition of human capital, and restricting revenue from exceeding demand. The output of this (piecewise) linear program provides an optimal decision vector from which we cull the optimal decision and state vectors for each period.

Illustrative Examples
To illustrate the complexities of supply evolution and optimization and the benefits of our approach, consider a simple application scenario starting with the evolution of a single skill. Using historical data, we predict a significant increase in attrition in November and December because of baby-boomer retirements. Furthermore, the demand forecast for this skill is relatively steady for the subsequent months (see the Demand curve in Figure 4). This suggests that there will be a significant shortage in this skill relative to demand under these retirement predictions. To mitigate this problem, we investigate the possibility of incentivizing some of this attrition to postpone retirement until enough people with this skill can be brought on board. We model the behavioral responses of baby-boomer retirees as a concave function of increasing incentives where the higher the incentives, the larger the fraction of people who are willing to postpone retirement, with diminishing returns. For this representative scenario, our supply evolution optimization solution determines the set of investments to maximize profit relative to forecasted demand (see the Optimal curve in Figure 4).

In contrast, many organizations would never realize the upcoming shortage because of future retirements under this scenario until it is too late; thus, they will lose considerable revenue while trying to catch up over the hiring lead time, as the Shortage-based curve in Figure 4 shows. Somewhat more enlightened organizations might realize this upcoming shortage in advance. However, they will hire enough people to fill this shortage as they believe to be the case in September without understanding that transitions into (out from) this skill will naturally occur over the next few months; thus, they will hire too many (too few) people and create a future skill overage (shortage) problem, as the Partial information curve in Figure 4 shows.

Multiskill Shortages and Overages
Given skill capacity targets and multiskill human capital predictions, we determine an optimal matching of
the latter to the former. The planning horizon (on the order of a year) consists of a sequence of stationary intervals, each on the order of a month or a quarter. We formulate and solve our skill shortage and overage analysis across these stationary periods to determine an optimal supply-demand matching that minimizes a weighted sum of shortages and overages and compute the corresponding skill shortages and overages. Any predetermined assignments can be incorporated as constraints; our solution then optimally matches the remaining multiskill human capital to the remaining demand.

We capture the fundamental aspects of this problem for each stationary period as a linear program. Our objective is to determine the matchings of human capital possessing multiple skills against the skill capacity targets to minimize a weighted sum of the skill shortages and overages. The shortages and overages must satisfy a balance equation in which the sum of skill matchings and expected shortage equate to the sum of corresponding skill targets and expected overage—constraint (16) in the appendix must be satisfied. Then our solution to the multiskill shortage and overage analysis across multiple stationary intervals under a general stochastic modulation process is obtained by solving the corresponding dynamic program.

Our use of weights for each skill shortage and overage in the objective function helps to facilitate skill priorities and importance factors when matching multiskill human capital to the skill capacity targets. For example, it might be desirable to push the optimal matching toward hot skills (to reduce shortages of such skills), all else being equal, and similarly to push the optimal matching away from commodity skills (to reduce overages of such skills). In addition, it can be desirable to include preferences among the subsets of skills in multiskill supply-demand matching. To this end, we introduce an additional family of variables and exploit the big-M method for solving linear programs (Bertsimas and Tsitsiklis 1997). More precisely, our complete general formulation for each stationary interval of the multiskill shortage and overage analysis with preferences is given by the linear program with objective (17) and constraints (16) in the appendix.

Skill Shortage and Overage Management

Once we have obtained skill shortages and overages by solving the multiperiod version of the above linear program, we next determine the investment decisions to manage these shortages and overages over the planning horizon. Our approach is based on an iterative combination of the OTM optimization capabilities. Initially, a first-phase application of our overall solution process results in: (1) risk-based capacity planning optimization, which provides ideal skill capacity targets based on business and economic considerations and the HCSC dynamics within each period; (2) supply evolution and optimization, which provides human capital capacity levels and required investments across future periods to realize these capacity levels; and (3) multiskill shortage and overage analysis, which provides optimal matchings of multiskill human capital to skill capacity targets.

