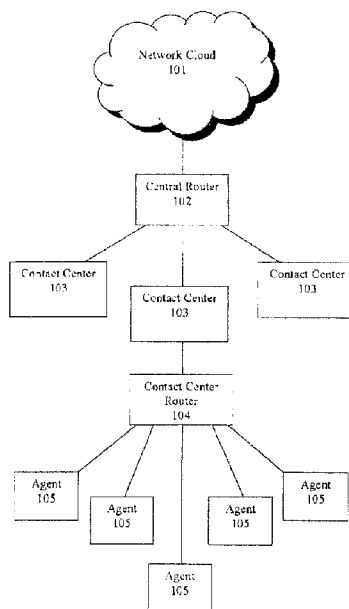




(22) **Date de dépôt/Filing Date:** 2013/03/21
(41) **Mise à la disp. pub./Open to Public Insp.:** 2013/10/03
(45) **Date de délivrance/Issue Date:** 2020/07/07
(62) **Demande originale/Original Application:** 2 868 802
(30) **Priorités/Priorities:** 2012/03/26 (US61/615,788);
2012/03/26 (US61/615,779); 2012/03/26 (US61/615,772);
2013/03/15 (US13/843,541)

(51) **Cl.Int./Int.Cl.** *H04M 3/523* (2006.01),
G06Q 10/06 (2012.01)
(72) **Inventeurs/Inventors:**
CHISHTI, ZIA, US;
KAN, ITTAI, US
(73) **Propriétaire/Owner:**
AFINITI INTERNATIONAL HOLDINGS, LTD., BM
(74) **Agent:** LAVERY, DE BILLY, LLP

(54) **Titre : SYSTEMES ET PROCEDES DE MAPPAGE D'APPELS UTILISANT UN ALGORITHME DE VARIANCE (VA) ET / OU UNE
COMPENSATION DE DISTRIBUTION**
(54) **Title: CALL MAPPING SYSTEMS AND METHODS USING VARIANCE ALGORITHM (VA) AND/OR DISTRIBUTION
COMPENSATION**



(57) **Abrégé/Abstract:**

Method, system and program product, comprising: obtaining agent parameter data; percentiling agents based on the agent parameter data, to obtain an agent distribution of agent percentiles; partitioning callers based on criteria into partitions; obtaining caller propensity data; percentiling the callers based on propensity for an outcome to obtain a caller distribution; performing distribution compensation using one algorithm selected from an edge compensation algorithm applied to the distribution of agent percentiles or the distribution of the caller percentiles, near at least one distribution edge to provide edge compensation, and a topology altering algorithm applied to either or both the agent distribution and the caller distribution to change one or more of the distributions to a different topology; and matching an agent to a caller in one of the partitions with a closest respective percentile, where one of the caller percentile or the agent percentile has been distribution compensated.

ABSTRACT

Method, system and program product, comprising: obtaining agent parameter data; percentiling agents based on the agent parameter data, to obtain an agent distribution of agent percentiles; partitioning callers based on criteria into partitions; obtaining caller propensity data; percentiling the callers based on propensity for an outcome to obtain a caller distribution; performing distribution compensation using one algorithm selected from an edge compensation algorithm applied to the distribution of agent percentiles or the distribution of the caller percentiles, near at least one distribution edge to provide edge compensation, and a topology altering algorithm applied to either or both the agent distribution and the caller distribution to change one or more of the distributions to a different topology; and matching an agent to a caller in one of the partitions with a closest respective percentile, where one of the caller percentile or the agent percentile has been distribution compensated.

CALL MAPPING SYSTEMS AND METHODS USING VARIANCE ALGORITHM (VA) AND/OR DISTRIBUTION COMPENSATION

BACKGROUND

[0001] The present invention relates generally to the field of routing phone calls and other telecommunications in a contact center system.

[0002] The typical contact center consists of a number of human agents, with each assigned to a telecommunication device, such as a phone or a computer for conducting email or Internet chat sessions, that is connected to a central switch. Using these devices, the agents are generally used to provide sales, customer service, or technical support to the customers or prospective customers of a contact center or a contact center's clients.

[0003] Typically, a contact center or client will advertise to its customers, prospective customers, or other third parties a number of different contact numbers or addresses for a particular service, such as for billing questions or for technical support. The customers, prospective customers, or third parties seeking a particular service will then use this contact information, and the incoming caller will be routed at one or more routing points to a human agent at a contact center who can provide the appropriate service. Contact centers that respond to such incoming contacts are typically referred to as "inbound contact centers."

[0004] Similarly, a contact center can make outgoing contacts to current or prospective customers or third parties. Such contacts may be made to encourage sales of a product, provide technical support or billing information, survey consumer preferences, or to assist in collecting debts. Contact centers that make such outgoing contacts are referred to as "outbound contact centers."

[0006] In both inbound contact centers and outbound contact centers, the individuals (such as customers, prospective customers, survey participants, or other third parties) that interact with contact center agents using a telecommunication device are referred to in this application as a “caller.” The individuals employed by the contact center to interact with callers are referred to in this application as an “agent.”

[0007] Conventionally, a contact center operation includes a switch system that connects callers to agents. In an inbound contact center, these switches route incoming callers to a particular agent in a contact center, or, if multiple contact centers are deployed, to a particular contact center for further routing. In an outbound contact center employing telephone devices, dialers are typically employed in addition to a switch system. The dialer is used to automatically dial a phone number from a list of phone numbers, and to determine whether a live caller has been reached from the phone number called (as opposed to obtaining no answer, a busy signal, an error message, or an answering machine). When the dialer obtains a live caller, the switch system routes the caller to a particular agent in the contact center.

[0008] Routing technologies have accordingly been developed to optimize the caller experience. For example, U.S. Patent No. 7,236,584 describes a telephone system for equalizing caller waiting times across multiple telephone switches, regardless of the general variations in performance that may exist among those switches. Contact routing in an inbound contact center, however, is a process that is generally structured to connect callers to agents that have been idle for the longest period of time. In the case of an inbound caller where only one agent may be available, that agent is generally selected for the caller without further analysis. In another example, if there are eight agents at a contact center, and seven are occupied with contacts, the switch will generally route the inbound caller to the one agent that is available. If all eight agents are occupied with contacts, the switch will typically put the contact on hold and then route it to the next agent that becomes available. More generally, the contact center will set up a queue of incoming callers and preferentially route the longest-waiting callers to the agents that become available over time. Such a pattern of routing contacts to either the first available agent or the longest-waiting agent is referred to as “round-robin” contact routing. In round robin contact routing, eventual matches and connections between a caller and an agent are essentially random.

[0009] Some attempts have been made to improve upon these standard yet essentially random processes for connecting a caller to an agent. For example, U.S. Patent No. 7,209,549 describes a telephone routing system wherein an incoming caller's language preference is collected and used to route their telephone call to a particular contact center or agent that can provide service in that language. In this manner, language preference is the primary driver of matching and connecting a caller to an agent, although once such a preference has been made, callers are almost always routed in "round-robin" fashion.

BRIEF SUMMARY OF THE EMBODIMENTS

[0010] Embodiments of a method are disclosed comprising: obtaining, by one or more computers, agent parameter data for a set of agents; ranking or percentiling, by the one or more computers, the agents based at least in part on the agent parameter data, to obtain an agent distribution of agent rankings or percentiles; partitioning, by the one or more computers, callers in a set of callers based on one or more criteria into a set of partitions; obtaining, by one or more computers, caller propensity data for the respective partitions; ranking or percentiling, by the one or more computers, the callers based at least in part on data relating to or predicting a caller propensity for a desired outcome based at least in part on the caller propensity data for the partitions of the respective callers, to obtain a caller distribution of caller rankings or percentiles; performing distribution compensation, by the one or more computers, using at least one algorithm selected from the group of: an edge compensation algorithm applied to at least one selected from the group of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles, near at least one edge of the respective distribution, to obtain edge compensated rankings or percentiles; and a topology altering algorithm applied to either or both of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles to change one or more of the distributions to a different topology; and matching, by the one or more computers, a respective one of the agents with a respective ranking or percentile to a respective one of the callers in one of the partitions with a closest respective ranking or percentile, where at least one of the caller ranking or percentile or the agent ranking or percentile has been distribution compensated.

[0011] In embodiments, the performing the distribution compensation step uses the edge compensation algorithm and provides edge compensation only to the agent rankings or percentiles.

[0012] In embodiments, the performing the distribution compensation step uses the edge compensation algorithm and provides edge compensation only to the caller rankings or percentiles.

[0013] In embodiments, the performing the distribution compensation step uses the edge compensation algorithm and provides edge compensation to both the caller rankings or percentiles and the agent rankings or percentiles.

[0014] In embodiments, the distribution compensation step uses the edge compensation algorithm and takes the agents that are free at runtime, and rescales the respective agent rankings or percentiles for these runtime available agents to provide more space/margin at the edges of the agent distribution.

[0015] In embodiments, the amount of the margin is based at least in part on a number of the agents that are free at runtime.

[0016] In embodiments, the distribution compensation step uses the edge compensation algorithm and takes the callers in a queue or other grouping at runtime and rescales the respective caller propensity rankings or percentiles, to provide more space/margin at the edges of the distribution.

[0017] In embodiments, the amount of the margin is based at least in part on a number of callers that are in the queue or in the grouping at runtime.

[0018] In embodiments, the distribution compensation step uses the edge compensation algorithm and weights multiple of the agents free at runtime near at least one edge of the agent distribution and weights multiple of the callers near at least one edge of the caller distribution to increase utilization of the agents at the at least one edge of the agent distribution.

[0019] In embodiments, the distribution compensation step uses the edge compensation algorithm and weights multiple of the agents free at runtime near both edges of the agent distribution or weights multiple of the callers near both edges of the caller distribution.

[0020] In embodiments, the partition for the callers is based at least in part on one or more selected from the group of demographic data, area code, zip code, NPANXX, VTN, geographic area, 800 number, and transfer number.

[0021] In embodiments, the agent performance data may comprise one or more selected from the group of sale, number of items sold per call, and revenue per call coupled with handle time.

[0022] In embodiments, the topology altering algorithm is used and may convert the distribution of the agent performances and/or the distribution of the caller to a circle topology.

[0023] In embodiments, the topology altering algorithm is used and may convert the distribution of the agent performances and the distribution of the caller to remove the edges of the distribution.

[0024] In embodiments, where there are multiple agents available and one caller, Kappa for the distribution of the agents may be greater than 1.0.

[0025] In embodiments, where there are multiple callers and one agent, the Rho applied to the callers may be greater than 1.0.

[0026] Embodiments of a system are disclosed comprising: one or more computers configured with program code that, when executed, causes performance of the steps: obtaining, by the one or more computers, agent parameter data for a set of agents; ranking or percentiling, by the one or more computers, the agents based at least in part on the agent parameter data, to obtain an agent distribution of agent rankings or percentiles; partitioning, by the one or more computers, callers in a set of callers based on one or more criteria into a set of partitions; obtaining, by one or more computers, caller propensity data for the respective partitions; ranking or percentiling, by the one or more computers, the callers based at least in part on data relating to or predicting a caller propensity for a desired outcome based at least in part on the caller propensity data for the partitions of the respective callers, to obtain a caller distribution of caller rankings or percentiles; performing distribution compensation, by the one or more computers, using at least one algorithm selected from the group of: an edge compensation algorithm applied to at least one selected from the group of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles, near at least one edge of the respective distribution, to obtain edge

compensated rankings or percentiles; and a topology altering algorithm applied to either or both of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles to change one or more of the distributions to a different topology; and matching, by the one or more computers, a respective one of the agents with a respective ranking or percentile to a respective one of the callers in one of the partitions with a closest respective ranking or percentile, where at least one of the caller ranking or percentile or the agent ranking or percentile has been distribution compensated.

[0027] In embodiments, a program product is provided comprising: a non-transitory computer-readable medium configured with computer-readable program code, that when executed, by one or more computers, causes the performance of the steps: obtaining, by the one or more computers, agent parameter data for a set of agents; ranking or percentiling, by the one or more computers, the agents based at least in part on the agent parameter data, to obtain an agent distribution of agent rankings or percentiles; partitioning, by the one or more computers, callers in a set of callers based on one or more criteria into a set of partitions; obtaining, by one or more computers, caller propensity data for the respective partitions; ranking or percentiling, by the one or more computers, the callers based at least in part on data relating to or predicting a caller propensity for a desired outcome based at least in part on the caller propensity data for the partitions of the respective callers, to obtain a caller distribution of caller rankings or percentiles; performing distribution compensation, by the one or more computers, using at least one algorithm selected from the group of: an edge compensation algorithm applied to at least one selected from the group of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles, near at least one edge of the respective distribution, to obtain edge compensated rankings or percentiles; and a topology altering algorithm applied to either or both of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles to change one or more of the distributions to a different topology; and matching, by the one or more computers, a respective one of the agents with a respective ranking or percentile to a respective one of the callers in one of the partitions with a closest respective ranking or percentile, where at least one of the caller ranking or percentile or the agent ranking or percentile has been distribution compensated.

[0028] The exemplary processes may further include or be supplemented by a pattern matching algorithm. For example, a pattern matching algorithm may use demographic data of the agents and/or callers to predict a chance of one or more outcome variables based on historical caller-agent pairings. The comparison via a pattern matching algorithm may be combined with the matching via corresponding agent performance and propensity of callers to determine a caller-agent match and routing decision.

[0029] Agent data may include agent grades or rankings, agent historical data, agent demographic data, agent psychographic data, and other business-relevant data about the agent (individually or collectively referred to in this application as “agent data”). Agent and caller demographic data can comprise any of: gender, race, age, education, accent, income, nationality, ethnicity, area code, zip code, marital status, job status, and credit score. Agent and caller psychographic data can comprise any of introversion, sociability, desire for financial success, and film and television preferences. It is further noted that certain data, such as an area code, may provide statistical data regarding probable income level, education level, ethnicity, religion, and so on, of a caller which may be used by the exemplary process to determine a propensity of caller for a particular outcome variable, for example.

[0030] The examples can be applied broadly to different processes for matching callers and agents. For instance, exemplary processes or models may include conventional queue routing, performance based matching (e.g., ranking a set of agents based on performance and preferentially matching callers to the agents based on a performance ranking or score), an adaptive pattern matching algorithm or computer model for matching callers to agents (e.g., comparing caller data associated with a caller to agent data associated with a set of agents), affinity data matching, combinations thereof, and so on. The methods may therefore operate to output scores or rankings of the callers, agents, and/or caller-agent pairs for a desired optimization of an outcome variable (e.g., for optimizing cost, revenue, customer satisfaction, and so on). In one example, different models may be used for matching callers to agents and combined in some fashion with the exemplary multiplier processes, e.g., linearly weighted and combined for different performance outcome variables (e.g., cost, revenue, customer satisfaction, and so on).

[0031] According to another aspect, computer-readable storage media and apparatuses are provided for mapping and routing callers to agents according to the various processes

described herein. Many of the techniques described here may be implemented in hardware, firmware, software, or combinations thereof. In one example, the techniques are implemented in computer programs executing on programmable computers that each includes a processor, a storage medium readable by the processor (including volatile and nonvolatile memory and/or storage elements), and suitable input and output devices. Program code is applied to data entered using an input device to perform the functions described and to generate output information. The output information is applied to one or more output devices. Moreover, each program is preferably implemented in a high level procedural or object-oriented programming language to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language.

BRIEF DESCRIPTION OF THE DRAWINGS

[0032] Figure 1 is a diagram reflecting the general setup of a contact center and its operation.

[0033] Figure 2 illustrates an exemplary routing system having a routing engine for routing callers based on performance and/or pattern matching algorithms.

[0034] Figure 3 illustrates an exemplary routing system having a mapping engine for matching callers to agents based on a probability multiplier process alone or in combination with one or more additional matching processes.

[0035] Figures 4A and 4B illustrate an exemplary probability matching process and random matching process, respectively.

[0036] Figures 5A and 5B illustrate exemplary probability matching processes for matching a caller to an agent.

[0037] Figures 6A and 6B illustrate exemplary three-dimensional plots of agent performance, client propensity, and the probability of a sale for particular caller-agent pairings.

[0038] Figure 7 illustrates an exemplary probability matching process or computer model for matching callers to agents based on probabilities of outcome variables.

[0039] Figure 8 illustrates an exemplary probability matching process or computer model for matching callers to agents based on probabilities of outcome variables.

[0040] Figure 9 illustrates an exemplary probability matching process or computer model for matching callers to agents based on probabilities of outcome variables.

[0041] Figure 10 illustrates a typical computing system that may be employed to implement some or all processing functionality in certain embodiments of the invention.

[0042] Figure 11 illustrates an exemplary probability matching process or computer model for matching callers to agents using a variance algorithm.

[0043] Figure 12 illustrates an exemplary probability matching process or computer model for matching callers to agents using distribution compensation.

[0044] Figure 13 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 20$, $Kappa = 1$ and no edge correction.

[0045] Figure 14 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 20$, $Kappa = 1$ and a topological edge correction (Alternating Circle Method).

[0046] Figure 15 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 20$, $Kappa = 1$ and the Adaptive Shift Method of edge correction.

[0047] Figure 16 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 10$, $Kappa = 1$ and the Static Shift Method v5 of edge correction.

[0048] Figure 17 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 20$, $Kappa = 1$ and the Static Shift Method v6 of edge correction.

[0049] Figure 18 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 40$, $Kappa = 1$ and the Static Shift Method v6 of edge correction.

[0050] Figure 19 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 10$, $Kappa = 1.4$ and the Adaptive Shift Method of edge correction..

[0051] Figure 20 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 10$, $Kappa = 1.4$ and no edge correction.

[0052] Figure 21 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 10$, $Kappa = 1.4$ and the Static Shift Method v5 of edge correction.

[0052] Figure 22 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 10$, $Kappa = 1.4$ and the Static Shift Method v6 of edge correction.

[0053] Figure 23 is an exemplary graph of agent utilization with $n_{agents} = 100$, $\mu = 10$, $Kappa = 1.4$ and the Alternating Circle Method of edge correction.

[0054] Figure 24 is an exemplary graph illustrating the time effects of mean revenue per call vs. days in a week in a month.

[0055] Figure 25 is an exemplary graph illustrating the time effects of mean revenue per call vs. days in one week.

[0056] Figure 26 is an exemplary graph illustrating the time effects of mean revenue per call vs. hours of the day.

[0057] Figure 27 is an exemplary graph illustrating normalized power spectral density vs. frequency (1/hour).

DETAILED DESCRIPTION OF EMBODIMENTS

[0058] Exemplary call mapping and routing systems and methods are described, for example, in U.S. Patent Application Nos. 12/267,471, entitled "Routing Callers to Agents Based on Time Effect Data," filed on November 7, 2008; 12/490,949, entitled "Probability Multiplier Process for Call Center Routing," filed on June 24, 2009; and 12/266,418, entitled, "Pooling Callers for Matching to Agents Based on Pattern Matching Algorithms," filed on November 6, 2008, U.S. Patent Application No. 12/051,251 filed on January 28, 2008, and U.S. Patent Application No. 12/267,471 filed on January 28, 2010, and provisional application no. 61/084,201 filed July 28, 2008.

[0059] The following description is presented to enable a person of ordinary skill in the art to make and use the invention, and is provided in the context of particular applications and their requirements. Various modifications to the embodiments will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the invention. Moreover, in the following description, numerous details are set forth for the purpose of explanation. However, one of ordinary skill in the art will realize that the invention might be

practiced without the use of these specific details. In other instances, well-known structures and devices are shown in block diagram form in order not to obscure the description of the invention with unnecessary detail. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

[0061] While the invention is described in terms of particular examples and illustrative figures, those of ordinary skill in the art will recognize that the invention is not limited to the examples or figures described. Those skilled in the art will recognize that the operations of the various embodiments may be implemented using hardware, software, firmware, or combinations thereof: as appropriate. For example, some processes can be carried out using processors or other digital circuitry under the control of software, firmware, or hard-wired logic. (The term “logic” herein refers to fixed hardware, programmable logic and/or an appropriate combination thereof, as would be recognized by one skilled in the art to carry out the recited functions.) Software and firmware can be stored on computer-readable storage media. Some other processes can be implemented using analog circuitry, as is well known to one of ordinary skill in the art. Additionally, memory or other storage, as well as communication components, may be employed in embodiments of the invention.

[0062] According to certain aspects of the present invention, systems and methods are provided for matching callers to agents within a call routing center based on similar rankings or relative probabilities for a desired outcome variable. In one example, an exemplary probability multiplier process include matching the best agents to the best callers, the worst agents to worst callers, and so on, based on the probability of a desired outcome variable. For instance, agents may be scored or ranked based on performance for an outcome variable such as sales, customer satisfaction, cost, or the like. Additionally, callers can be scored or ranked for an outcome variable such as propensity or statistical chance to purchase (which may be based on available caller data, e.g., phone number, area code, zip code, demographic data, type of phone used, historical data, and so on). Callers and agents can then be matched according to their respective rank or percentile rank; for example, the highest ranking agent matched with the highest ranking caller, the second highest ranked agent with the second highest caller, and so on.