The intraperiod dynamics are captured by each instance of risk-based capacity planning optimization, and the interperiod dynamics are captured by supply evolution and optimization, whereas intraperiod and interperiod dynamics are captured by single- and multiple-period instances of shortage and overage analysis, respectively. This partitioning of the overall OTM solution is based on various business and organizational aspects of the ITS HCSC. Given these dependencies and interrelationships, we combine all three optimization capabilities through a subsequent-phase iterative procedure as follows. First, the costs of hiring, training, and retaining are included in risk-based capacity planning optimization. The resulting skill capacity output and the skill composition output of supply evolution and optimization are then optimally matched via multiskill shortage and overage analysis. This skill-matching output is provided as input to both risk-based capacity planning optimization and supply evolution and optimization for updates on the recommended investment decisions. The entire process is then repeated until all solutions converge to their equilibria.

OTM Business Benefits

The OTM suite has been implemented and deployed in several ITS organizations worldwide to support
the management and planning of major portions of the ITS HCSC. The business processes supported by OTM include engagement delivery, resource capacity planning, engagement portfolio management, sales-delivery interlock, business decision support, and strategic planning. In this section, we summarize our experience with the OTM worldwide deployment over the past couple of years. We first discuss validation studies and then present a representative set of business applications based on recent ITS data. We conclude with reviews of the OTM impact on the ITS business, and ongoing efforts to support future IBM solution offerings.

Validation Studies
Our experience includes an analysis of the accuracy of each OTM capability on a regular basis. From the start, we instituted a feedback loop to validate and calibrate OTM predictions and results against actual business outcomes and ITS data. This involves quarterly reviews with IBM executives to track and evaluate the accuracy of the OTM methodology. Such validation studies demonstrate strong agreement between OTM model and optimization outputs and the corresponding business outcomes realized in practice. For example, we observe a typical accuracy of 85–90 percent for our pipeline revenue forecasts, and an accuracy generally higher than 90 percent for overall human capital demand forecasts (relative to quarterly revenue targets). Our experience demonstrates loss-risk probability estimates from OTM risk-based capacity planning models to be within a few percentage points of ITS data on real-world engagement losses. We observe the accuracy of OTM supply evolution models to be within a few percentage points in predicting skill compositions of large organizations (with many thousands of people) over planning horizons of up to one year. The accuracy of such predictions for smaller organizations, correspondingly, are within 12 percent of the real-world evolution of skill composition.

Representative Business Applications
The OTM deployment supports ITS management and planning processes. To highlight such applications of OTM in IT services delivery, we briefly present scenarios based on recent ITS data. These include representative comparisons that demonstrate the benefits of our OR models and methods over the approaches used prior to the OTM deployment.

OTM risk-based capacity planning is applied in three ways to provide distinct insights: (1) fix revenue targets and determine the optimal skill capacity levels that minimize cost and achieve the targets (or gross profit margins); (2) fix cost targets and determine the optimal skill capacity levels that maximize revenue and achieve the targets (or gross profit margins); (3) determine the optimal skill capacity levels that maximize profit. Now consider a quarterly capacity planning problem with 132 skill types and 216 engagement types under both risk-based capacity planning and a previous approach based on linear skill-demand projections.

In a cost-minimizing application, the same revenue targets and similar aggregate loss-risk probabilities are realized under both solutions. However, the skill capacity portfolios differ considerably, with risk-based optimization providing lower loss probabilities for engagements with relatively high profit margins and lower capacity levels for engagements with relatively low profit margins. The risk-based optimization solution yields more than a 15 percent reduction in the individual loss-risk probabilities for different engagement types and more than a 40 percent reduction in expected capacity costs.

In a revenue-maximizing application, risk-based optimization increases the capacity levels for those skills that are critical to engagements with relatively high profit margins or relatively high usage, across a wide range of engagement types. Hence, the aggregate loss-risk probability is 5 percent under risk-based optimization, as compared to 23 percent under the linear-projection approach, which in turn renders a 35 percent increase in expected revenue.

In a profit-maximizing application, risk-based optimization balances the trade-offs among engagement revenues and capacity costs, yielding a 100 percent increase in expected profit over the previous approach. Although the risk-based solution provides lower expected revenue, it also renders significantly lower capacity levels than the linear-projection approach, both for individual skill capacity levels and in the aggregate. These cost reductions in turn yield a significant increase in expected profit, as Figure 3 similarly illustrated.
OTM supply evolution and optimization supports supply-side decision management. Consider a skills capacity prediction problem, based on an ITS data set consisting of nearly 500 skills, under both our supply evolution models and a previous approach of linear skill capacity projections. The supply evolution models are parameterized by historical quarterly hiring, attrition, and internal transition data, whereas the linear projections are calculated from historical quarterly hiring and attrition trends. Figure 5 compares the staffing-level predictions for three skills from both approaches looking eight quarters into the future. We observe that our supply evolution models provide better predictions of future staffing levels by preserving various business cycles and dynamics, as evidenced by the moderate degrees of periodicity and growth in the supply predictions. However, the linear-projection approach leads to overstaffing and understaffing because it ignores important dynamics such as the probabilities associated with internal transitions between states of the HCSC; for example, refer to skill 3 in Q4 and skill 1 in Q4 in Figure 5.

Next, consider the corresponding optimization problem, given demand forecasts for the next four quarters. Our supply evolution optimization uses human capital costs and service-delivery revenues to determine the investment decisions and human-talent evolution that maximize profit, in comparison with a
previously used myopic approach that adjusts capacity levels up to a fixed fraction $\alpha$ of the skill shortages and overages. Figure 6 compares the expected profits and staffing levels for three skills obtained under both approaches. We observe that the myopic policy exhibits an interesting concave behavior as a function of $\alpha$. Even under the best $\alpha$ setting of 40 percent, our stochastic optimization provides an expected profit increase of nearly 50 percent. These profit improvements are related to our observations that unnecessary responses to temporary dips in demand can be counterproductive. Given the costs of staffing decisions, the time-varying system dynamics over the planning horizon, and the statistical nature of demand forecasts, it is often preferable to retain human capital in excess of demand estimates. For example, Figure 6(b) depicts the significant overstaffing and understaffing obtained with the previous myopic policy; refer to skill B in Q1 and Q3 and skill A in Q1 and Q4, respectively. This in turn entails unnecessary investment costs and reduced profit, which is clearly an undesirable solution from both a financial and personal perspective.

OTM shortage and overage analysis determines the optimal matching of multiskill human talent against capacity targets to minimize a weighted sum of skill shortages and overages. Consider a supply-demand matching problem based on an ITS data set comprising 950 employees and 307 skills to compare OTM optimization with results from a previous approach in which multiskill human capital are partitioned into individual skills using historical averages of their skill deployment in the prior quarter. This business application shows that our shortage and overage analysis yields a 25 (42) percent reduction in skill shortages (overages) and a 32 percent reduction in the sum of skill shortages and overages.

OTM shortage and overage management identifies important business decisions with respect to impact on expected revenues through higher skill capacity levels and on expected costs to realize these capacity levels. Applying this OTM capability to the above business application yields a 200 percent improvement in expected profit over that of previous approaches. This solution represents a 7 percent decrease in expected revenues relative to the first-phase application of risk-based capacity planning, under which the costs to realize the corresponding ideal skill capacity targets are ignored. When such costs are factored into the analysis, the idealized risk-based capacity planning solution yields only a 177 percent improvement in expected profit over that of the previous approaches. The additional relative expected profit improvements of 200 percent are the direct result of our iterative methodology.
Business Impact
OTT’s business impact has been evaluated through quarterly reviews with IBM executives based on comparisons with previous approaches used by ITS: (1) linear projections to forecast ongoing demand and pipeline demand; (2) skill shortages and overages obtained by partitioning multiskill human capital among their individual skills using historical averages of skill deployment in the previous quarter; (3) investment decisions to hire and train human talent to reduce a certain fraction (α) of skill shortage and overage estimates. In addition, the efficiency and efficacy of our automated solution enables execution of OTM capabilities in response to unexpected events to adjust and fine-tune HCM investments and strategies.