[0063] The exemplary probability multiplier process takes advantage of the inherent geometric relationship of multiplying the different probabilities, for example, a 30% sales rate agent with a 30% buying customer (giving you a total chance of 9%) as opposed to matching a 20% or 10% sales rate agent with that same customer (resulting in a 6% or 3% chance). When used across all agents and callers, the process results in a higher overall predicted chance of a particular outcome variable, such as sales, than a random matching process.

[0064] In one example, in addition to using relative ranks of agents and callers to match callers to agents, a pattern matching algorithm using agent and/or caller demographic data may be used. For instance, agents and callers may be matched based on demographic data via a pattern matching algorithm, such as an adaptive correlation algorithm. The caller-agent matches from the probability multiplier algorithm and pattern matching algorithms can be combined, e.g., linearly combined with or without weightings, to determine a final match for routing a caller to an agent.

[0065] Initially, exemplary call routing systems and methods are described for matching callers to available agents. This description is followed by exemplary systems and methods for ranking or ordering callers and agents based on an outcome variable, e.g., sales, customer satisfactions, or the like, and matching agents to callers based on the relative rankings. For instance, matching the highest ranking agent for a particular outcome variable with the highest ranking caller for a particular outcome variable, matching the lowest ranking agent with the lowest ranking caller, and so on.

[0066] Figure 1 is a diagram reflecting the general setup of a typical contact center operation 100. The network cloud 101 reflects a specific or regional telecommunications network designed to receive incoming callers or to support contacts made to outgoing callers. The network cloud 101 can comprise a single contact address, such as a telephone number or email address, or multiple contact addresses. The central router 102 reflects contact routing hardware and software designed to help route contacts among call centers 103. The central router 102 may not be needed where there is only a single contact center deployed. Where multiple contact centers are deployed, more routers may be needed to route contacts to another router for a specific contact center 103. At the contact center level 103, a contact center router 104 will route a contact to an agent 105 with an individual telephone or other

telecommunications equipment 105. Typically, there are multiple agents 105 at a contact center 103.

[0066] Figure 2 illustrates an exemplary contact center routing system 200 (which may be included with contact center router 104 of Figure 1). Broadly speaking, routing system 200 is operable to match callers and agents based, at least in part and in one example, on a probability multiplier process based on agent performance and caller propensity (e.g., statistical chance or likelihood) for a particular outcome variable. Routing system 200 may further be operable to match callers based on pattern matching algorithms using caller data and/or agent data alone or in combination with the probability multiplier process. Routing system 200 may include a communication server 202 and a routing engine 204 for receiving and matching callers to agents (referred to at times as “mapping” callers to agents).

[0067] In one example, and as described in greater detail below, routing engine 204 is operable to determine or retrieve performance data for available agents and caller propensity for an outcome variable from callers on hold. The performance data and caller propensity data may be converted to percentile ranks for each and used to match callers to agents based on the closest match of percentile ranks, respectively, thereby resulting in high performing agents matched to callers with a high propensity to purchase, for example.

[0068] Additionally, in some examples, routing engine 204 may additionally include pattern matching algorithms and/or computer models, which may adapt over time based on the performance or outcomes of previous caller-agent matches. The additional pattern matching algorithms may be combined in various fashions with a probability multiplier process to determine a routing decision. In one example, a pattern matching algorithm may include a neural network based adaptive pattern matching engine as is known in the art; for example, a resilient backpropagation (RProp) algorithm, as described by M. Riedmiller, H.Braun: “A Direct Adaptive Method for Faster backpropagation Learning: The RPROP Algorithm,” Proc. of the IEEE Intl. Cont’. on Neural Networks 1993. Various other exemplary agent performance and pattern matching algorithms and computer model systems and processes which may be included with contact routing system and/or routing engine 204 are described, for example, in U.S. Serial No. 12/021,251, filed January 28, 2008, and U.S. Serial No. U.S. Patent application serial no. 12/202,091, filed August 29, 2008. Of course, it

will be recognized that other performance based or pattern matching algorithms and methods may be used alone or in combination with those described here.

[0069] Routing system 200 may further include other components such as collector 206 for collecting caller data of incoming callers, data regarding caller-agent pairs, outcomes of caller-agent pairs, agent data of agents, historical performance data of agents, and the like. Further, routing system 200 may include a reporting engine 208 for generating reports of performance and operation of routing system 200. Various other servers, components, and functionality are possible for inclusion with routing system 200. Further, although shown as a single hardware device, it will be appreciated that various components may be located remotely from each other (e.g., communication server 202 and routing engine 204 need not be included with a common hardware/server system or included at a common location). Additionally, various other components and functionality may be included with routing system 200, but have been omitted here for clarity.

[0070] Figure 3 illustrates further detail of exemplary routing engine 204. Routing engine 204 includes a main mapping engine 304, which may include one or more mapping engines therein for use alone or in combination with other mapping engines. In some examples, routing engine 204 may route callers based solely or in part on performance data associated with agents and caller data associated with the propensity or chances of a particular outcome variable. In other examples, routing engine 204 may further make routing decisions based solely or in part on comparing various caller data and agent data, which may include, e.g., performance based data, demographic data, psychographic data, type of phone/phone number, BTN-data, and other business-relevant data. Additionally, affinity databases (not shown) may be used and such information received by routing engine 204 and/or mapping engine 304 for making or influencing routing decisions. Database 312 may include local or remote databases, third party services, and so on (additionally, mapping engine 304 may receive agent data from database 314 if applicable for the particular mapping process).

[0071] In one example, relative agent performance may be determined by ranking or scoring a set of agents based on performance for a particular outcome variable (such as revenue generation, cost, customer satisfaction, combinations thereof, and the like). Further, the relative agent performance may be converted to a relative percentile ranking. Processing

engine 320-1, for example, may determine or receive relative agent performance data for one or more outcome variables. Additionally, processing engine 320-1 may receive or determine a propensity of a caller for a particular outcome variable (such as propensity to purchase, length of call, to be satisfied, combinations thereof, and the like). The propensity of a caller may be determined from available caller data. The relative performance data of the agents and propensity data of the callers may then be used to match a caller and an agent based on corresponding ranking. In some examples, the performance and propensity data is converted to relative percentile rankings for the callers and agents, and matching callers and agents based on the closest respective relative percentiles.

[0073] Processing engine 320-2, in one example, includes one or more pattern matching algorithms, which operate to compare available caller data with a caller to agent data associated a set of agents and determine a suitability score of each caller-agent pair. Processing engine 320-2 may receive caller data and agent data from various databases (e.g., 312 and 314) and output caller-agent pair scores or a ranking of caller-agent pairs, for example. The pattern matching algorithm may include a correlation algorithm such as a neural network algorithm, genetic algorithm, or other adaptive algorithm(s).

[0074] Additionally, a processing engine may include one or more affinity matching algorithms, which operate to receive affinity data associated with the callers and/or agents. Affinity data and/or affinity matching algorithms may be used alone or in combination with other processes or models discussed herein.

[0075] Routing engine 204 may further include selection logic (not shown) for selecting and/or weighting one or more of the plurality of processing engines 320-1 and 320-2 for mapping a caller to an agent. For example, selection logic may include rules for determining the type and amount of caller data that is known or available and selecting an appropriate processing engine 320-1, 320-2, etc., or combinations thereof. Selection logic may be included in whole or in part with routing engine 204, mapping engine 304, or remotely to both.

[0076] Further, as indicated in Figure 3 at 350, call history data (including, e.g., caller-agent pair data and outcomes with respect to cost, revenue, customer satisfaction, and so on) may be used to retrain or modify processing engines 320-1 and 320-2. For instance, the agent

performance data may be updated periodically (e.g., daily) based on historical outcomes to re-rank the agents. Further, historical information regarding callers may be used to update information regarding caller propensities for particular outcome variables.

[0076] In some examples, routing engine 204 or main mapping engine 304 may further include a conventional queue based routing processes, which may store or access hold or idle times of callers and agents, and operate to map callers to agents based on a hold time or queue order of the callers (and/or agents). Further, various function or time limits may be applied to callers on hold to ensure that callers are not held too long awaiting an agent. For instance, if a caller's time limit (whether based on a predetermined value or function related to the caller) is exceeded the caller can be routed to the next available agent.

[0077] Additionally, an interface may be presented to a user allowing for adjustment of various aspects of the exemplary systems and methods, for example, allowing adjustments of the number of different models, degrees, and types of caller data. Further, an interface may allow for the adjustment of the particular models used for different degrees or types, for example, adjusting an optimization or weighting of a particular model, changing a model for a particular degree or type of caller data, and so on. The interface may include a slider or selector for adjusting different factors in real-time or at a predetermined time. Additionally, the interface may allow a user to turn certain methods on and off, and may display an estimated effect of changes. For instance, an interface may display the probable change in one or more of cost, revenue generation, or customer satisfaction by changing aspects of the routing system. Various estimation methods and algorithms for estimating outcome variables are described, for example, in copending U.S. provisional Patent application serial no. 611084,201, filed on July 28, 2008. In one example, the estimate includes evaluating a past time period of the same (or similar) set of agents and constructing a distribution of agent/caller pairs. Using each pair, an expected success rate can be computed via the performance based matching, pattern matching algorithm, etc., and applied to current information to estimate current performance (e.g., with respect to one or more of sales, cost, customer satisfaction, etc.). Accordingly, taking historical call data and agent information the system can compute estimates of changing the balance or weighting of the processing methods. It is noted that a comparable time (e.g., time

of day, day of the week etc.) for the historical information may be important as performance will likely vary with time.

[0079] Figure 4A schematically illustrates an exemplary probability multiplier process for matching callers and agents and Figure 4B illustrates a random matching process (e.g., queue based or the like). These illustrative examples assume that there are five agents and five callers to be matched. The agents can be ranked based on performance of a desired outcome variable. For instance, the agents may be scored and ordered based on a statistical chance of completing a sale based on historical sales rate data. Additionally, the callers can be scored and ranked based on a desired outcome variable, for example, on a propensity or likelihood to purchase products or services. The callers may be ranked and ordered based on known or available caller data including, for example, demographic data, zip codes, area codes, type of phone used, and so on, which are used to determine a statistical or historical chance of the caller making a purchase.

[0080] The agents and callers are then matched to each other based on the ranking, where the highest ranked agent is matched to the highest ranked caller, the second highest ranked agent matched to the second highest ranked caller, and so on. Matching the best to the best and worst to the worst results in an increase product of the matched pairs compared to randomly matching callers to agents as shown in Figure 4B. For instance, using illustrative sales rates for agents A1-A5 (e.g., based on past agent performance) and the chance of callers C1-C5 making a purchase (e.g., based on caller data such as demographic data, caller data, and so on), the product of the matches shown in Figure 4A is as follows:

$$(0.09*0.21) + (0.07*0.12) + (0.06*0.04) + (0.05*0.03) + (0.02*0.02) 0.0316$$

[0081] In contrast, for a random matching, as illustrated in Figure 4B and using the same percentages, the product is as follows:

$$(0.09*0.12) + (0.07*0.02) + (0.06*0.21) + (0.05*0.03) + (0.02*0.04) 0.0271$$

[0082] Accordingly, matching the highest ranking agent with the highest ranking caller and the worst ranking agent with the worst ranking caller increases the overall product, and thus chances of optimizing the desired outcome variable (e.g., sales).

[0083] Figure 5A schematically illustrates an exemplary process for matching callers on hold to an agent that becomes free. In this example, all agents A1-A5 on duty or all that might become free within a reasonable hold time of callers C1-C5 are scored or ranked as previously described. Additionally, callers C1-C5 are scored or ranked as previously described. As an agent, e.g., agent A2 becomes free the process determines that caller C2 is the same (or similar) rank as agent A2 and caller C2 is matched thereto. The remaining callers on hold may then be re-ranked for matching when the next agent becomes free. Additionally, as new callers are placed on hold the callers can be reranked in a real-time fashion. The exemplary process operates in a similar fashion for multiple free agents and a caller becomes free (both for inbound and outbound call centers).

[0084] It will be recognized that in most instances the number of agents and callers will not be equal. Accordingly, the callers (and/or agents) can be ranked and converted to relative percentile rankings for the callers; for example, a normalized ranking or setting the highest ranked caller as the 100th percentile and the lowest ranked caller as the 0th percentile. The agents may be similarly converted to relative percentile rankings. As an agent becomes free, the agent may be matched to the caller having the closest relative percentile rank to the agent's relative percentile rank. In other examples, as an agent becomes free the agent can be compared to the ranking of at least a portion of callers on hold to compute Z-scores for each agent-caller pair. The highest Z-score may correspond to the smallest difference in relative percentile rankings. Further, as noted herein, the Zscores may be used to combine the matching with other algorithms such as pattern matching algorithms, which may also output a Z-score.

[0085] Figure 5B schematically illustrates exemplary methods for matching callers and agents when an agent becomes free and multiple callers are on hold. In this example, the callers (and in some examples the agents) are grouped in sub-groups of performance. For instance, a range of caller performance may be divided into multiple sub-groups and callers bucketed within each group. The top 20% of callers by performance might be grouped together as C1₁-C1_N as illustrated, followed by the next 20%, and so on. As an agent becomes free, e.g., A2, a caller from an appropriate sub-group is matched to the caller, in this example from C2₁-C2_N. Within the sub-group the caller may be chosen by a queue order,

best-match, a pattern matching algorithm, or the like. The appropriate sub-group from which to route a caller may be determined based on the agent ranking or score, for example.

[0086] In one example, suppose it is desired to optimize a call center performance for Outcome variable O . O can include one or more of sales rate, customer satisfaction, first call resolution, or other variables. Suppose further that at some time there are N_A agents logged in and N_C callers in queue. Suppose that agents have performances in generating O of

$$A_i^O \quad i = 1, \dots, N_A$$

and callers, partitioned by some property P , have a propensity to O of

$$C_i^O \quad i = 1, \dots, N_C$$

[0087] For example, in the case where O is sales rate and P is caller area code, A^O is each agent's sales rate and C^O is the sales rate for callers in a particular area code. Calculating the percentile ranked agent performances (with respect to the set of logged in agents) and the percentile ranked caller propensities (with respect to the set of callers in queue at some instant of time) as follows:

$$A_{Pi}^O = pr(A_i^O, A^O) \quad (i = 1, \dots, N_A)$$

$$C_{Pi}^O = pr(C_i^O, C^O) \quad (i = 1, \dots, N_C)$$

where $pr(a, B)$ is the percentile rank function which returns the rank of value a with respect to the set of values B scaled into the range $[0, 100]$.

[0088] Suppose that all the agents are on calls when the k 'th agent becomes available. Then to determine which caller in the queue they should be connected to, compute the difference between the percentile ranks of the newly free k 'th agent and those of the callers in queue:

$$D_j = A_{Pk}^O - C_{Pj}^O \quad (j = 1, \dots, N_C)$$

[0089] The value of j indexing the minimum element of the set $\{D_j\}$ gives the member of the queue to connect to the k 'th agent. A Z-score can also be derived from the D_j . This has the advantages that the highest value agent-caller pairing is the best fit of the set and that the output from this algorithm can be combined with Z-score outputs from other algorithms since they have the same scale.

$$Z_j = (T_j - \mu) / \sigma$$

where μ and σ and the mean and standard deviation of T which is given by:

$$T_j = \text{Min}(D_j) - D_j$$

[0090] It will be recognized by those of skill in the art that the above example and algorithm described for the case of two variables is not restricted to the case of two variables, but can be extended in an obvious way to the case of more than two variables which are monotonically related to the desired outcome. Furthermore the increase in call center performance can be shown to increase with more variables, as will be understood and contemplated by those of ordinary skill in the art.

[0091] Figure 6A illustrates an exemplary three-dimensional plot of agent performance versus caller propensity along the x-axis and y-axis and the probability of a sale along the z-axis. In this example, agent performance and caller propensity are defined as linear functions of x and y . For instance, without loss of generality, $x \in [0,1]$ and $y \in [0,1]$, such that the agent propensity is:

$$a = ca + ma x$$

where “ca” and “ma” represent the intercept and slope for the agent performance linear function (note that in certain instances herein, multiple letters represent a single variable e.g., “ca” and “ma” are each single variables, and are offset by other letters to indicate a multiplication operator). Similarly for the caller propensity:

$$c = cc + mc y$$

where “cc” and “mc” represent the intercept and slope for the caller propensity linear function. The multiplicative model probability of sale p of the product of a and c is:

$$p = a c$$

[0092] The average height d of a surface of the probability of a sale, which is graphically illustrated in Figure 6A as surface 600, can be computed as follows:

$$d = \int_0^1 \int_0^1 p \, dx \, dy = \frac{1}{4} (2ca + ma)(2cc + mc)$$

[0093] Determining the average height n of the diagonal of the surface height “pdiag” (a single variable), which corresponds to multiplying similar percentile ranking agents against callers is as follows:

$$p_{diag} = (ca + ma\lambda)(cc + mc\lambda);$$

$$n = \int_0^1 p_{diag} \, d\lambda = ca \, cc + \frac{ca \, mc}{2} + \frac{cc \, ma}{2} + \frac{ma \, mc}{3}$$

where λ parametrises the diagonal function so one can integrate down the line.

[0094] The boost b , or the potential increase in performance or sales rate according to the probability matching by corresponding rates, as illustrated by the diagonal shaded band 602 on surface 600, can be computed as follows:

$$b = n / d - 1 = \frac{4 \left(ca \, cc + \frac{ca \, mc}{2} + \frac{cc \, ma}{2} + \frac{ma \, mc}{3} \right)}{(2ca + ma)(2cc + mc)} - 1$$

where the theoretical maximum boost of matching according to probability, for this illustrative example, is 1/3. Accordingly, matching callers to agents on or near diagonal shaded band 602 increases the probability of sales.

[0095] Figure 6B illustrates an exemplary analysis and plot for normal distribution of callers and agents for agent performance and caller propensity (as opposed to a uniform

distribution of performance). Assuming the same definitions of agent performance and caller propensity, a two dimensional Gaussian function can be used to represent the distribution of call frequencies across the agents' performance and callers' propensities:

$$g2d(x, y, x0, y0, \sigma) = e^{\frac{-(x-x0)^2 - (y-y0)^2}{2\sigma^2}}$$

[0096] The sales rate can then be represented by a function of a and c, where A and σ give the amplitude and standard deviation of the Gaussian component respectively. Assuming that the Gaussian is centered at {0.5, 0.5}, the probability can be written as:

$$p = ac \left(1 + Ag2d\left(x, y, \frac{1}{2}, \frac{1}{2}, \sigma\right) \right) = (ca + max)(cc + mcx) \left(Ae^{\frac{\left(x-\frac{1}{2}\right)^2 + \left(y-\frac{1}{2}\right)^2}{2\sigma^2}} + 1 \right)$$

[0097] The diagonal sales rate, d, can then be determined directly from the sales rate as:

$$d(x, ca, ma, cc, mc, \sigma, x) = \left(Ae^{-\frac{\left(x-\frac{1}{2}\right)^2}{\sigma^2}} + 1 \right) (ca + max)(cc + mcx)$$

[0098] Integrating this with respect to x over [0,1] gives the sales rate for call agent pairs occurring on the diagonal giving:

$$\frac{1}{12} e^{-\frac{1}{4\sigma^2}} \left(e^{\frac{1}{4\sigma^2}} \left(3 \sqrt{\pi} A \sigma \operatorname{erf}\left(\frac{1}{2\sigma}\right) ((2ca + ma)(2cc + mc) + 2ma mc \sigma^2) + 6ca(2cc + mc) + 6cc ma + 4ma mc \right) - 6Ama mc \sigma^2 \right)$$

[0099] The sales rate for random client agent pairings can be computed as:

$$\text{totalsalesrate}[ca_ , ma_ , cc_ , mc_ , \sigma_ , A_] = \int_0^1 \int_0^1 \text{salesrate}[x, y, ca, ma, cc, mc, \sigma, A] dx dy$$

which expands to:

$$\frac{1}{4} (2 ca + ma) (2 cc + mc) \left(2 \pi A \sigma^2 \text{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right)^2 + 1 \right)$$

and the boost of the algorithm can be computed as follows:

$$\text{normalboost}[ca_ , ma_ , cc_ , mc_ , \sigma_ , A_] = \text{diagintegral}[ca, ma, cc, mc, \sigma, A] / \text{totalsalesrate}[ca, ma, cc, mc, \sigma, A] - 1$$

which results in a boost in sales of:

$$\frac{e^{-\frac{1}{4\sigma^2}} \left(e^{\frac{1}{4\sigma^2}} \left(3 \sqrt{\pi} A \sigma \text{erf}\left(\frac{1}{2\sigma}\right) ((2 ca + ma) (2 cc + mc) + 2 ma mc \sigma^2) + 6 ca (2 cc + mc) + 6 cc ma + 4 ma mc \right) - 6 A ma mc \sigma^2 \right)}{3 (2 ca + ma) (2 cc + mc) \left(2 \pi A \sigma^2 \text{erf}\left(\frac{1}{2\sqrt{2}\sigma}\right)^2 + 1 \right)}$$

-1

[0100] Accordingly, and similar to the normal distribution of Figure 6A, matching callers to agents on or near diagonal shaded band 602 increases the probability of sales. Of course, it will be understood that the exemplary functions, assumptions, and distributions of caller performance and agent propensity are illustrative, and will vary based on, e.g., historical data, feedback, and the like. Further, additional considerations and variables may be incorporated into the general processes. Note also that while referring to Sales as the variable to optimize in the above example, the same procedure can be applied to other variables or combinations thereof which are to be optimized such as call handle time (e.g. cost), or first call resolution, or many others.