The quarterly executive reviews demonstrate potential for reductions in skill shortages and overages by 10–80 and 30–150 percent, respectively. The business benefits of OTM solutions over previous approaches are then computed based on: (1) increased costs for human capital investments and decreased profits for lower utilization under skill shortages; and (2) increased costs for hiring and training and decreased revenues for insufficient human capital to fulfill the demand under skill shortages. The initial benefit comparisons reviewed with IBM executives demonstrated the financial impact in the first year of OTM deployment to be over $11 million within a single quarter in the United States in terms of cost savings and increased revenue. Similar results have been repeated within other ITS organizations in later stages of worldwide deployment, translating to relative revenue-cost benefits of OTM solutions over previous approaches commensurate with 2–4 percent of ITS quarterly revenue targets.

Although our focus has been on the ITS deployment, the possibilities for exploiting the OTM framework and methodology in a broader application setting are numerous. Concepts of labor and people are central to all services organizations; companies in health care, finance, insurance, retail, and the public sector are asking many of the same questions addressed by the OTM solution. What is the outlook for a nursing workforce three years from now if current hiring, training, and attrition patterns continue? Will enough insurance agents with a particular set of skills be available to meet product demand? What levels of sales can be driven with the current sales force? How many financial analysts should be hired next year to deliver on a set of growth objectives?

As a result of these and related questions, the issue of HCM is becoming one of the most important factors on CEO agendas. The ability to manage and plan human talent more effectively and efficiently is a critical driver of success for most services organizations, especially those with a large number of employees and a diverse portfolio of products and services. Analyst research indicates that, despite the majority of organizations making significant financial investments in the recruitment, training, and development of their people, they spend up to an additional 8 percent of total wages and salaries (on average) to manage human-talent issues; many of these issues could be avoided or turned into bottom-line contributions by exploiting advanced HCM solutions. The OTM suite is an important and innovative step in this direction. We continue to enrich our solution methodology in collaboration with various IBM product groups to create a general HCM platform for a broad client base.

Appendix
Risk-Based Capacity Planning
Let $\mathcal{S} := \text{set of skills, indexed by } i, \text{ and } \mathcal{K} := \text{set of service engagements, indexed by } k$. Define $E_i := \text{stationary skill } i \text{ blocking event probability}, L_k := \text{stationary engagement } k \text{ loss-risk probability}, n_i := \text{number of active engagements } k \text{ in equilibrium}, n := (n_i)$. By definition, we then have $n$ as an element of the polytope

$$\mathcal{S}(C) := \{n \in \mathbb{Z}^{[S]}_+ : A n \leq C\},$$

where $A := [A_{i,k}], C := (C_i), A_{i,k} := \text{amount of skill } i \text{ capacity required to deliver engagements } k, C_i := \text{amount of skill } i \text{ capacity overall}, \forall i \in \mathcal{S}, \forall k \in \mathcal{K}$.

Assuming Poisson arrival processes with rate vector $\nu = (\nu_k)$, we address the challenges of EFPA by observing that the mode $n^*$ of the stationary distribution $\pi(n)$ corresponds to a solution of the optimization problem

$$\max_n \sum_{k=1}^{[S]} n_k \log \nu_k - \log n_k! \quad \text{s.t. } n \in \mathcal{S}(C),$$

and by defining a continuous relaxation $\bar{\mathcal{S}}(C) := \{x \in \mathbb{R}^{[S]}_+ : A x \leq C\}$ of the polytope $\mathcal{S}(C)$ and subsets of
in this relaxation \( \mathcal{F}_v(x) := \mathcal{F}(C) \cap \{ x : x_k = \ell \} \), with \( x := (x_k) \) the corresponding continuous relaxation of \( n \). For engagements \( k \), we derive a convex relaxation of optimization problem (2) by exploiting Stirling’s approximation, ignoring \( O(\log x_k) \) terms, and restricting to the slice \( n_k = \ell \in \{ n_k : n \in \mathcal{F}(C) \} \). We solve this optimization problem to obtain the mode \( x^*(\ell, k) \) of the corresponding distribution, which we use to obtain our approximation of \( \mathbb{P}[n_k = \ell] \) as

\[
\mathbb{E}[n_k] = \sum_{\ell \in \{ n_k : n \in \mathcal{F}(C) \}} \ell \mathbb{P}[n_k = \ell],
\]

subject to constraints of the form \( L_k \leq \beta_k \in (0, 1] \), where \( u_i \) is the base revenue rate for engagement \( k \); \( v_i \) is the base cost rate for skill \( i \); \( v_{i1}^H, v_i^T, v_i^R \) are the cost rates for hiring, training, and retaining skill \( i \); \( C^H_i, C^T_i, C^R_i \) are the amounts of existing, hired, trained, and retained capacity for skill \( i \); and \( C_i = C^H_i + C^T_i + C^R_i \) for all \( i \in \mathcal{J} \) and \( k \in \mathcal{K} \).