[0101] To the extent that there is a discrepancy in any of the foregoing equations as compared to Application Serial Number 12/490,949, the equations of Application Serial Number 12/490,949 are the correct equations and take precedence. Figure 7 illustrates an exemplary process for matching callers to agents within a call routing center. In this example, agents are ranked based on a performance characteristic associated with an outcome variable such as sales or customer satisfaction at 702. In some examples agent performance may be determined for each agent from historical data over a period of time. In other examples, the method may merely retrieve or receive agent performance data or agent ranking for the agents.

[0102] In one example, agents are graded on an optimal interaction, such as increasing revenue, decreasing costs, or increasing customer satisfaction. Grading can be accomplished by collating the performance of a contact center agent over a period of time on their ability to achieve an optimal interaction, such as a period of at least 10 days. However, the period of time can be as short as the immediately prior contact to a period extending as long as the agent's first interaction with a caller. Moreover, the method of grading agents can be as simple as ranking each agent on a scale of 1 to N for a particular optimal interaction, with N being the total number of agents. The method of grading can also comprise determining the average contact handle time of each agent to grade the agents on cost, determining the total sales revenue or number of sales generated by each agent to grade the agents on sales, or conducting customer surveys at the end of contacts with callers to grade the agents on customer satisfaction. The foregoing, however, are only examples of how agents may be graded; many other methods may be used.

[0103] Callers are ranked or scored based on an outcome variable based on caller data at 704. Callers may be ranked or scored based on a predicted chance of a particular outcome based on known or available caller data. The amount and type of caller data may vary for each caller but can be used to determine a statistical chance for a particular outcome based on historical outcomes. For instance, the only data known for a caller might be an area code, which is associated with a particular propensity to purchase based on past interactions with callers from the particular area code. In some examples, there may be no data associated with the caller, in which case an average propensity or statistical chance for the particular outcome when no data is known may be used.

[0104] Callers and agents are then matched based on their respective rankings at 706. For example, matching the better agents to the better callers and so on as described. Additionally, to account for an uneven number of callers and agents either or both rankings can be adjusted or normalized and the callers and agents routed based on a closest match. For instance, the rank of an agent may be divided by the number of agents, and similarly for the callers, and the callers matched to agents based on a closest match (or within a certain range). The process may then route, or cause the routing, of the caller to the agent at 708. In other examples, the process may pass the match on to other apparatuses or processes that may use the match in other processes or use to weight with other routing processes.

[0105] Figure 8 illustrates another exemplary process for matching callers to agents within a call routing center. In this example, agents are ranked based on a performance characteristic associated with an outcome variable such as sales or customer satisfaction and converted to a relative percentile ranking at 702. For example, the raw performance values of the agents can be converted into a relative percentile ranking; for example, a 9% sales rate might be converted to an 85% performance ranking. In other examples, the raw performance values can be converted to a standardized score or Z-score.

[0106] Callers are ranked or scored based on an outcome variable based on caller data and converted to a relative percentile ranking at 804. Similar to that of the agents, raw predicted values for the callers can be converted into a percentile ranking; for example, a 20% propensity or likelihood to purchase might be converted to a 92% percentile ranking amongst callers. In other examples, the raw values can be converted to a standardized score or Z-score.

[0107] Callers and agents are then matched based on their respective relative percentile rankings at 806. For example, the relative percentile ranking of a caller can be compared to relative percentile ranking of agents and the caller matched to the closest agent available. In examples where an agent becomes free and multiple callers are on hold the agent may be matched to the closest matching caller. In other examples, a caller may be held for a predetermined time for the best matching agent to become free and then matched and routed to the closest matching agent.

[0108] It will be recognized that various other fashions of ranking callers and agents, and matching callers to agents based on their respective rankings, are contemplated. For example,

generally speaking, the exemplary processes result in higher ranking callers being routed to higher ranking agents and lower ranking callers being routed to lower ranking agents.

[0108] Figure 9 illustrates another exemplary process for matching callers to agents within a call routing center based on both a probability multiplier process and a pattern matching algorithm. The process includes determining relative agent performance of a set of agents for an outcome variable at 902 and determining relative caller propensity of a set of callers for the outcome variable at 904. The relative agent performance and relative caller propensity may further be normalized or converted to relative percentile rankings at 906.

[0109] A portion or all of available agent data and caller data may be passed through a pattern matching algorithm at 908. In one example, the matching algorithm includes an adaptive pattern matching algorithm such as a neural network algorithm that is trained on previous caller-agent pairing outcomes.

[0110] The matching algorithm may include comparing demographic data associated with the caller and/or agent for each caller-agent pair and computing a suitability score or ranking of caller-agent pairs for a desired outcome variable (or weighting of outcome variables). Further, a Z-score can be determined for each caller-agent pair and outcome variable(s); for instance, co-pending U.S. Patent Application No. 12/202,091, filed August 29, 2009, describes exemplary processes for computing Z-scores for caller-agent pairs.

[0111] Exemplary pattern matching algorithms and computer models can include a correlation algorithm, such as a neural network algorithm or a genetic algorithm. In one example, a resilient backpropagation (RProp) algorithm may be used, as described by M. Riedmiller, H.Braun: "A Direct Adaptive Method for Faster backpropagation Learning: The RPROP Algorithm," Proc. of the IEEE Intl. Conf. on Neural Networks 1993. To generally train or otherwise refine the algorithm, actual contact results (as measured for an optimal interaction) are compared against the actual agent and caller data for each contact that occurred. The pattern matching algorithm can then learn, or improve its learning of, how matching certain callers with certain agents will change the chance of an optimal interaction. In this manner, the pattern matching algorithm can then be used to predict the chance of an optimal interaction in the context of

matching a caller with a particular set of caller data, with an agent of a particular set of agent data. Preferably, the pattern matching algorithm is periodically refined as more actual data on caller interactions becomes available to it, such as periodically training the algorithm every night after a contact center has finished operating for the day.

[0113] The pattern matching algorithm can be used to create a computer model reflecting the predicted chances of an optimal interaction for each agent and caller matching. For example, the computer model may include the predicted chances for a set of optimal interactions for every agent that is logged in to the contact center as matched against every available caller. Alternatively, the computer model can comprise subsets of these, or sets containing the aforementioned sets. For example, instead of matching every agent logged into the contact center with every available caller, exemplary methods and systems can match every available agent with every available caller, or even a narrower subset of agents or callers. The computer model can also be further refined to comprise a suitability score for each matching of an agent and a caller.

[0114] In other examples, exemplary models or methods may utilize affinity data associated with callers and/or agents. For example, affinity data may relate to an individual caller's contact outcomes (referred to in this application as "caller affinity data"), independent of their demographic, psychographic, or other business-relevant information. Such caller affinity data can include the caller's purchase history, contact time history, or customer satisfaction history. These histories can be general, such as the caller's general history for purchasing products, average contact time with an agent, or average customer satisfaction ratings. These histories can also be agent specific, such as the caller's purchase, contact time, or customer satisfaction history when connected to a particular agent.

[0115] As an example, a certain caller may be identified by their caller affinity data as one highly likely to make a purchase, because in the last several instances in which the caller was contacted, the caller elected to purchase a product or service. This purchase history can then be used to appropriately refine matches such that the caller is preferentially matched with an agent deemed suitable for the caller to increase the chances of an optimal interaction. Using this embodiment, a contact center could preferentially match the caller with an agent who does not have a high grade for generating revenue or who would not otherwise be an acceptable match, because the chance of a sale is still likely given the caller's past purchase

behavior. This strategy for matching would leave available other agents who could have otherwise been occupied with a contact interaction with the caller. Alternatively, the contact center may instead seek to guarantee that the caller is matched with an agent with a high grade for generating revenue; irrespective of what the matches generated using caller data and agent demographic or psychographic data may indicate.

[0116] In one example, affinity data and an affinity database developed by the described examples may be one in which a caller's contact outcomes are tracked across the various agent data. Such an analysis might indicate, for example, that the caller is most likely to be satisfied with a contact if they are matched to an agent of similar gender, race, age, or even with a specific agent. Using this example, the method could preferentially match a caller with a specific agent or type of agent that is known from the caller affinity data to have generated an acceptable optimal interaction.

[0117] Affinity databases can provide particularly actionable information about a caller when commercial, client, or publicly-available database sources may lack information about the caller. This database development can also be used to further enhance contact routing and agent-to-caller matching even in the event that there is available data on the caller, as it may drive the conclusion that the individual caller's contact outcomes may vary from what the commercial databases might imply. As an example, if an exemplary method was to rely solely on commercial databases in order to match a caller and agent, it may predict that the caller would be best matched to an agent of the same gender to achieve optimal customer satisfaction. However, by including affinity database information developed from prior interactions with the caller, an exemplary method might more accurately predict that the caller would be best matched to an agent of the opposite gender to achieve optimal customer satisfaction.

[0118] Callers can then be matched to agents at 910 based on a comparison of relative rankings determined in 906 and the pattern matching algorithm at 908. For instance, outcomes of both processes may be combined, e.g., via a linear or non-linear combination, to determine the best matching caller-agent pair.

[0119] The selection or mapping of a caller to an agent may then be passed to a routing engine or router for causing the caller to be routed to the agent at 912. The routing engine or

router may be local or remote to a system that maps the caller to the agent. It is noted that additional actions may be performed, the described actions do not need to occur in the order in which they are stated, and some acts may be performed in parallel.

[0120] VARIANCE ALGORITHM MAPPING: A further embodiment is next described for call mapping based at least in part on a variance algorithm. The example and description below is generally described in terms of agent performance (AP), however, an analogous problem exists for estimating caller propensities, for instance to purchase, and the same methodology applies. Accordingly, to avoid repeating terms, examples will be expressed in terms of agent performances (AP) with it being understood that this could refer to caller propensity (CP) equally based on various partitions.

[0121] An exemplary call mapping and routing system can utilize three different mathematical types of target data. Binomial, for instance conversion rate (CR) that is sale/no sale, multinomial, e.g., a number of revenue generation per unit (RGU's) sold per call, and continuous, e.g., revenue per call and handle time. All the techniques described here apply to all 3 kinds of data though they need differences in the mathematical techniques used, particularly in the Bayesian Mean Regression (BMR) case, as will be recognized by those of ordinary skill in the art. Again, to avoid cluttering the argument with repetition, term CR is used throughout but this should be understood to be a stand in for binomial, multinomial, or continuous data.

[0122] Systems and methods are provided herein that can be used to improve or optimize the mapping and routing of callers to agents in a contact center, where the mapping and routing of callers may use performance based routing techniques. In one aspect of the present invention, a method and system attempts to map or assign high performing agents to callers belonging to groups wherein agent performance makes a large difference to call outcomes and map or assign poor performing agents to callers in groups where agent performance makes relatively less of a difference. In one example, the method and system assumes that a partitioning of the callers has been made, e.g., {P}, which may be defined in many ways. Possible partitions may be based at least in part on caller demographics, caller NPA (aka area code) or NPANXX (first 6 digits on phone number), or vector directory number (VDN), or geographic area, or 800 number, or transfer number, to name a few. By whatever method, the

callers are partitioned into sets for each of which various statistics for the partition may be calculated.

[0123] An exemplary process may be carried out as follows:

[0124] 1. Calculate the best estimate of Agent Performance (AP's), for instance by a Bayesian mean regression method or the like. Divide the set of agents by their AP values into a top half set of best performing agents {T} and a bottom half set of agents {B}. This may be done by splitting the agents at the median AP and assigning those with AP greater than the median to {T} and the remainder to {B}. Note that in embodiments, the agents may be divided in more than two splits or groups based on performance, or based on one or more other parameters relating to demographics or personality.

[0125] 2. Partition, by one or more computers, callers in a set of callers based on one or more criteria into a set of partitions. For each partition, P_i , calculate the mean Conversion Rate (CR) for calls in P_i taken by agents in {T} and perform the same calculation for agents in {B}. Then calculate a difference between these quantities, Δ . Partitions with a large Δ are those in which agents make a lot of difference, that is, high performing agents obtain a much higher CR than low performing agents. Conversely, partitions with small or zero Δ 's are ones where agents make little or no difference.

[0126] 3. Calculate a ranking or percentile for the partitions by Δ from the highest Δ to the lowest Δ , e.g., a partition in the 97th percentile has 97% of the calls in partitions with a lower Δ .

[0127] 4. Calculate a percentile of the agent performances (AP's). (These percentilings are by numbers of calls in each group, caller partition or agent, and assure that there are approximately equal numbers of each available at matching time, thus avoiding biasing the pool of agents or callers). Thus, in embodiments an agent with a 97th percentile has 97 percent of the calls going to agents with lower performance ranking.

[0128] 5. In embodiments, calls may be assigned to agent - caller pairs where the difference between the percentile Δ for the partition of the caller and the percentile of the agent's AP is minimized. Thus, in embodiments, calls from partitions with a higher Δ will be matched to higher performing agents and calls with a lower Δ will be matched to lower performing agents.

[0129] Referring to Fig. 11, implementations of embodiments are disclosed. Block 1100 represents an operation of obtaining, by one or more computers, agent performance data for a set of agents. In embodiments, the agent performance data may comprise one selected from the group of sale/no sale, number of items sold per call, and revenue per call coupled with handle time. In embodiments, this performance data may be obtained by accessing a database containing such performance data

[0130] Block 1110 represents an operation of ranking, by the one or more computers, the agents based at least in part on the agent performance data.

[0131] Block 1120 represents an operation of dividing, by the one or more computers, the agents into agent performance ranges based at least in part on the ranking step.

[0132] Block 1130 represents an operation of partitioning, by the one or more computers, callers in a set of callers based on one or more criteria into a set of partitions. In embodiments, the partition may be based at least in part on one or more selected from the group of demographic data, area code, zip code, NPANXX, VTN, geographic area, 800 number, and transfer number, to name a few.

[0133] Block 1140 represents an operation of determining for each of the partitions, by the one or more computers, an outcome value at least for a first one of the agent performance ranges and an outcome value for a second one of the agent performance ranges. In embodiments, the outcome value comprises one or more selected from the group of sale, number of items sold per call, and revenue per call coupled with handle time. In embodiments, the determining an outcome value step may comprise determining a mean outcome value for a particular outcome for the agent performance range.

[0134] Block 1150 represents an operation of calculating, by the one or more computers, for each of the partitions an outcome value difference indicator based at least in part on a difference between the outcome value for the first agent performance range and the outcome value for the second agent performance range.

[0135] Block 1160 represents an operation of matching, by the one or more computers, a respective agent with respective performance data to a respective caller in one of the partitions, based at least in part on the outcome value difference indicators for the partitions.

[0136] In embodiments, the matching step may be based at least in part on a rule to assign a higher performing agent to a caller from one of the partitions where the outcome value

difference indicator is higher relative to other of the outcome value difference indicators, wherein the higher performing agent is determined based at least in part on the agent performance data for the respective agent relative to the agent performance data for other of the agents. In embodiments, one or more threshold values may be used for the outcome value difference indicators and different ranges of agent performances may be associated with these respective thresholds. In embodiments, rankings or percentiles may be used, and agents with rankings or percentiles that are closest to the ranking or percentile of the outcome value difference indicator for the caller's partition may be matched.

[0137] In embodiments, the matching step further comprises the steps of calculating, by the one or more computers, a ranking or percentile for the partitions by Δ from a high Δ to a low Δ based at least in part on a first number of number of calls; and calculating, a ranking or percentile of the agent performances (AP's), by the one or more computers, for the respective agents in the set of agents based at least in part on the respective agent performance data and a second number of calls. In embodiments, these percentilings or rankings are by numbers of calls in each group, caller partition or agent, and assure that there are approximately equal numbers of each available at matching time, thus avoiding biasing the pool of agents or callers. In embodiments, the matching step may further comprise matching, by the one or more computers, the respective agent to the respective caller based at least in part on a rule to minimize a difference between the partition percentile or ranking of the outcome value difference indicator for the partition of the respective caller and the performance percentile or ranking of the respective agent.

[0138] In embodiments, a mix a performance data may be used. In embodiments, higher dimensional matching may be performed, wherein the matching step is performed for a conversion rate or a conversion rate delta, and a handle time, or a customer satisfaction score.

[0139] In embodiments, a mix of matching algorithms may be used. For example, the one or more computers may be configured with program code to perform, when executed, the steps of: switching, by the one or more computers, between the outcome value difference indicator matching algorithm and a second matching algorithm that is different from the outcome value difference indicator matching algorithm, based on one or more criteria. In this example, when there has been switching to the second matching algorithm, then performing, by the one or more computers, the second matching algorithm to match a respective one of

the agents with a respective one of the callers. In embodiments, the second matching algorithm may be a pattern matching algorithm, or a closest percentile or ranking matching of agents and callers, or other matching type algorithm. In embodiments, the system may switch to a random matching algorithm or a pure queue based algorithm to demonstrate the relative effectiveness of the use of the outcome value difference indicator matching algorithm as compared to a random or queue-based matching.

[0140] In embodiments, the one or more computers are configured with program code to perform, when executed, the steps of: obtaining results data, by the one or more computers, using each of the outcome value difference indicator matching algorithm and the second matching algorithm in the matching step; obtaining or receiving, by the one or more computers, a switch-over point in the distribution for agent performance and/or the distribution for caller propensity where better results are obtained from one of the algorithms relative to the other of the algorithms; and using, by the one or more computers, this switch-over point to switch between the algorithms. The switch point may be determined empirically, by reviewing the performance data of the algorithm models based on the two different types of performance data. This review may be performed manually or may be determined automatically, on a periodic or aperiodic basis or at runtime, based on a comparison of performance results and a determination that a predetermined difference threshold between the performance results for the respective algorithms has been exceeded. In embodiments, the use of a mix of performance data may also be used for handle time, or customer satisfaction, or another parameter.

[0141] In embodiments, a pure conversion rate ranking or percentile may be used for low agent conversion rates (e.g., poor sales). Then the system may switch to using rankings or percentiles for delta. Accordingly, in embodiments, the one or more computers may be configured with program code to perform, when executed, the steps of: switching, by the one or more computers, between the outcome value difference indicator matching algorithm and a second matching algorithm configured to match a respective agent with a respective ranking or percentile to a caller in one of the partitions with a closest ranking or percentile, based on one or more criteria; and when there has been switching to the second matching algorithm, then performing, by the one or more computers, performing the second matching algorithm to match a respective one of the agents with a respective one of the callers..

[0142] TIME EFFECTS COMPENSATION: A further time effects embodiment for call mapping is disclosed that compensates for time effects which may stand on its own or be incorporated into other embodiments herein. As noted previously, the example and description below is generally described in terms of agent performance (AP), however, an analogous problem exists for estimating caller propensities, for instance to purchase, and the same methodology applies. Accordingly, to avoid repeating terms, examples will be expressed in terms of agent performances (AP) with it being understood that this could refer to caller propensity (CP) equally.

[0143] Conversion Rates (CR's) in call centers typically vary over time. For example, there are periodic changes, such as variations in CR through the day, day of week effects, and longer cycles. There may also be secular changes. Causes for the latter include marketing campaigns, seasonal effects such as spending in the holiday season, and the state of the economy.

[0144] These effects, if not compensated for, can impact accurate calculation of both Agent Performance (AP) and CP. For example, if a particular call center had a higher CR for nighttime calls than during the daytime, then agents who only worked a night shift would appear to have a higher performance than those that worked during the day. But this higher performance would be an artifact and if uncorrected, would result in an inaccurate AP. See Figs. 12-14.

[0145] Systems and methods are provided herein that can be used to improve or optimize the mapping and routing of callers to agents in a contact center, where the mapping and routing of callers may use performance based routing techniques. In one aspect of the present invention, systems and methods are provided for mapping callers to agents based on agent and/or caller data, wherein the agent and/or caller data includes or is based on timing effects (e.g., time data or information that affects one or more of the agent and/or caller data used to determine the mapping of callers to agents. For instance, agent data and caller data utilized by a pattern matching algorithm may include time effect data associated with performance, probable performance, or output variables as a function of one or more of time of day, day of week, time of month, time of year, to name a few. The pattern matching algorithm may operate to compare caller data associated with each caller to agent data associated with each agent to determine an optimal matching of a caller to an agent, and include an analysis of

time effects on the performance of agents or probable outcomes of the particular matching when performing the matching algorithm.

[0145] In embodiments, secular trends in CR may be detected by smoothing of the call center data, for instance with a one day or one week moving window average. Periodic effects may be detected both by a power spectral analysis techniques such as the FFT or Lomb Scargle Periodogram (William H. Press and George B. Rybicki Fast Algorithm for Spectral Analysis of Unevenly Sampled Data *Astrophysical Journal*, 338:277-280, March 1989), day of week CR, hour of day CR, across many epochs. For example, a data sample of 3 months of data may be used, and a mean and standard errors (SE) of the mean may be calculated for all the Mondays, Tuesdays, etc.), and by simply computing means and standard errors of the mean.

[0146] Figs. 24-27 show evidence of hour of day, day of week and other cycle timing effects and exemplify the above. Fig. 24 illustrates mean revenue variations over a month. Fig. 25 illustrates mean revenue variations over days of the week. Fig. 26 illustrates mean revenue variations over hours of the day. Fig. 27 illustrates a normalized power spectral density vs. frequency (1 hour).

[0147] Exemplary Methods of Correcting for Secular Trends are provided. Daily means of CR from a period, typically some months long, are fitted either with a linear least squares regression or a locally weighted least squares smoother like lowess (see e.g., Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* 74, 829-836) or loess. Corrected CR data can then be obtained by dividing the raw daily time series by the fitted values.

[0148] Exemplary Methods of Correcting for Periodic Variations are also provided. Epoch averages of CR (conversion rate) by hour of day (HOD), or by day of week (DOW) or by hour of week (HOW), which combines HOD and DOW effects in one variable, may be computed from a suitably long data sample.

[0149] Below are 3 exemplary methods of correction:

[0150] 1. Additive: A new target variable is defined as the raw target variable less the epoch average for the epoch in which the call occurs. In this method and the methods to follow, the term "target variable" means whatever variable the call mapping system is optimizing. In the case of conversion rate, it would be {O, I}, in the case of revenue

generating units (RGU), the number of RGUs per call, or in the revenue case the revenue per call.

[0152] 2. Multiplicative: A new target variable is defined as the raw target variable divided by the epoch average for the epoch in which the call occurs.

[0153] 3. Microskill Method: The relevant time effect, HOD, DOW, or HOW is combined with the skills (or VDNs) used in the AP calculation by forming the outer product. For example, if are 2 skills, A and B, and one were considering the DOW correction only, one could form $2 \times 7 = 14$ new microskills comprising A & Monday, A & Tuesday, ... , A & Sunday, B & Monday, B & Tuesday, ... , B & Sunday. These 14 microskills would then be used as the factor in the BMR calculation of agent performance. For example, when time effects are applied at the skill level or the VDN level, and there are 10 skill for an agent that are measured, then a time effects adjustment for the respective agent may be made for each hour of the 24 day for each of the 10 skills, e.g., $10 \times 24 = 240$ time effects adjustments made to the performance data, i.e., a microskill adjustment.

[0154] When corrected data with a new target variable has been computed, AP and CP and all other mapping calculations may be done using the corrected target variable.

[0155] In embodiments, a method for correcting, by the one or more computers, for time effects comprises subtracting from a target outcome variable an epoch average from a desired outcome data sample, wherein the epoch average comprises an average of one selected from the group of hour-of-day, day-of-week, and hour-of-week desired outcome data from an epoch in which a respective call occurs, to obtain a new target outcome variable.

[0156] In embodiments, a method for correcting, by the one or more computers, for time effects comprises dividing a target outcome variable by an epoch average from a desired outcome data sample, wherein the epoch average comprises an average of one selected from the group of hour-of-day, day-of-week, and hour-of-week desired outcome data from an epoch in which a respective call occurs, to obtain a new target outcome variable

[0157] In embodiments, a method for correcting, by the one or more computers, for time effects comprises forming an outer product by combining selected epoch data, such as hour-of-day, day-of-week, and hour-of-week desired outcome data from an epoch in which a respective call occurs, to obtain a factor to be used in the Bayesian Mean Regression (BMR) calculation of agent performance.

[0158] DISTRIBUTION COMPENSATION: In embodiments, matching algorithms may be used to match agents to callers where a given number reflecting a parameter of an agent is matched to a caller with the closest number reflecting a different parameter of the caller. For example, in embodiments, the number for the agent may be a rank or a percentile reflecting the performance of the respective agent relative to other agents in a set of agents. Likewise, the number for the caller may be a rank or percentile of the caller relative to other callers in a queue or other grouping, such as for caller propensity for something. In embodiments, the rank or percentile may be for a partition (e.g., demographic, zip code, area code, etc.) of the callers based at least in part on whether or not the level of performance of the agent makes a difference in the outcome of the call.

[0159] When such algorithms are used, it has been discovered that, using the example of agent performance percentiles, the matching algorithm tends to assign/cluster callers to agents in the middle percentiles in the distribution of performances at the expense of agents at the ends of the distribution (e.g., the worst agents and the best agents). In embodiments, to compensate for this effect, an edge compensation algorithm is provided to increase the likelihood that agents at the edges of the performance distribution are utilized more. In one embodiment, the edge compensation algorithm takes the agents that are free at runtime, and rescales the respective agent performances for these runtime available agents to provide more space/margin at the edges of the performance distribution, e.g., weighting the performance numbers for the worst agents to increase their respective performance levels and/or to decrease the respective performance levels of the best agents. The amount of the margin of weighting is based on the number of agents that are free at runtime. For example, if there are n agents working of which k are free to take the call, we linearly rescale the n agent percentiles so that the bottom agent has percentile $100/(2*k)$ and the top agent has percentile $100*(1-(2/k))$. In another embodiment, the edge compensation algorithm takes the callers in a queue or other grouping and rescales the respective caller propensity rankings or percentiles, to provide more space/margin at the edges of the distribution, e.g., weighting the propensity numbers for the worst callers to increase their respective propensity levels and/or to decrease the respective propensity levels of the best callers. The amount of the margin of weighting is based on the number of callers that are in the queue or in the grouping at runtime.

[0160] Various examples of edge compensation algorithm are now provided.

[0161] Notation & Setup

Percentiles have all been divided by 100 and lie between 0 and 1.

This experimentation is for a one dimensional unit interval caller to agent matching algorithm, where the goals are to:

1. Minimize the average difference between agent and caller percentiles.
2. In the absence of explicit interpolation towards performance routing have a uniform utilization of agents or average wait time for callers.
3. Incorporate a simple algorithm to interpolate between uniform utilization of agents and performance based routing.

L1 = A situation where there are multiple free agents and a single call arrives and is routed to one of the free agents

L2 = A situation where there are multiple calls in the queue and a single agent becomes available. One of the calls in the queue is routed to the free agent.

Kappa = In the L1 situation: A method of interpolating between pure performance routing (when a call comes it is assigned to the best performing free agent) and a more uniform agent utilization, An embodiment is the rescaling of agent percentiles by raising the percentile to the kappa power for some kappa greater than 1.

Rho = In the L2 situation: A method for when there are of interpolating between pure performance routing (When an agent becomes comes available and is assigned to the best call in the queue. In this case calls with low percentile may have waiting times much longer than calls with high percentiles) and a more uniform call expected waiting time algorithm. ,An embodiment is the rescaling of agent percentiles by raising the percentile to the rho power for some rho greater than 1.

[0162] The code to experiment was written in R.

T = time in seconds, time is discrete

method =has values *now* for the current methodology, *lah* for Look Ahead Method, *asm* for Adaptive Shift Method, *ssm* for Static Shift Method, *ssm6* for Static Shift Method v6, *adm* for Average Distance Method, and *acl* for Alternating Circle, for the new methods being analyzed

nagents = number of agents

aht = average handling time

mu = equilibrium average number of free agents

base_ap = agent percentile before any adjustments or transformations, evenly distributed in [0,1) starting at $1 / (2 * nagents)$

base_cp = caller percentile before any adjustments or transformations, uniformly distributed in [0,1)

method_ap = vector of agent percentiles after any adjustment or transformation due to the method applied.

method *cp* = vector of callers percentiles who are in the queue waiting for agents in the order in which they entered the queue after any adjustment or transformation due to the method applied.

base_ap[freeagents] = the vector of agent percentiles for free agents only

base_cp[freeagents] = the vector of caller percentiles for call in the queue

[0163] Adaptive Shift Method L1

Method: This is the situation where there are one or more free agents and at most one call in the queue. At any point in time, shift the percentiles of the agents by $\{(1 - \max(\text{base_ap}[\text{freeagents}])) - \min(\text{base_ap}[\text{freeagents}])\}/2$. That is:

$$\text{asm_ap} = \text{base_ap} + (1 - \max(\text{base_ap}[\text{freeagents}]))/2 - \min(\text{base_ap}[\text{freeagents}])/2$$

[0164] Theory: The basis for this method is the observation that there would be no edge effect if there were no edge. That is, if the agent and caller percentiles were evenly distributed along a circle of length 1. Naively applied, this would result in some awful matches as callers with *base_cp* close to 1 might be matched with agents whose *base_ap* is close to 0. The critical observation for avoiding this is that under the naïve adaptive shift methodology, the probability that the call is assigned to the i^{th} free agent is $(\text{asm_ap}_{i+1} - \text{asm_ap}_{i-1})/2$ for an internal (neither the top (n^{th}) nor the bottom (i^{th})) free agent and $(1 - \text{asm_ap}_n + \text{asm_ap}_2)/2$ for the bottom agent and $(\text{asm_ap}_1 + 1 - \text{asm_ap}_{n-1})/2$ for the top free agent. If the free agent percentiles are rigidly shifted at runtime so that the bottom and top edges are the same, then the probabilities are preserved and the zero call is always matched to the bottom free agent and the 1 call is always matched to the top free agent. A downside of this method is that some interior matches will be less than optimal due to the shift.

[0165] Kappa: Implementation of Kappa to first order, $\text{Kappa} > 1$ results in higher utilization of the top agents and lower utilization of the bottom agents. However, the adaptive shift means that this imbalance in utilization results in a larger edge region allocated to the bottom – which results in a second order increase in utilization of the bottom. Numerical simulation confirms this, resulting in an edge tendency to the mean. Numerous functional forms were numerically attempted to compensate for the secondary effects. The final functional form selected was:

```

ap = base_ap - min(base_ap[freeagents]) +
(min(base_ap[freeagents])/1)^(kappa-1)*0.5*(min(base_ap[freeagents])+(1-
max(base_ap[freeagents])))
ap = ap^kappa.

```

[0166] Adaptive Shift Method L2:

Method: This is the situation where there are one or more callers in the queue and at most one free agent. At any point in time, shift the percentiles of the call queue by $\{(1 - \max(\text{base_cp})) - \max(\text{base_cp})\}/2$. That is:

```
asm_cp = base_cp + (1-max(base_cp))/2 - max(base_cp)/2.
```

[0167] Theory: The basis for this method is the observation that there would be no edge effect if there were no edge. That is, if the agent and caller percentiles were evenly distributed along a circle of length 1. Naively applied, this would result in some awful matches as callers with base_cp close to 1 might be matched with agents whose base_ap is close to 0. When there is a single agent, the adaptive shift method prevents these awful matches by rotating the callers' percentiles so that the agent is equally likely to be to the left of the caller with the lowest base_cp or to the right of the caller with the highest base_cp . This results in the same probability that any give caller in the queue is matched as when done on the circle. The downside of this method is that some interior matches will be less than optimal.

Rho: Rho is simple to implement with the formula below and there are no complications, although there is a slight decrease in utilization of the top agents.

```
asm_cp with rho = asm_cp^rho.
```

[0168] Note: When there is a single free agent and multiple callers in the queue, one should not apply a shift to the agent's percentile because that shift will move the agent's percentile to 0.5 and automatically match to the median percentile caller. One should leave the agent percentile unchanged and only shift the callers' percentile. Similarly, when there is

a single caller and multiple free agents, one should only shift the agents' percentile. When there are multiple free agents and multiple callers in the queue many possibilities come to mind three are: 1) no shift until enough callers and agents have been paired so that there is either one caller or one agent left unmatched and then revert to the adjusted shift method; or 2) shift the larger group until enough callers and agents have been paired so that there is either one caller or one agent left unmatched and then revert to the adjusted shift method; or 3) shift both groups until enough callers and agents have been paired so that there is either one caller or one agent left unmatched and then revert to the adjusted shift method. The particular method chosen will be dependent on a number of factors, including, but not limited to: match accuracy; simplicity; abandoned call behavior; and external service level agreements (e.g. limiting call wait time or regarding agent utilization).

[0169] Note (Theoretical Performance Limits): Because there is interaction between the top and bottom of the agent stack, it is clear that this method is not the best possible utilization homogenization solution from the view of accuracy of match.

[0170] Static Shift Method: L1

Method: This is the situation where there are one or more free agents and at most one call in the queue. At any point in time, rescale the percentiles of the agents to a length of $(1 - 1/\text{count_freeagents})$ and shift the percentiles by $1/(2 * \text{count_freeagents})$. That is:

$$\text{ssm_ap} = \text{base_ap} * (1 - 1/\text{count_freeagents}) + 1/(2 * \text{count_freeagents})$$

[0171] Theory: This method is an attempt to address the major weakness of the Adaptive Shift Method. That is -- the direct interaction between the top and bottom agents which lowers match performance and makes kappa difficult to implement. We attempt to average out the cumulative effect of the adaptive shifts and perform a static affine transformation. The transformation is quite ad hoc. Additional variations are possible.

Kappa: Kappa is simple to implement via the formula below and there are no complications, although there is a slight decrease in utilization of the top agents.

$$\text{ssm_ap with kappa} = \text{ssm_ap}^{\text{kappa}}$$

[0172] Static Shift Method: L2:

Method: The same theory as for the Adaptive Shift Methodology.

Theory: The same method as for the Adaptive Shift Methodology.

Rho: Rho is simple to implement with the formula below and there are no complications, although there is a slight decrease in utilization of the top agents.

$ssm_cp \text{ with } \rho = ssm_cp^{\rho}.$

[0173] Static Shift Method: v5 and v6: L1:

Method: This is the situation where there are one or more free agents and at most one call in the queue. At any point in time, first rescale the base_ap to start at 0 and end at 1 rather than start at $1/(2*nagents)$ and end at $1 - (1/(2*nagents))$, second rescale the percentiles of the agents to a length of $(1 - 1/count_freeagents)$ and shift the percentiles by $1/(2*count_freeagents)$. That is:

$ssm6_ap = (base_ap - 1/(2*nagents)) * (1 - 1/nagents) * (1 - 1/count_freeagents) + 1/(2*count_freeagents).$

Theory: This method is an attempt to address the major weakness of the Static Shift Method. A pattern emerges in the utilization when there are relatively few free agents. A number of variations were tested – through v8, but v5 and v6 are listed here. The transformation selected is quite simple although it is ad hoc. Some patterns remain, although they are muted.

Kappa: For v5, kappa is applied directly.

$asm5_ap \text{ with } \kappa = asm5_ap^{\kappa}$

However, it was found that the edge buffers inserted by ssm5 should be adjusted when kappa is applied – otherwise the utilization of the bottom agent is too high and of the top agent is

too low. An ad hoc adjustment that seemed to work reasonably well is to divide the bottom edge buffer by kappa and add the difference to the top edge buffer. This is what was chosen for ssm6.

asm6_ap with kappa = (asm6_ap - (1/(2*count_freeagents)) + (1/(2*count_freeagents))/(kappa))^kappa.

[0174] Static Shift Method: v5 and v6: L2:

Method: The same theory as for the Adaptive Shift Methodology.

Theory: The same method as for the Adaptive Shift Methodology.

Rho: For v5, rho is applied directly.

asm5_cp with rho = asm5_ap^rho.

[0175] Average Distance Method: L1:

Method: This is the situation where there are one or more free agents and at most one call in the queue. This method is a variation of the Adaptive Shift and Static Shift Methods that attempts to adjust the affine shift to the configuration at each point in time. The point of view is that all gaps should be considered in determining the edge adjustment – not just the two edge gaps.

```
avg_distance = 0
for(j in 1:(length(ap_freeagents)-1))
{
  avg_distance = avg_distance + ((ap_freeagents[j+1]-ap_freeagents[j])^2)/2
}
avg_distance = avg_distance + ((min(ap_freeagents)+(1-max(ap_freeagents)))^2)/2
ap = (base_ap - min(ap_freeagents))*(1-2*avg_distance)/(max(ap_freeagents)-min(ap_freeagents))
+ avg_distance.
```

Theory: This method is an attempt to address the major weakness of the Adaptive Shift Method. That is – the direct interaction between the top and bottom agents which lowers match performance and makes kappa difficult to implement. This method is a variation of the Adaptive Shift and Static Shift Methods that attempts to adjust the affine shift to the

configuration at each point in time. The point of view is that all gaps should be considered in determining the edge adjustment – not just the two edge gaps. This method likely could be improved with tweaks.

Kappa: For adm, kappa is applied directly.

adm_ap with kappa = $\text{adm_ap}^{\text{kappa}}$

[0176] Average Distance Method: L2:

Method: The same theory as for the Adaptive Shift Methodology.

Theory: The same method as for the Adaptive Shift Methodology.

Rho: For adm, rho is applied directly via the formula below.

adm_cp with rho = $\text{adm_ap}^{\text{rho}}$.

[0177] Alternating Circle Method: L1:

Method: This is the situation where there are one or more free agents and at most one call in the queue. At any point in time, map the percentiles of the free agents to the circle $[-1,1)$ by alternating the signum. In addition, we randomize the starting signum. First we tried only varying the signum on the freeagents. That is:

$(\text{acl_ap}[\text{freeagents}])[i] = \text{rbinom}(1,1,0.5) * (\text{base_ap}[\text{freeagents}])[i] * (-1)^i$.

The caller cp is not altered. Distances are computed based on identifying -1 and +1.

Unfortunately, numerical experimentation showed a higher order masking effect due the reason for which is not clear. For this reason we tried a static variation of the signum. That is:

$\text{acl_ap}[i] = \text{rbinom}(1,1,0.5) * \text{acl_ap}[i] * (-1)^i$

Again, the caller cp is not altered and distances are computed based on identifying -1 and +1. This resulted in perfect homogenization of utilization.

Theory: The basis for this method is the observation that there would be no edge effect if there were no edge. In this case we change the topology so that the agent and caller percentiles are evenly distributed along a circle of length 2 (take the interval [-1,+1) and identify the endpoints -1 and +1). By randomizing the starting signum for the free agents, we can leave the caller on the [0,1) portion of the circle. The shortcoming of this method is that sometimes the caller cp is far from the endpoints 0 and 1, but the “best” match is on the negative side of the circle and the second “best” match is not very good.

Kappa: For acl, kappa is applied directly with the formula below.

$acl_ap \text{ with } \kappa = |acl_ap|^\kappa * \text{signum}(acl_ap).$

[0178] Alternating Circle Method: L2:

Method: This is the situation where there are one or more callers in the queue and at most one free agent. At any point in time, map the percentiles of the callers in the queue to the circle [-1,+1) by alternating the signum. In addition, we randomize the starting signum. That is:

$(acl_cp[callqueue])[i] = \text{rbinom}(1,1,0.5) * (base_cap[callqueue])[i] * (-1)^i$

Distances are computed based on identifying -1 and +1.

Theory: The basis for this method is the observation that there would be no edge effect if there were no edge. That is, if the agent and caller percentiles were evenly distributed along a circle of length 2. By randomizing the starting signum for the call queue, we can leave the agent on the [0,1) portion of the circle. The shortcoming of this method is that sometimes the agent ap is far from the endpoints 0 and 1, but the “best” match is on the negative side of the circle and the second “best” match is not very good.

Rho: For acl, rho is applied directly via the formula below.

acl_cp with rho = $|acl_cp|^\kappa * \text{signum}(acl_cp)$

[0179] Look Ahead Method: L1:

Method: This is a fairly complex method.

```
bestagent_now = bestagent
count_freeagents = sum(agents == 0)
freeagent_score = rep(0, count_freeagents)
penalty = rep(0, count_freeagents)
discount_factor = 1.0
for(j in 1:count_freeagents)
{
  penalty[j] = abs(ap_freeagents[j]-thecall)
  post_ap_freeagents = ap_freeagents[-j]
  freeagent_score[j] = (post_ap_freeagents[1]^2)/2 + (1-
post_ap_freeagents[count_freeagents-1]^2)/2
  if(count_freeagents > 2)
  {
    for(k in 1:(count_freeagents-2))
    {
      freeagent_score[j] = freeagent_score[j] + ((post_ap_freeagents[k+1]-
post_ap_freeagents[k])^2)/4
    }
  }
  bestagent2 = which(freeagents)[which.min(penalty + freeagent_score*discount_factor)]
  if(bestagent[1] != bestagent2[1]){now_ne_lah_counter = now_ne_lah_counter + 1}.
```

Theory: This method is an attempt modify the current greedy method into one that penalizes the greed to the extent that it makes the next caller's expected distance greater.

Kappa: For adm, kappa is applied directly.

lah_ap with kappa = lah_ap^κ .

[0180] Look Ahead Method: L2:

Theory: The same method as for the Adaptive Shift Methodology.

Rho: For lah, rho is applied directly via the formula below.

lah_cp with $\rho = \text{lah_ap}^\rho$.

[0181] Edge Utilization: The graphs for $\kappa = 1$, were examined. The results are summarized here and several of these graphs are included to make the visualization clear.

The first graph illustrates the situation without an edge correction.

- For $\mu = 40$, acl, and asm agent utilization graphs have the desired flat profile. The profile for ssm6 is V shaped (however, only a small portion of the variability is attributable to this shape – the range of the utilization is 90 for ssm6 and 69 / 70 for asm / acl). The agent utilization graph for adm is W shaped. For now and lah the shape is an upside down U.
- For $\mu = 20$: acl, asm, ssm6 agent utilization graphs have the desired flat profile, the agent utilization graph for adm is W shaped. For now and lah the shape is an upside down U.
- For $\mu = 10, 5, 2$: acl, asm, and ssm6 have the desired flat profile. The rest do not.
- For $\mu = 0$: acl and asm have the desired flat profile. ssm6 has a bit of a shape to it, but the difference between highest and lowest agent utilization is only slightly higher than for acl and asm.