### Supply Evolution and Optimization

Let \( \mathcal{J} := \) family of subsets of the set of skills \( \mathcal{J} \) that are possessed by people, indexed by \( j \). For all \( j \in \mathcal{J}, t \in \{ 0, \ldots, T \} \), define \( y_j(t) := \) expected number of people in state \( j \) at time \( t \), \( h_j(t) := \) expected amount of hiring into state \( j \) over \( [t, t+1) \), \( a_i(t) := \) expected amount of attrition from state \( j \) over \( [t, t+1) \), \( y(t) := (y_j(t)) \), \( h(t) := (h_j(t)) \), and \( a(t) := (a_j(t)) \). In addition, define \( p_{j,j}'(t) := \) stationary probability of transitions from state \( j \) to state \( j' \) over \( [t, t+1) \), where \( \sum_{j' \in \mathcal{J}} p_{j,j}'(t) \leq 1 \). When state \( j \) attrition is strictly positive, this inequality is strict and \( 1 - \sum_{j' \in \mathcal{J}} p_{j,j}'(t) \) represents the stationary probability that an individual leaves state \( j \) through attrition, such that \( a_j(t) = y_j(t)(1 - \sum_{j' \in \mathcal{J}} p_{j,j}'(t)) \). The dynamics of our supply evolution models are derived in matrix form (using row vector notation) to render

\[
y(t+1) = \sum_{s=0}^{t} h(s) \prod_{s'=s+1}^{t} P(s') + y(0) \prod_{s=0}^{t} P(s),
\]

where \( P(t) := [p_{j,j}(t)], \forall j, j \in \mathcal{J} \) and \( \forall t \in \{ 0, \ldots, T \} \).

Define \( \mathcal{E}_j(t) := \) expected costs of state \( j \) over \( [t, t+1) \), \( \mathcal{R}_j(t) := \) expected revenues of state \( j \) over \( [t, t+1) \), \( \mathcal{E}(t) := (\mathcal{E}_j(t)) \), and \( \mathcal{R}(t) := (\mathcal{R}_j(t)) \). Then, the expected cumulative costs of people over \( [0, T] \) is given by \( \sum_{t=1}^{T} (\mathcal{E}(t) \cdot y(t)) \), under appropriate independence assumptions. The corresponding expected cumulative revenues and profits are obtained in an analogous manner as functionals of both \( \mathcal{R}(t) \) and \( y(t) \), subject to the componentwise demand constraints. We rewrite Equation (13) into the following discrete-time linear dynamical system recursion (using column vector notation)

\[
y(t+1) = y(t) + \mathbf{B}(t)u(t),
\]

where \( \mathbf{B}(t) \) captures the sparsity patterns of \( \mathbf{P}(t) \), and \( u(t) := \) vector of decision variables for transitions between states and for hiring and retention actions.
Our goal is to solve the stochastic optimization and control problem
\[
\min_{u(0), \ldots, u(T-1)} \sum_{t=0}^{T-1} -\mathcal{P}(t+1, y(t+1)) + d(t) \cdot u(t),
\]
where \(\mathcal{P}(t+1, y(t+1)) := \text{expected profits at time } t+1 \) as a function of the system state \(y(t+1)\) subject to demand, and \(d(t) := \text{expected decision costs at time } t, \) for all \(t \in \{0, \ldots, T-1\}\). Finally, we stack the summation in the objective function (15) and formulate this optimization problem as a (piecewise) linear program in terms of a decision variable vector \(z := [y(1), \ldots, y(T), u(0), \ldots, u(T-1)]\) and a weight vector \(w := [-\mathcal{P}(1, \cdot), \ldots, -\mathcal{P}(T, \cdot), d(0), \ldots, d(T-1)]\), subject to the constraints (14) expressed in a corresponding matrix form together with a related set of constraints also in matrix form. The resulting optimal decision vector \(z^*\) yields the optimal state vectors \(y^*(1), \ldots, y^*(T)\) and optimal decision vectors \(u^*(0), \ldots, u^*(T-1)\).