[0182] The graphs for $\kappa = 1.2$ and 1.4 were examined. The results are summarized and several graphs are included to make the visualization clear.

- Acl accommodates κ seamlessly, with a monotonically increasing utilization
- now, asm, ssm5 and ssm6 have utilization that decreases for high ap. The decrease in utilization for highest ap agents is greatest for now and least for ssm6. The decrease for ssm6 is sufficiently small that we believe there is no reason to look for a better solution at this time.

[0183] The method selected for actual implementation depends on a variety of factors.

Including:

- Importance of minimizing agent and caller percentile differences.
- Agent utilization requirements of call center
- Service level agreements concerning caller wait times
- Technological implementation limitations

[0184] Higher dimensional analogues occur when additional agent and caller parameters are matched. Other topologies are also possible. For instance, the most straight forward

extension is for the agent and caller to each be characterized by two independent parameters, the first being the performance percentile as previously discussed and the second being a percentile based on the expected handle time for the agent and caller, respectively. The edge compensation algorithm analogues of the methods discussed above would then correspond to insertion of a margin at the edge (additionally widened near the vertices) or alteration of the topology from a square to a cylinder or torus). If categorical parameters are matched, the topology may be disconnected.

[0185] Referring to Fig. 12. embodiments of a method are disclosed for distribution compensation. Block 1200 represents an operation of obtaining, by one or more computers, agent parameter data for a set of agents.

[0186] Block 1210 represents an operation of ranking or percentiling, by the one or more computers, the agents based at least in part on the agent parameter data, to obtain an agent distribution of agent rankings or percentiles. In embodiments, agent performance comprises one or more selected from the group of sale or no sale, number of items sold per call, and revenue per call coupled with handle time.

[0187] Block 1220 represents an operation of partitioning, by the one or more computers, callers in a set of callers based on one or more criteria into a set of partitions. In embodiments, the partition for the callers is based at least in part on one or more selected from the group of demographic data, area code, zip code, NPANXX, VTN, geographic area, 800 number, and transfer number.

[0188] Block 1230 represents an operation of obtaining, by one or more computers, caller propensity data for the respective partitions.

[0189] Block 1240 represents an operation of ranking or percentiling, by the one or more computers, the callers based at least in part on data relating to or predicting a caller propensity for a desired outcome based at least in part on the caller propensity data of

partitions of the respective callers, to obtain a caller distribution of caller rankings or percentiles.

[0190] Block 1250 represents an operation of performing distribution compensation, by the one or more computers, using at least one algorithm selected from the group of: an edge compensation algorithm applied to at least one selected from the group of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles, near at least one edge of the respective distribution, to obtain edge compensated rankings or percentiles; and a topology altering algorithm applied to either or both of the distribution of the agent rankings or percentiles and the distribution of the caller rankings or percentiles to change one or more of the distributions to a different topology.

[0191] In embodiments, the performing the distribution compensation step uses the edge compensation algorithm and may provide edge compensation only to the agent rankings or percentiles. In embodiments, the performing the distribution compensation step uses the edge compensation algorithm and may provide edge compensation only to the caller rankings or percentiles. In embodiments, the performing the distribution compensation step uses the edge compensation algorithm and may provide edge compensation to both the caller rankings or percentiles and the agent rankings or percentiles.

[0192] In embodiments, the distribution compensation step uses the edge compensation algorithm and may takes the agents that are free at runtime, and rescale the respective agent rankings or percentiles for these runtime available agents to provide more space/margin at the edges of the agent distribution. In embodiments, the amount of the margin is based at least in part on a number of the agents that are free at runtime.

[0193] In embodiments, the distribution compensation step uses the edge compensation algorithm and may takes the callers in a queue or other grouping at runtime and rescale the respective caller propensity rankings or percentiles, to provide more space/margin at the edges of the distribution. In embodiments, the amount of the margin is based at least in part on a number of callers that are in the queue or in the grouping at runtime.

[0194] In embodiments, the distribution compensation step uses the edge compensation algorithm and may weight multiple of the agents free at runtime near at least one edge of the agent distribution and weight multiple of the callers near at least one edge of the caller distribution to increase utilization of the agents at the at least one edge of the agent distribution.

[0195] In embodiments, the distribution compensation step uses the edge compensation algorithm and may weight multiple of the agents free at runtime near both edges of the agent distribution or weight multiple of the callers near both edges of the caller distribution.

[0196] Block 1260 represents an operation of matching, by the one or more computers, a respective one of the agents with a respective ranking or percentile to a respective one of the callers in one of the partitions with a closest respective ranking or percentile, where at least one of the caller ranking or percentile or the agent ranking or percentile has been distribution compensated.

[0197] In embodiments, the distribution compensation step uses the topology altering algorithm, which algorithm may convert the distribution of the agent performances and the distribution of the caller to a circle.

[0198] In embodiments, the distribution compensation step uses the topology altering algorithm, which algorithm may convert the distribution of the agent performances and the distribution of the caller to remove the edges of the distribution.

[0199] In embodiments where the distribution compensation step uses the edge compensation algorithm, the Kappa for the distribution of the agents may be greater than 1.0.

[0200] In embodiments, rho applied to the callers in a queue may be 1.0. Note that rho is the analogue for Kappa as applied to calls. A rho = 1.0 is the analogue of a flat utilization of calls, e.g., the same expected L2 wait time for all call CP's. A rho greater than 1.0 is a setting in which there is a bias to answer the highest ranked or percentiled calls first.

[0201] Many of the techniques described here may be implemented in hardware or software, or a combination of the two. Preferably, the techniques are implemented in computer programs executing on programmable computers that each includes a processor, a storage medium readable by the processor (including volatile and nonvolatile memory and/or storage elements), and suitable input and output devices. Program code is applied to data entered using an input device to perform the functions described and to generate output information. The output information is applied to one or more output devices. Moreover, each program is preferably implemented in a high level procedural or object-oriented programming language to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language.

[0202] Each such computer program is preferably stored on a storage medium or device (e.g., CD-ROM, hard disk or magnetic diskette) that is readable by a general or special purpose programmable computer for configuring and operating the computer when the storage medium or device is read by the computer to perform the procedures described. The system also may be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner.

[0203] Figure 10 illustrates a typical computing system 1000 that may be employed to implement processing functionality in embodiments of the invention. Computing systems of this type may be used in clients and servers, for example. Those skilled in the relevant art will also recognize how to implement the invention using other computer systems or architectures. Computing system 1000 may represent, for example, a desktop, laptop or notebook computer, hand-held computing device (PDA, cell phone, palmtop, etc.), mainframe, server, client, or any other type of special or general purpose computing device as may be desirable or appropriate for a given application or environment. Computing system 1000 can include one or more processors, such as a processor 1004. Processor 1004 can be implemented using a general or special purpose processing engine such as, for example, a microprocessor, microcontroller or other control logic. In this example, processor 1004 is connected to a bus 1002 or other communication medium.

[0204] Computing system 1000 can also include a main memory 1008, such as random access memory (RAM) or other dynamic memory, for storing information and instructions to be executed by processor 1004. Main memory 1008 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 1004. Computing system 1000 may likewise include a read only memory (“ROM”) or other static storage device coupled to bus 1002 for storing static information and instructions for processor 1004.

[0205] The computing system 1000 may also include information storage system 1010, which may include, for example, a media drive 1012 and a removable storage interface 1020. The media drive 1012 may include a drive or other mechanism to support fixed or removable storage media, such as a hard disk drive, a floppy disk drive, a magnetic tape drive, an optical disk drive, a CD or DVD drive (R or RW), or other removable or fixed media drive. Storage media 1018 may include, for example, a hard disk, floppy disk, magnetic tape, optical disk, CD or DVD, or other fixed or removable medium that is read by and written to by media drive 1012. As these examples illustrate, the storage media 1018 may include a computer-readable storage medium having stored therein particular computer software or data.

[0206] In alternative embodiments, information storage system 1010 may include other similar components for allowing computer programs or other instructions or data to be loaded into computing system 1000. Such components may include, for example, a removable storage unit 1022 and an interface 1020, such as a program cartridge and cartridge interface, a removable memory (for example, a flash memory or other removable memory module) and memory slot, and other removable storage units 1022 and interfaces 1020 that allow software and data to be transferred from the removable storage unit 1018 to computing system 1000.

[0207] Computing system 1000 can also include a communications interface 1024. Communications interface 1024 can be used to allow software and data to be transferred between computing system 1000 and external devices. Examples of communications interface 1024 can include a modem, a network interface (such as an Ethernet or other NIC card), a communications port (such as for example, a USB port), a PCMCIA slot and card, etc. Software and data transferred via communications interface 1024 are in the form of signals which can be electronic, electromagnetic, optical or other signals capable of being received by communications interface 1024. These signals are provided to communications

interface 1024 via a channel 1028. This channel 1028 may carry signals and may be implemented using a wireless medium, wire or cable, fiber optics, or other communications medium. Some examples of a channel include a phone line, a cellular phone link, an RF link, a network interface, a local or wide area network, and other communications channels.

[0208] In this document, the terms “computer program product,” “computer-readable medium” and the like may be used generally to refer to physical, tangible media such as, for example, memory 1008, storage media 1018, or storage unit 1022. These and other forms of computer-readable media may be involved in storing one or more instructions for use by processor 1004, to cause the processor to perform specified operations. Such instructions, generally referred to as “computer program code” (which may be grouped in the form of computer programs or other groupings), when executed, enable the computing system 1000 to perform features or functions of embodiments of the present invention. Note that the code may directly cause the processor to perform specified operations, be compiled to do so, and/or be combined with other software, hardware, and/or firmware elements (e.g., libraries for performing standard functions) to do so.

[0209] In an embodiment where the elements are implemented using software, the software may be stored in a computer-readable medium and loaded into computing system 1000 using, for example, removable storage media 1018, drive 1012 or communications interface 1024. The control logic (in this example, software instructions or computer program code), when executed by the processor 1004, causes the processor 1004 to perform the functions of the invention as described herein.

[0210] It will be appreciated that, for clarity purposes, the above description has described embodiments of the invention with reference to different functional units and processors. However, it will be apparent that any suitable distribution of functionality between different functional units, processors or domains may be used without detracting from the invention. For example, functionality illustrated to be performed by separate processors or controllers may be performed by the same processor or controller. Hence, references to specific functional units are only to be seen as references to suitable means for providing the described functionality, rather than indicative of a strict logical or physical structure or organization.

[0210] The above-described embodiments of the present invention are merely meant to be illustrative and not limiting. Various changes and modifications may be made without departing from the invention in its broader aspects.

CLAIMS

1. A method for handling contacts and agents in a contact center system comprising:
 - determining, by at least one computer processor communicatively coupled to and configured to operate in the contact center system, a rank for each of a plurality of available agents using a model based on outcomes of previous agent-contact pairings;
 - adjusting, by the at least one computer processor and before an arrival of a next available contact, a first agent-selection percentile for a higher-ranked agent of the plurality of available agents, which controls the probability that the higher-ranked agent will be selected for connection to the next available contact in the contact center system, relative to a second agent-selection percentile for a lower-ranked agent of the plurality of available agents, which controls the probability that the lower-ranked agent will be selected for connection to the next available contact in the contact center system; and
 - selecting, based on the first agent-selection percentile and the second agent-selection percentile, by the at least one computer processor and after the arrival of the next available contact, the lower-ranked agent or the higher-ranked agent for connection to the next available contact in the contact center system.
2. The method of claim 1, wherein an expected utilization of each of the plurality of available agents is substantially uniform before adjusting the first agent-selection percentile and the second agent-selection percentile.
3. The method of claim 1, wherein the adjusting further comprises:
 - adjusting, by the at least one computer processor, an expected utilization of each of the plurality of available agents by rescaling the rank of each of the plurality of available agents.
4. The method of claim 1, wherein adjusting the first agent-selection percentile relative to the second agent-selection percentile further comprises:
 - adjusting, by the at least one computer processor, the rank of each of the plurality of available agents using an exponential function of the substantially uniformly distributed percentile ranking.

5. The method of claim 1, further comprising decreasing, by the at least one computer processor, a first expected wait time of the higher-ranked agent relative to a second expected wait time of the lower-ranked agent.

6. The method of claim 1, further comprising:

determining, by the at least one computer processor, an agent utilization requirement of the contact center system; and

decreasing, by the at least one computer processor, a uniformity of an expected utilization for each of the plurality of available agents while complying with the agent utilization requirement of the contact center system.

7. A system for handling contacts and agents in a contact center system comprising:

at least one computer processor communicatively coupled to and configured to operate in the contact center system, wherein the computer processor is further configured to:

determine a rank for each of a plurality of available agents using a model based on outcomes of previous agent-contact pairings;

adjust, before an arrival of a next available contact, a first agent-selection percentile for a higher-ranked agent of the plurality of available agents, which controls the probability that the higher-ranked agent will be selected for connection to the next available contact in the contact center system, relative to a second agent-selection percentile for a lower-ranked agent of the plurality of available agents, which controls the probability that the lower-ranked agent will be selected for connection to the next available contact in the contact center system; and

select, based on the first agent-selection percentile and the second agent-selection percentile, and after the arrival of the next available contact, the lower-ranked agent or the higher-ranked agent for connection to the next available contact in the contact center system.

8. The system of claim 7, wherein an expected utilization of each of the plurality of available agents is substantially uniform before adjusting the first agent-selection percentile and the second agent-selection percentile.

9. The system of claim 7, wherein the at least one computer processor is further configured to adjust an expected utilization of each of the plurality of available agents by rescaling the rank of each of the plurality of available agents.

10. The system of claim 7, wherein adjusting the first agent-selection percentile relative to the second agent-selection percentile comprises adjusting the rank of each of the plurality of available agents using an exponential function of the substantially uniformly distributed percentile ranking.

11. The system of claim 7, wherein the at least one computer processor is further configured to decrease a first expected wait time of the higher-ranked agent relative to a second expected wait time of the lower-ranked agent.

12. The system of claim 7, wherein the at least one computer processor is further configured to:

determine an agent utilization requirement of the contact center system; and

decrease a uniformity of an expected utilization for each of the plurality of available agents while complying with the agent utilization requirement of the contact center system.

13. An article of manufacture for handling contacts and agents in a contact center system comprising:

a non-transitory computer processor readable medium; and

instructions stored on the medium;

wherein the instructions are configured to be readable from the medium by at least one computer processor communicatively coupled to and configured to operate in the contact center system and thereby cause the at least one computer processor to operate so as to:

determine a rank for each of a plurality of available agents using a model based on outcomes of previous agent-contact pairings;

adjust, before an arrival of a next available contact, a first agent-selection percentile for a higher-ranked agent of the plurality of available agents, which controls the probability that the higher-ranked agent will be selected for connection to the next available contact in the contact center system, relative to a second agent-selection percentile for a lower-ranked agent of the plurality of available agents, which controls the probability that the lower-ranked agent will be selected for connection to the next available contact in the contact center system; and

select, based on the first agent-selection percentile and the second agent-selection percentile, and after the arrival of the next available contact, the lower-ranked agent or the higher-ranked agent for connection to the next available contact in the contact center system.

14. The article of manufacture of claim 13, wherein an expected utilization of each of the plurality of available agents is substantially uniform before adjusting the first agent-selection percentile and the second agent-selection percentile.

15. The article of manufacture of claim 13, wherein the at least one computer processor is further caused to operate so as to adjust an expected utilization of each of the plurality of available agents by rescaling the rank of each of the plurality of available agents.

16. The article of manufacture of claim 13, wherein adjusting the first agent-selection percentile relative to the second agent-selection percentile comprises adjusting the rank of each of the plurality of available agents using an exponential function of the substantially uniformly distributed percentile ranking.

17. The article of manufacture of claim 13, wherein the at least one computer processor is further caused to operate so as to decrease a first expected wait time of the higher-ranked agent relative to a second expected wait time of the lower-ranked agent.

18. The article of manufacture of claim 13, wherein the at least one computer processor is further caused to operate so as to:

- determine an agent utilization requirement of the contact center system; and
- decrease a uniformity of an expected utilization for each of the plurality of available agents while complying with the agent utilization requirement of the contact center system.

19. A method for handling contacts and agents in a contact center system comprising:

- determining, by at least one computer processor communicatively coupled to and configured to operate in the contact center system, a rank for each of a plurality of contacts in queue based on a contact type of each of the plurality of contacts in queue using a model based on outcomes of previous agent-contact pairings;

- adjusting, by the at least one computer processor and before an arrival of a next available agent, a first contact-selection percentile for a higher-ranked contact of the plurality of contacts in queue, which controls the probability that the higher-ranked contact will be selected for connection to the next available agent in the contact center system, relative to a second contact-selection percentile for a lower-ranked contact of the plurality of contacts in queue, which controls the probability that the lower-ranked contact will be selected for connection to the next available agent in the contact center system; and

- selecting, based on the first contact-selection percentile and the second contact-selection percentile, by the at least one computer processor and after the arrival of the next available agent, the lower-ranked contact or the higher-ranked contact for connection to the next available agent in the contact center system.

20. The method of claim 19, wherein an expected utilization of each of the plurality of contacts in queue is substantially uniform before adjusting the first contact-selection percentile and the second contact-selection percentile.

21. The method of claim 19, wherein the adjusting further comprises: adjusting, by the at least one computer processor, an expected utilization of each of the plurality of contacts in queue by rescaling the rank of each of the plurality of contacts in queue.

22. The method of claim 19, wherein the adjusting further comprises:

adjusting, by the at least one computer processor, the rank of each of the plurality of contacts in queue using an exponential function of the substantially uniformly distributed percentile ranking.

23. The method of claim 19, further comprising decreasing, by the at least one computer processor, a first expected wait time of the higher-ranked contact relative to a second expected wait time of the lower-ranked contact.

24. The method of claim 19, further comprising:

determining, by the at least one computer processor, a contact wait time service level requirement of the contact center system; and

decreasing, by the at least one computer processor, a uniformity of an expected utilization of each of the plurality of contacts in queue while complying with the contact wait time service level requirement of the contact center system.

25. A system for handling contacts and agents in a contact center system comprising:

at least one computer processor communicatively coupled to and configured to operate in the contact center system, wherein the computer processor is further configured to:

determine a rank for each of a plurality of contacts in queue based on a contact type of each of the plurality of contacts in queue using a model based on outcomes of previous agent-contact pairings;

adjust, before an arrival of a next available agent, a first contact-selection percentile for a higher-ranked contact of the plurality of contacts in queue, which controls the probability that the higher-ranked contact will be selected for connection to the next available agent in the contact center system, relative to a second contact-selection percentile for a lower-ranked contact of the plurality of contacts in queue, which controls the probability that the lower-ranked contact will be selected for connection to the next available agent in the contact center system; and

select, based on the first contact-selection percentile and the second contact-selection percentile, and after the arrival of the next available agent, the lower-ranked contact or the higher-ranked contact for connection to the next available agent in the contact center system.

26. The system of claim 25, wherein an expected utilization of each of the plurality of contacts in queue is substantially uniform before adjusting the first contact-selection percentile and the second contact-selection percentile.

27. The system of claim 25, wherein the at least one computer processor is further configured to adjust an expected utilization of each of the plurality of contacts in queue by rescaling the rank of each of the plurality of contacts in queue.

28. The system of claim 25, wherein adjusting the first contact-selection percentile relative to the second contact-selection percentile comprises adjusting the rank of each of the plurality of contacts in queue using an exponential function of the substantially uniformly distributed percentile ranking.

29. The system of claim 25, wherein the at least one computer processor is further configured to decrease a first expected wait time of the higher-ranked contact relative to a second expected wait time of the lower-ranked contact.

30. The system of claim 25, wherein the at least one computer processor is further configured to:

- determine a contact wait time service level requirement of the contact center system; and

- decrease a uniformity of an expected utilization of each of the plurality of contacts in queue while complying with the contact wait time service level requirement of the contact center system.

31. An article of manufacture for handling contacts and agents in a contact center system comprising:

a non-transitory computer processor readable medium; and
instructions stored on the medium;

wherein the instructions are configured to be readable from the medium by at least one computer processor communicatively coupled to and configured to operate in the contact center system and thereby cause the at least one computer processor to operate so as to:

determine a rank for each of a plurality of contacts in queue based on a contact type of each of the plurality of contacts in queue using a model based on outcomes of previous agent-contact pairings;

adjust, before an arrival of a next available agent, a first contact-selection percentile for a higher-ranked contact of the plurality of contacts in queue, which controls the probability that the higher-ranked contact will be selected for connection to the next available agent in the contact center system, relative to a second contact-selection percentile for a lower-ranked contact of the plurality of contacts in queue, which controls the probability that the lower-ranked contact will be selected for connection to the next available agent in the contact center system; and

select, based on the first contact-selection percentile and the second contact-selection percentile, and after the arrival of the next available agent, the lower-ranked or the higher-ranked contact for connection to the next available agent in the contact center system.

32. The article of manufacture of claim 31, wherein an expected utilization of each of the plurality of contacts in queue is substantially uniform before adjusting the first contact-selection percentile and the second contact-selection percentile.

33. The article of manufacture of claim 31, wherein the at least one computer processor is further caused to operate so as to adjust an expected utilization of each of the plurality of contacts in queue by rescaling the rank of each of the plurality of contacts in queue.

34. The article of manufacture of claim 31, wherein adjusting the first contact-selection percentile relative to the second contact-selection percentile comprises adjusting the rank

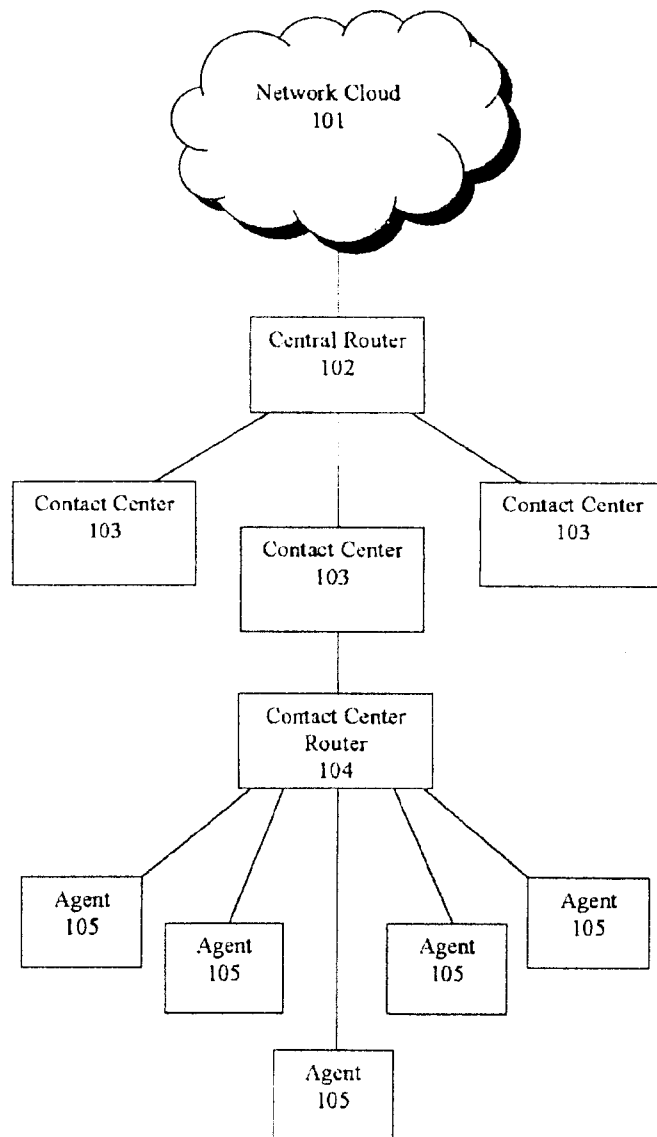
of each of the plurality of contacts in queue using an exponential function of the substantially uniformly distributed percentile ranking.

35. The article of manufacture of claim 31, wherein the at least one computer processor is further caused to operate so as to decrease a first expected wait time of the higher-ranked contact relative to a second expected wait time of the lower-ranked contact.

36. The article of manufacture of claim 31, wherein the at least one computer processor is further caused to operate so as to:

determine a contact wait time service level requirement of the contact center system; and

decrease a uniformity of an expected utilization of each of the plurality of contacts in queue while complying with the contact wait time service level requirement of the contact center system.

**FIG. 1**

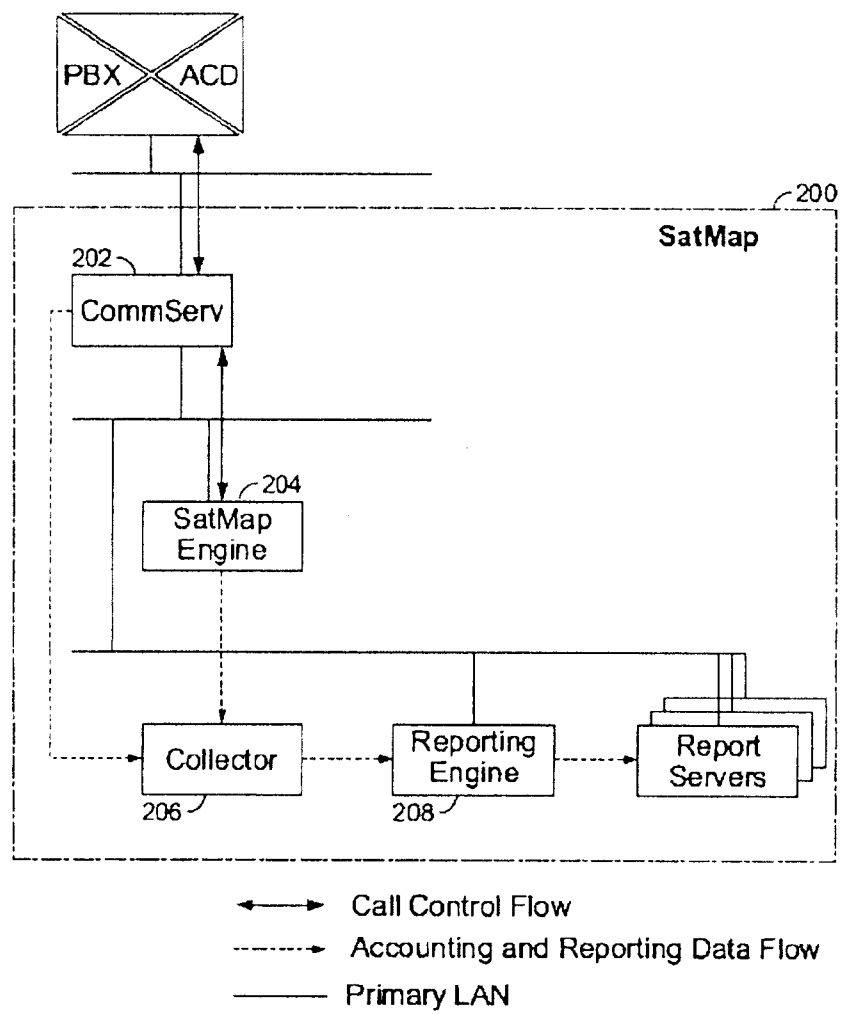


FIG. 2

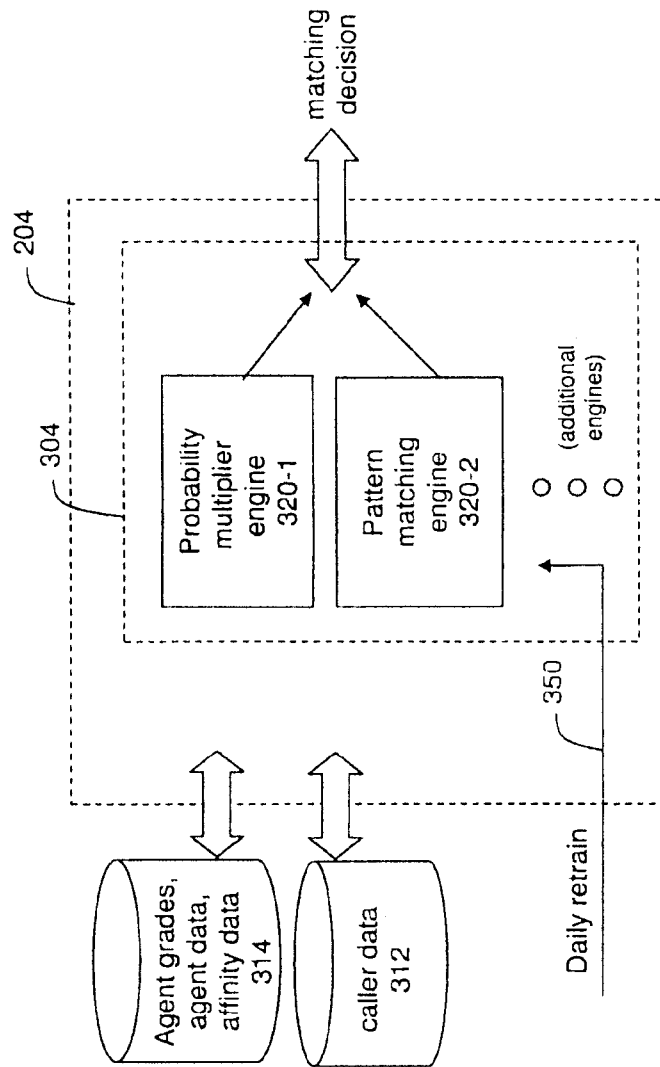


FIG. 3

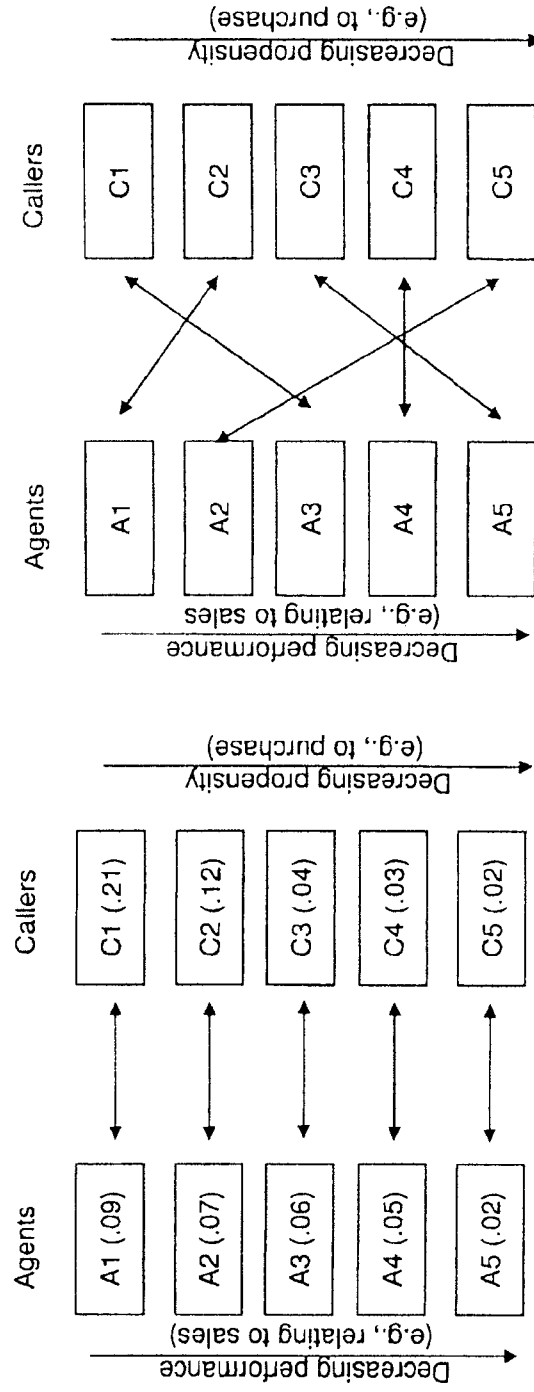
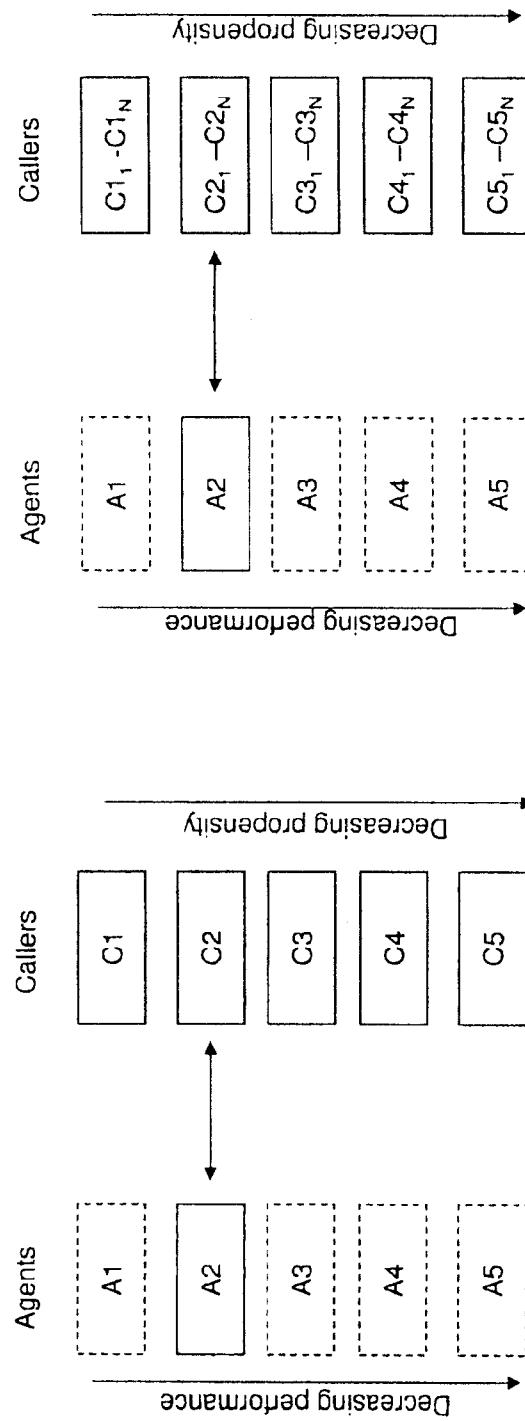


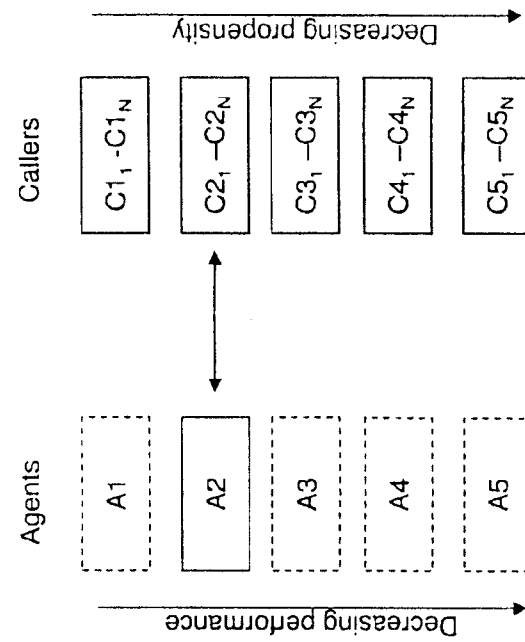
FIG. 4A

FIG. 4B



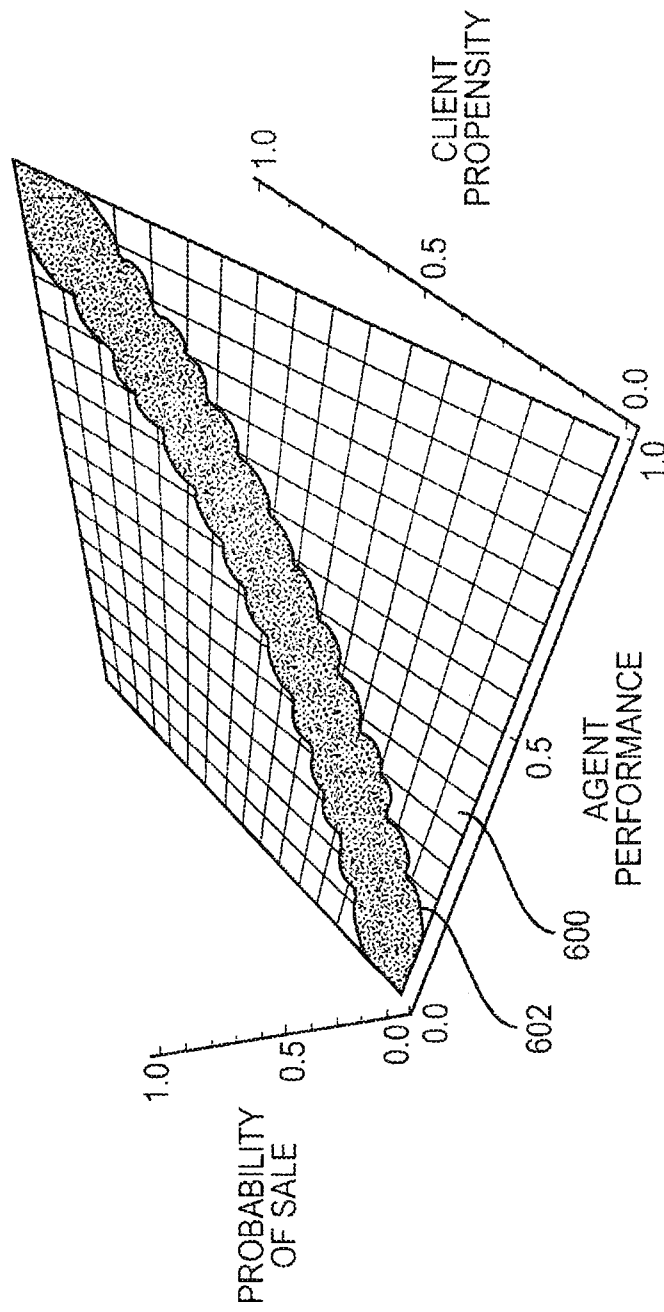
As agent A2 becomes free,
matched with caller C2 on hold
having same or closest rank

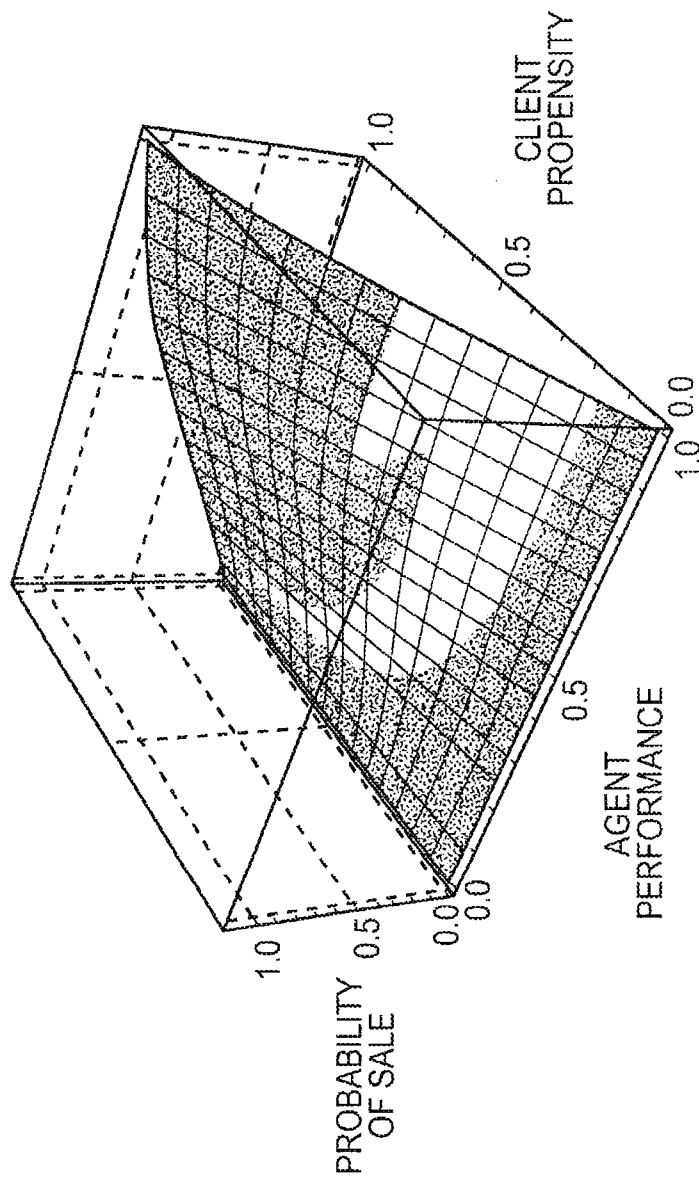
FIG. 5A



As agent A2 becomes free,
matched with caller from sub-group C2₁-C2_N,
e.g., based on queue order within
predetermined range of propensity