**Multiskill Shortages and Overages**

Define \(\mathcal{F}_i := \text{capacity target for skill } i, \mathcal{O}_i := \text{expected shortage for skill } i, \Theta_i := \text{expected overage for skill } i, \mathcal{M}_{i,j} := \text{expected number of people capable of employing skill subset } j \text{ matched to employ skill } i, \) for all \(i \in \mathcal{I}, j \in \mathcal{J}\). Our objective is to determine the matchings \(\mathcal{M}_{i,j}\) of \(\mathcal{O}_i\) against \(\mathcal{F}_i\) to minimize a weighted sum of \(\mathcal{F}_i\) and \(\Theta_i\). People cannot be matched to a skill they do not possess; thus, \(\mathcal{M}_{i,j} = 0\) for all \(i \in \mathcal{I}\) that are not an element of \(j \in \mathcal{J}\), with \(\mathcal{M}_{i,j} \geq 0\) otherwise. All people need to be matched, implying \(\sum_{j=1}^{||J||} \mathcal{M}_{i,j} = \mathcal{O}_i\). The shortages and overages must satisfy the balance equations
\[
\sum_{j=1}^{||J||} \mathcal{M}_{i,j} + \mathcal{F}_i = \mathcal{F}_i + \Theta_i, \quad \forall i \in \mathcal{I}.
\]

To support preferences among subsets of skills, we define \(\mathcal{L}_{i,j} := \text{remaining number of people capable of employing skill subset } j \in \mathcal{J} \text{ not matched to employ skill } i \in \mathcal{I}\). This expands the above matching constraints to \(\sum_{j=1}^{||J||} \mathcal{M}_{i,j} + \mathcal{L}_{i,j} = \mathcal{O}_i\), where \(\mathcal{L}_{i,j} = 0\) for all \(i \in \mathcal{I}\) that are not elements of the subset \(j \in \mathcal{J}\) and \(\mathcal{L}_{i,j} \geq 0\) otherwise. Then, when \(j\) is preferred over \(j'\) for skill \(i\), we exploit the big-M method of linear programming (Bertsimas and Tsitsiklis 1997) and add \(M\mathcal{L}_{i,j} + \mathcal{L}_{i,j}\) in the objective function. More precisely, our general multiskill shortage and overage optimization with preferences for each stationary interval of the planning horizon consists of solving the linear program
\[
\min \sum_{i=1}^{||I||} (w_i^f \mathcal{F}_i + w_i^o \Theta_i) + \sum_{i,j} (\mathcal{M}_{i,j} \mathcal{L}_{i,j} + \mathcal{L}_{i,j}),
\]
subject to the foregoing constraints, where \(w_i^f\) and \(w_i^o\) are the weights associated with skill \(i\) shortages and overages. The solution to our overall multiskill shortage and overage analysis across multiple stationary intervals under a general stochastic modulation process is then obtained by solving the corresponding dynamic program.

**Acknowledgments**

We thank the following people for helpful comments on earlier drafts of the paper: B. Denton, J. Milne, D. Yao, S. Bollapragada, M. Campbell, A. Degbo, T. Ervolina, M. Ettl, and C. Hoffmann. We thank IBM executives B. Dietrich, A. Greenland, S. Johnson, K. Kitagawa, S. Kumar, A. Manuel, and E. Tsukada for their support throughout the Wagner Prize application process. Finally, we thank the following people for their collaborations and support throughout the development and deployment of OTM: R. Brown, M. Callahan, M. Collins, M. Eaton, A. Gill, S. Koh, W. Ng, A. Radovanović, B. Ramachandran, L. Sproull, A. Waechter, and M. Yan. The comments of the guest editor and editor in chief helped turn very different earlier versions of the paper into its more readable, so it is hoped, present form.

**References**