FIG. 5B

**FIG. 6A**

**FIG. 6B**

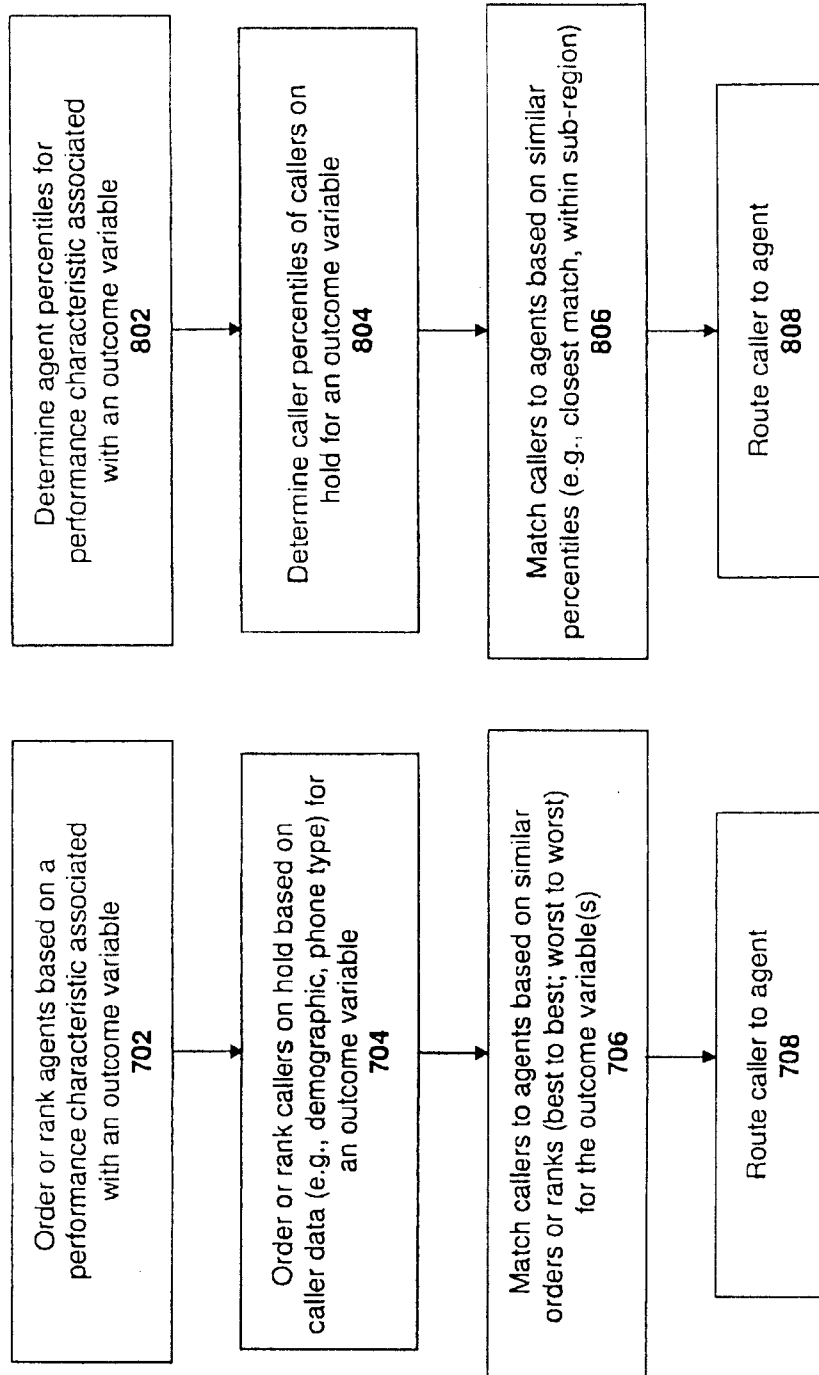
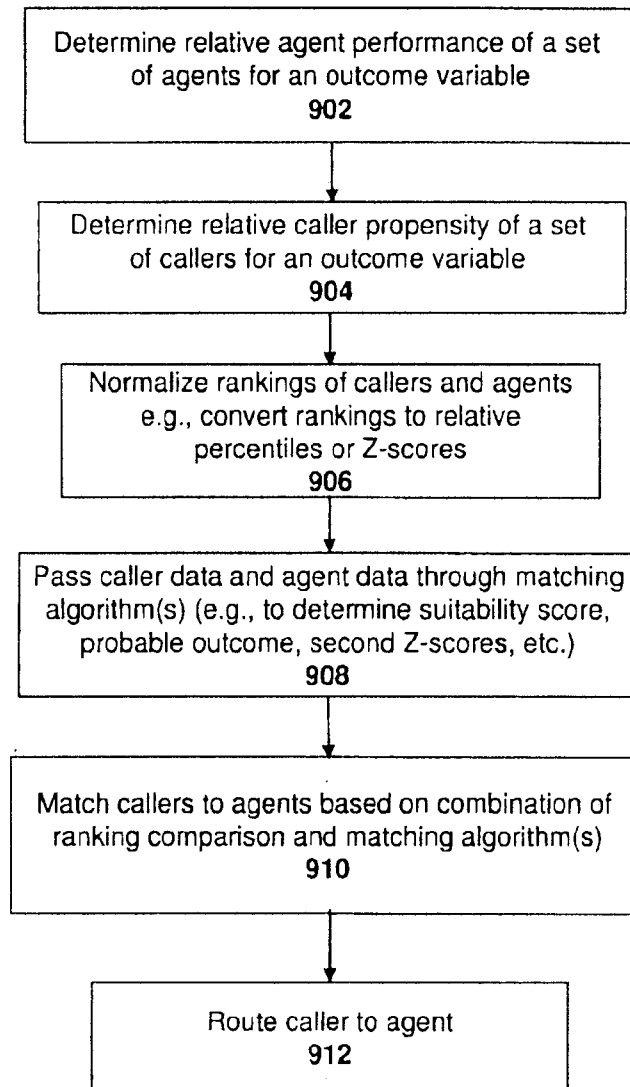


FIG. 7

FIG. 8

**FIG. 9**

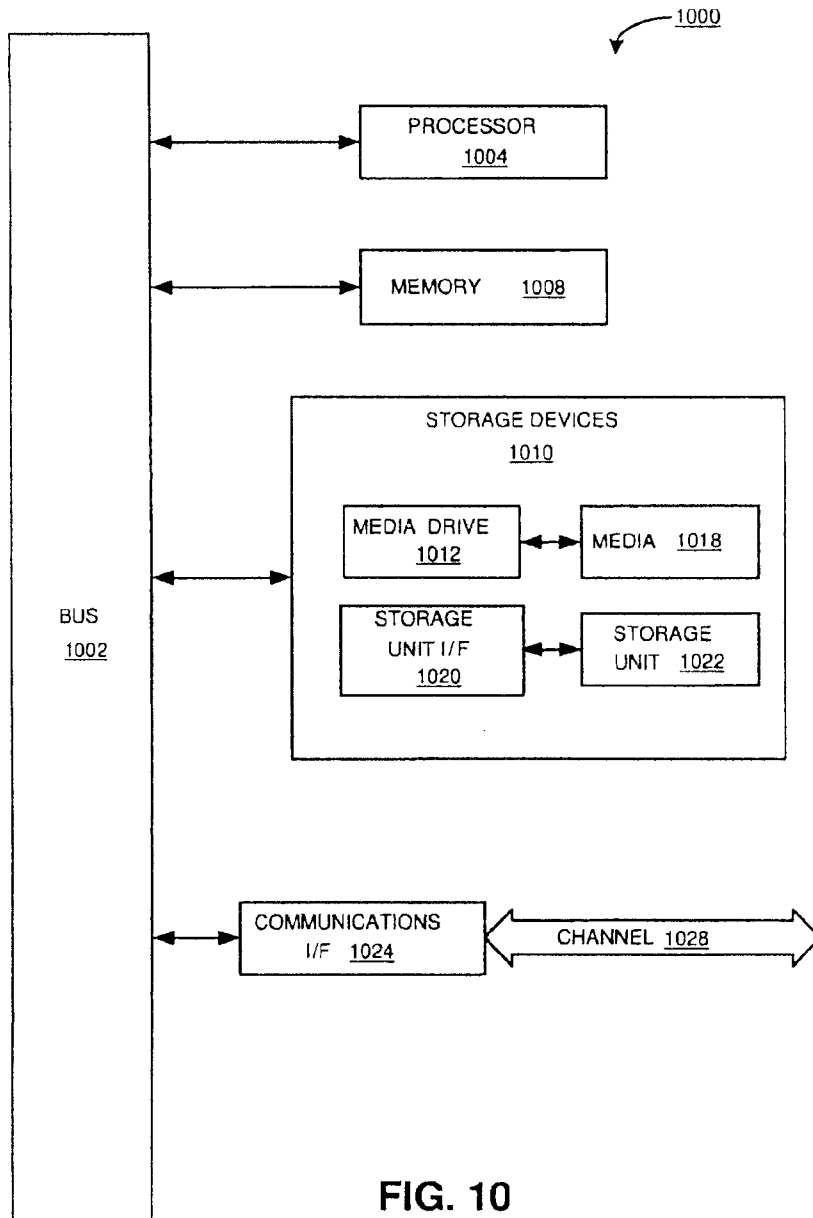
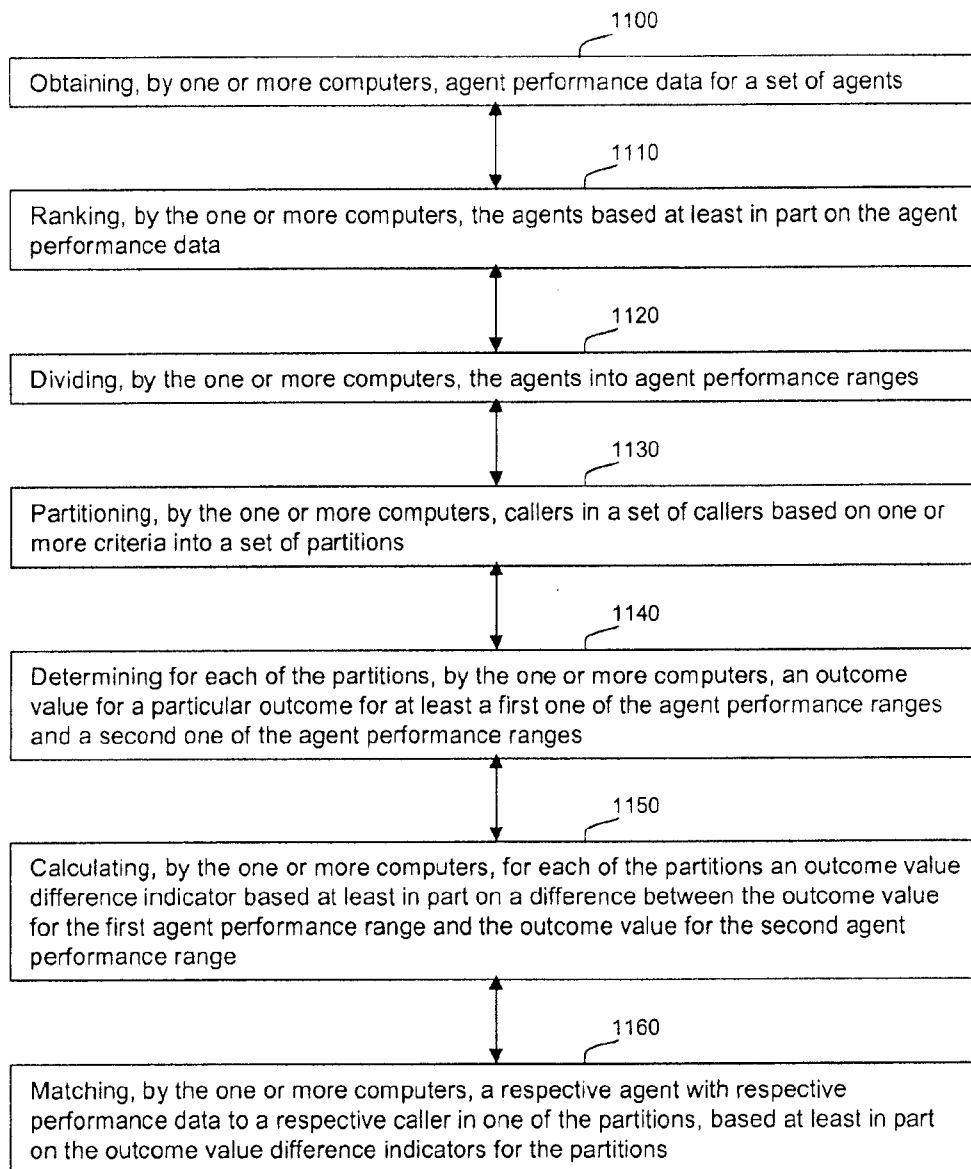


FIG. 10

Fig. 11

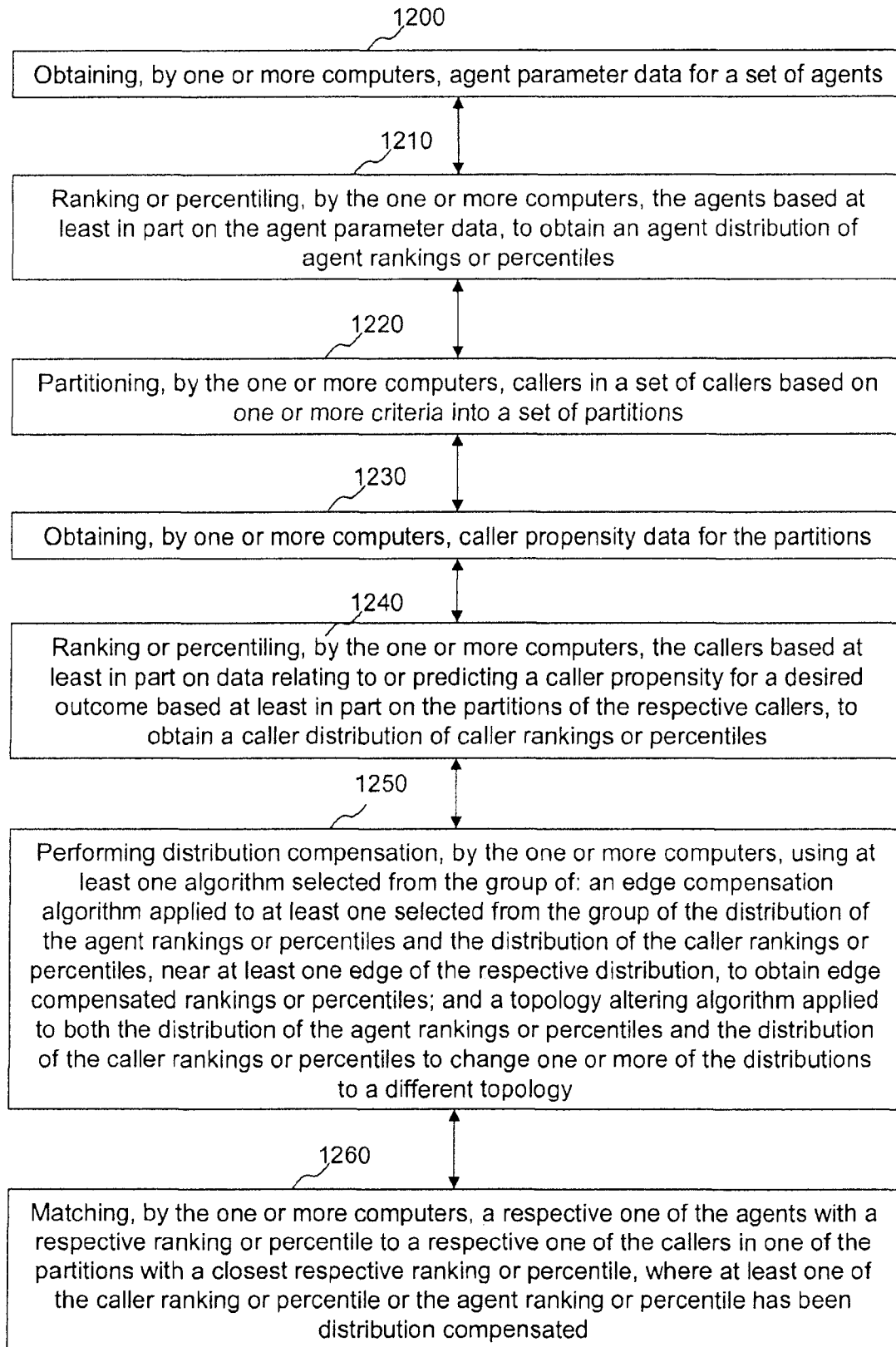


FIG. 12

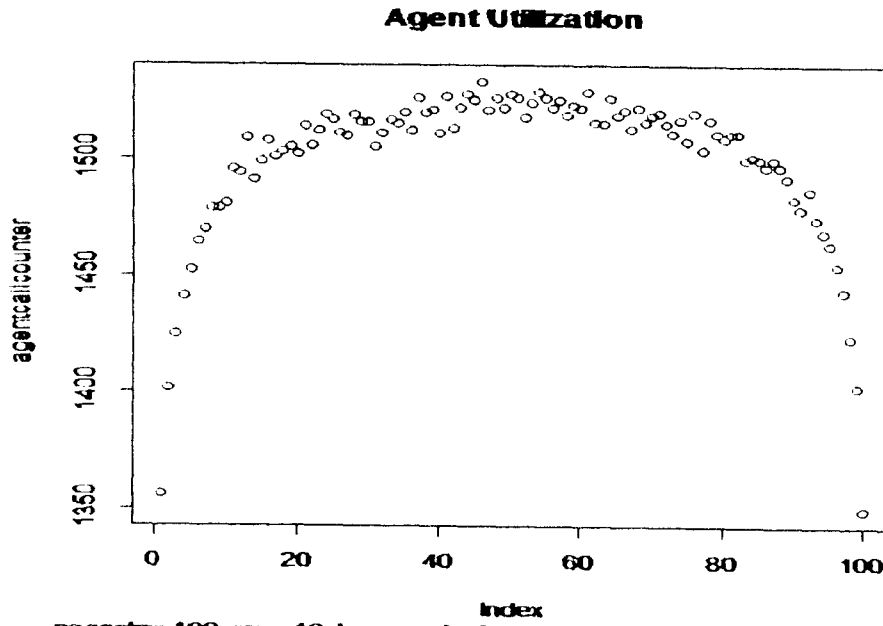


FIG. 13

• `nagents= 100 ,mu= 10 ,kappa= 1 ,ckappa= 1 ,method= now ,callbasen= 100`

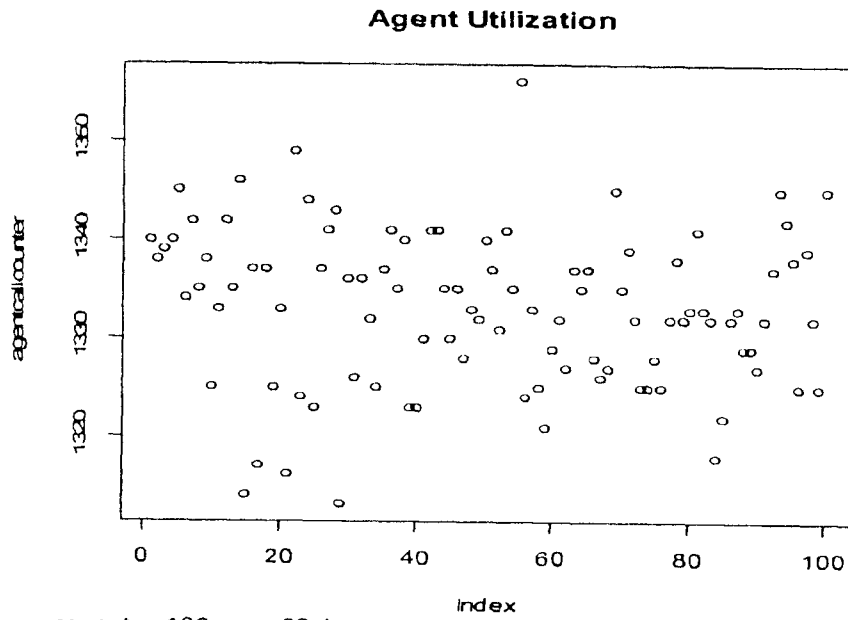
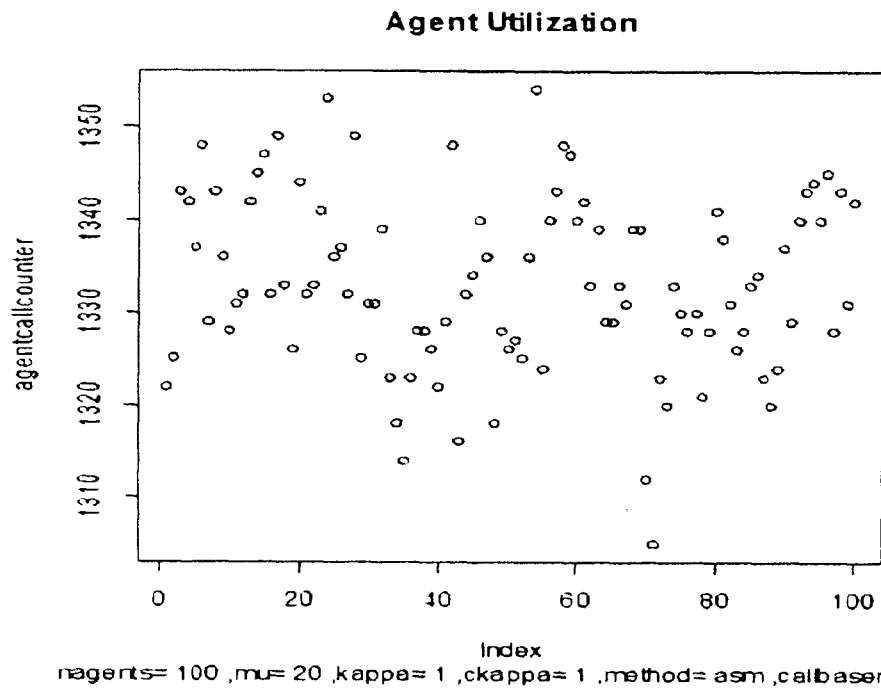


FIG. 14

`nagents= 100 ,mu= 20 ,kappa= 1 ,ckappa= 1 ,method= acl ,callbasen= 100`

**FIG. 15**

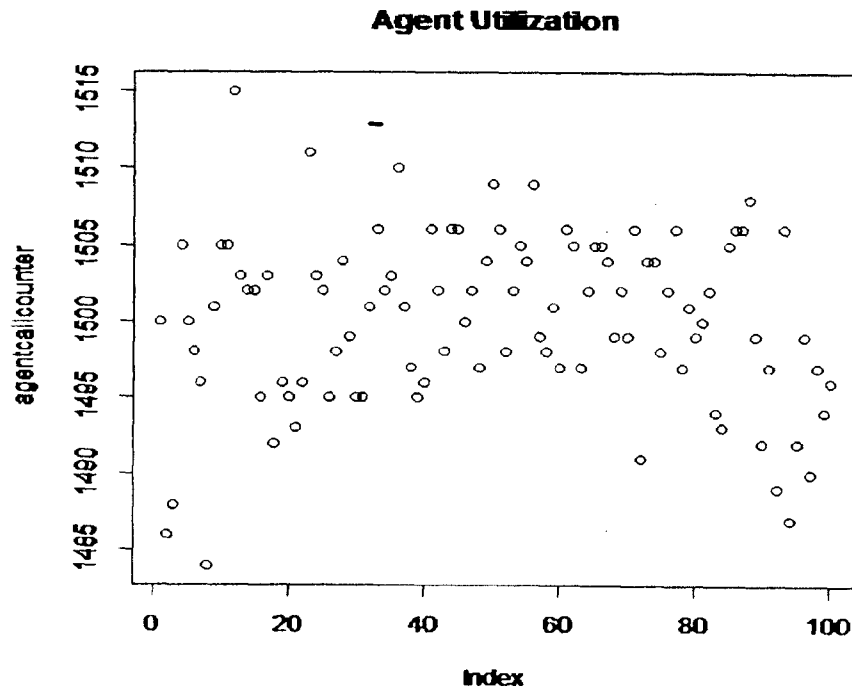


FIG. 16

nagents= 100 ,mu= 10 ,kappa= 1 ,ckappa= 1 ,method= ssm5 ,calbasen= 10

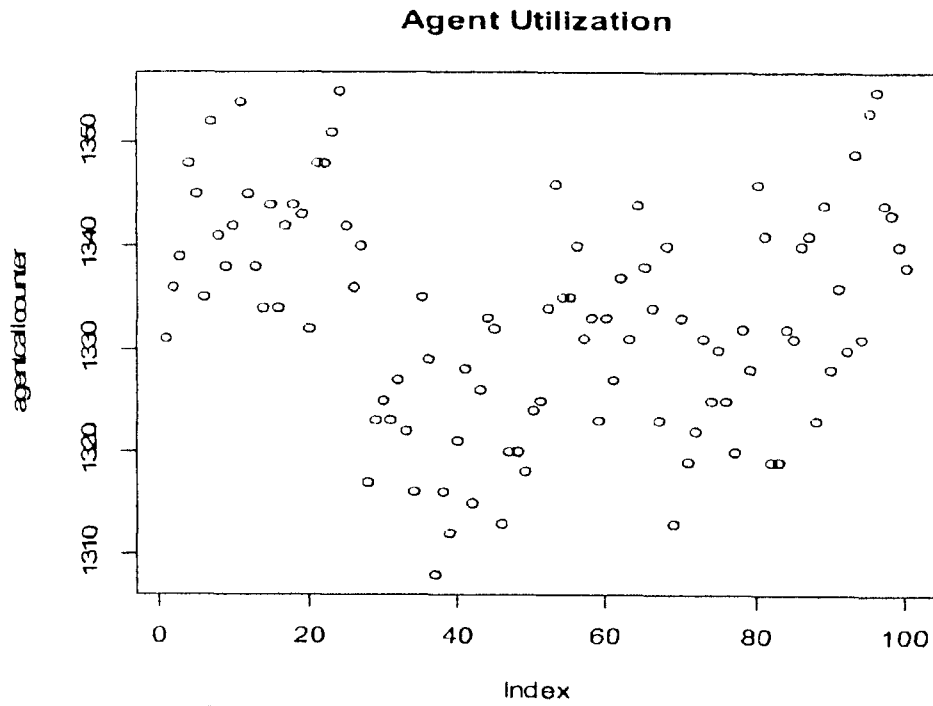


FIG. 17

nagents= 100 ,mu= 20 ,kappa= 1 ,ckappa= 1 ,method= ssm6 ,callbasen= 10

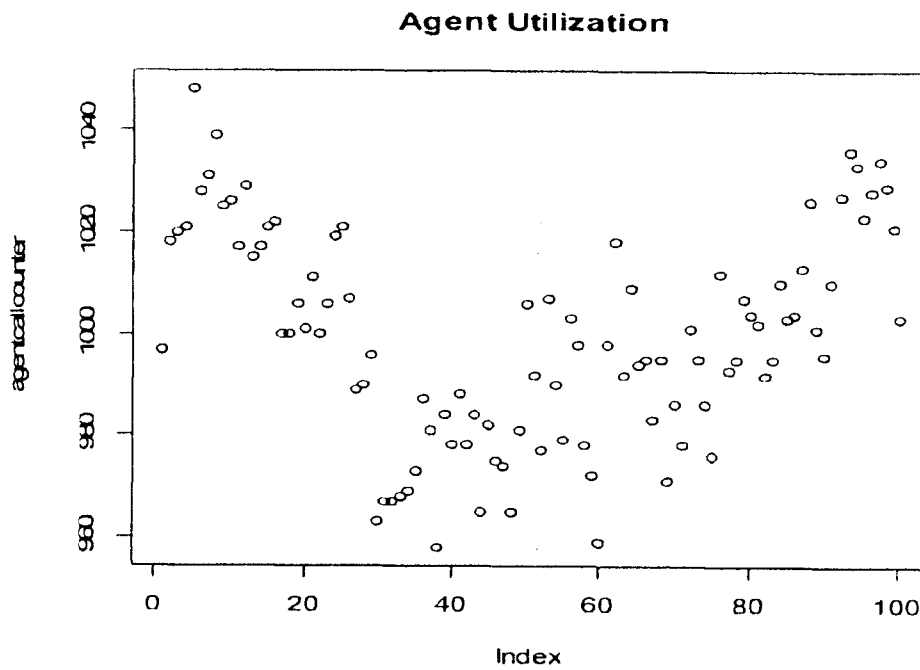
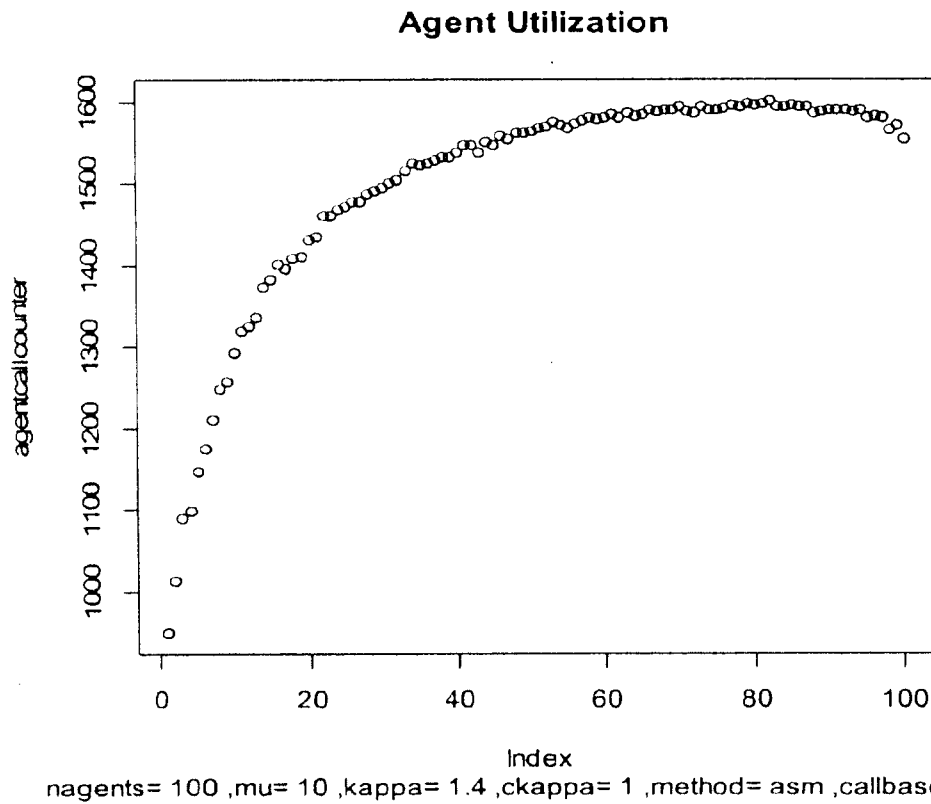


FIG. 18

nagents= 100 ,mu= 40 ,kappa= 1 ,ckappa= 1 ,method= ssm6 ,callbasen= 10

**FIG. 19**

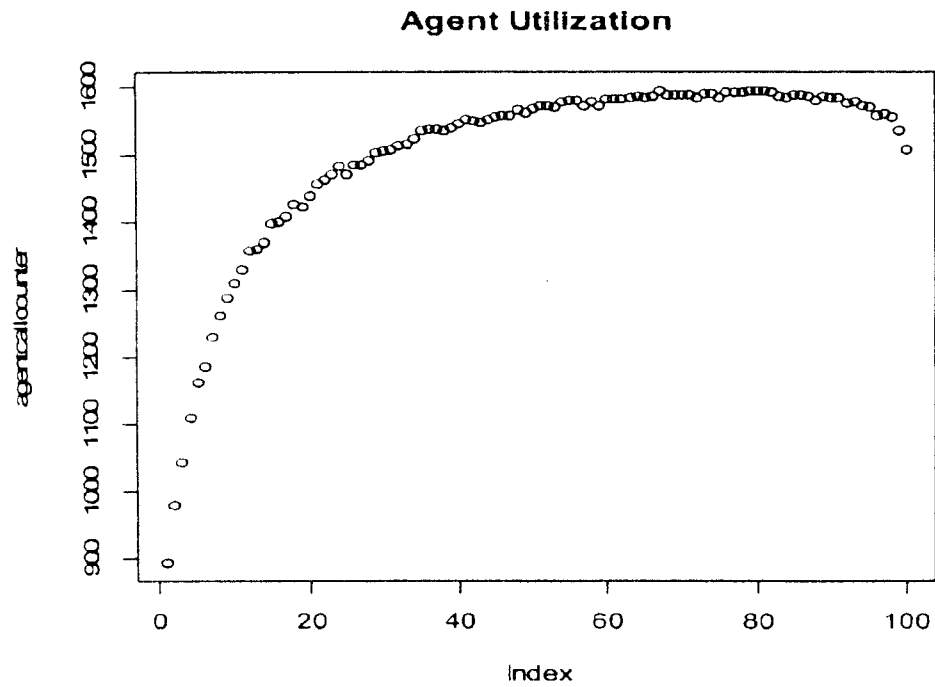


FIG. 20

nagents= 100 ,mu= 10 ,kappa= 1.4 ,ckappa= 1 ,method= now ,callbasen= 10

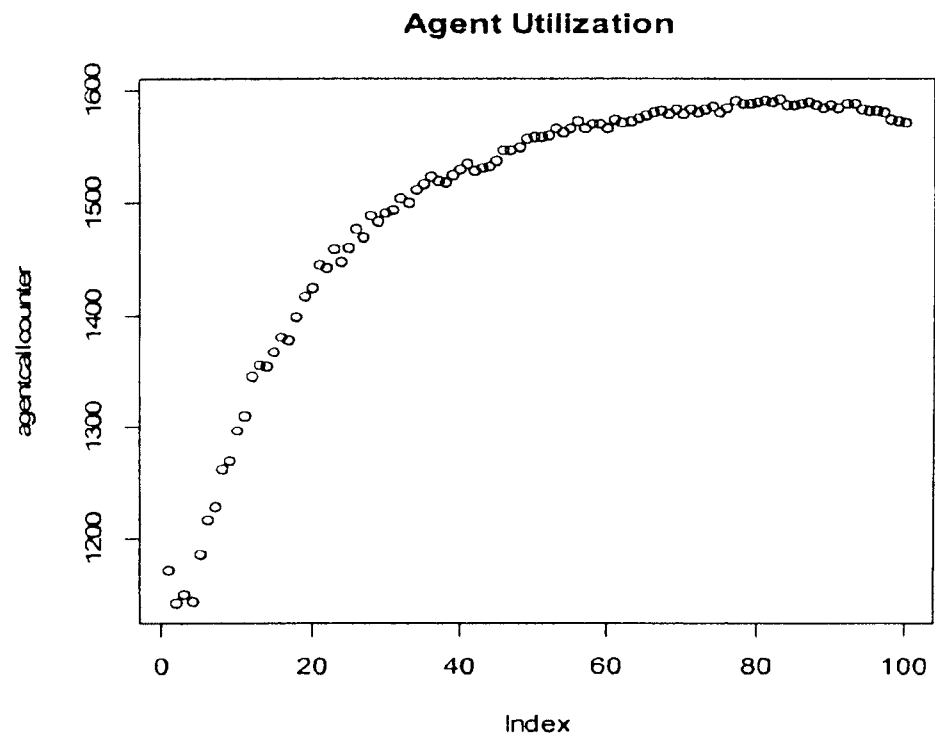


FIG. 21

nagents= 100 ,mu= 10 ,kappa= 1.4 ,ckappa= 1 ,method= ssm5 ,callbasen= 1

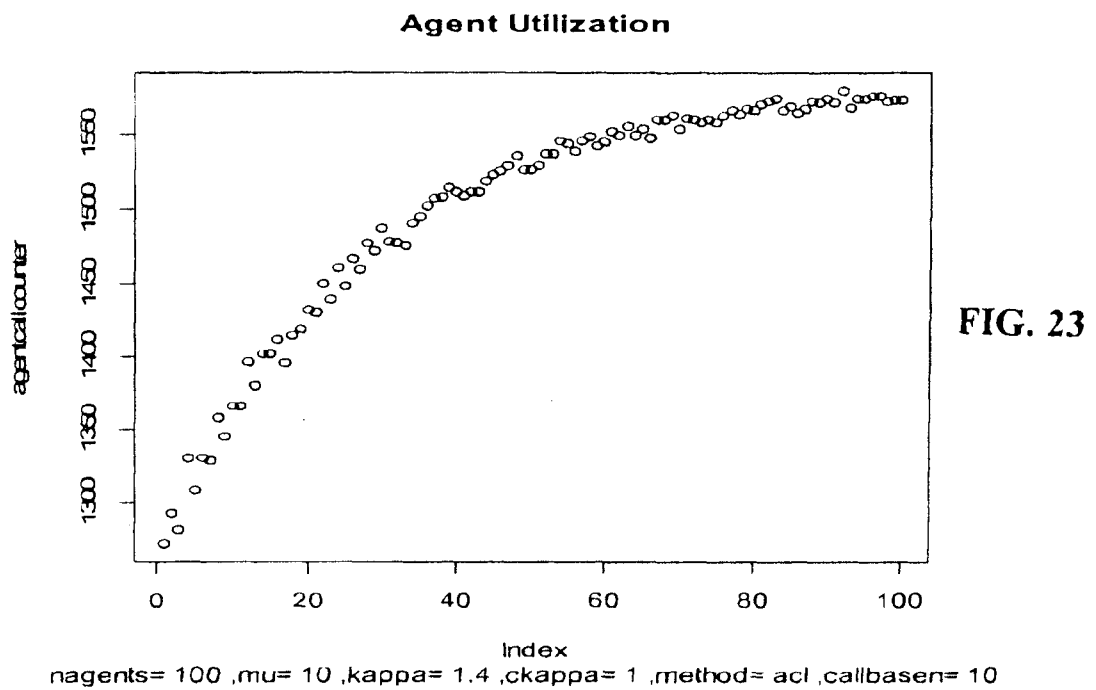
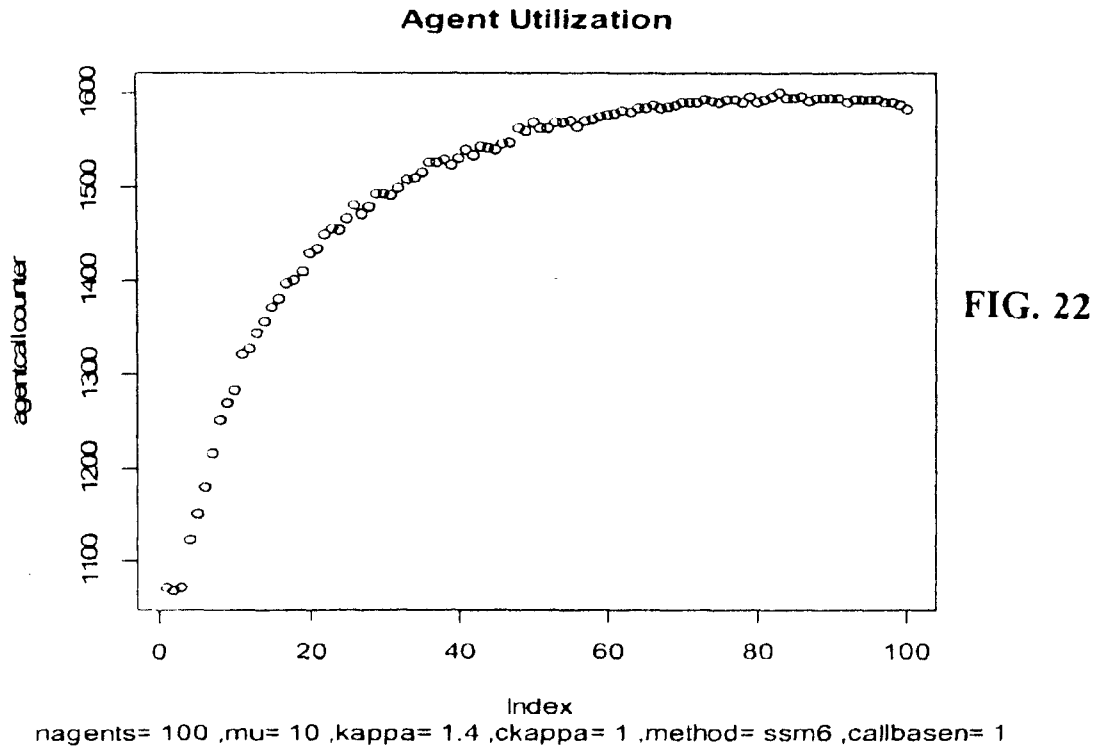


FIG. 24

Revenues Naturally Vary

Mean Revenue per Call per Day vs. Date

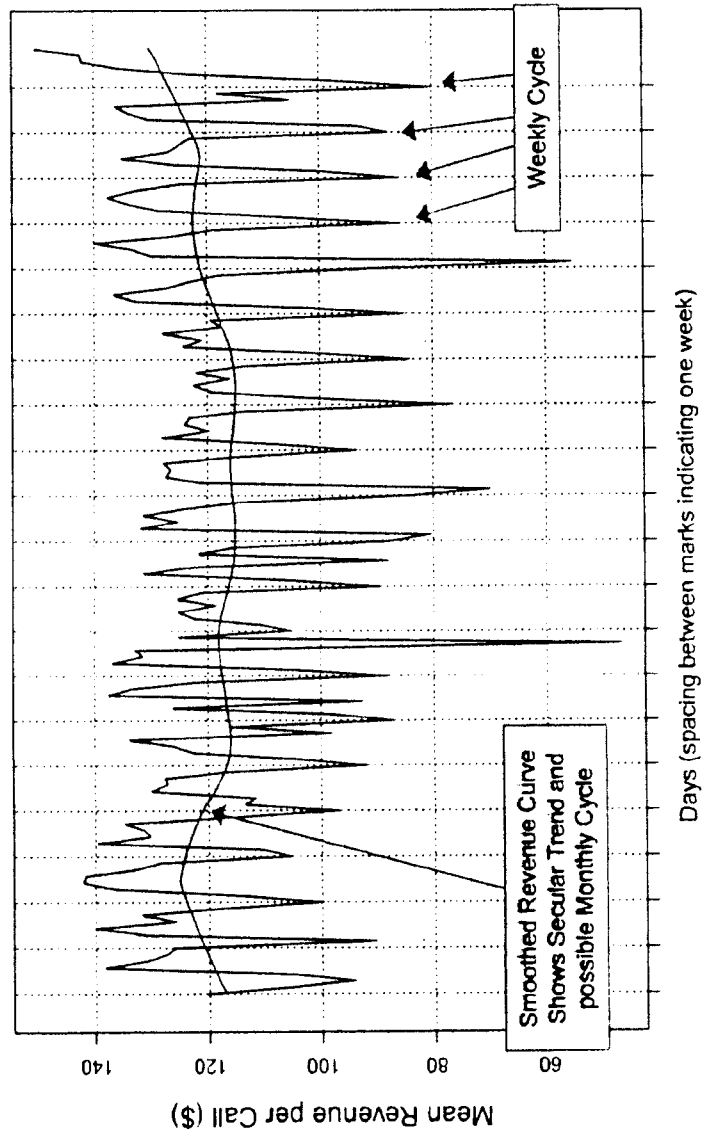


FIG. 25

Day of Week Cycle – Epoch Averaged

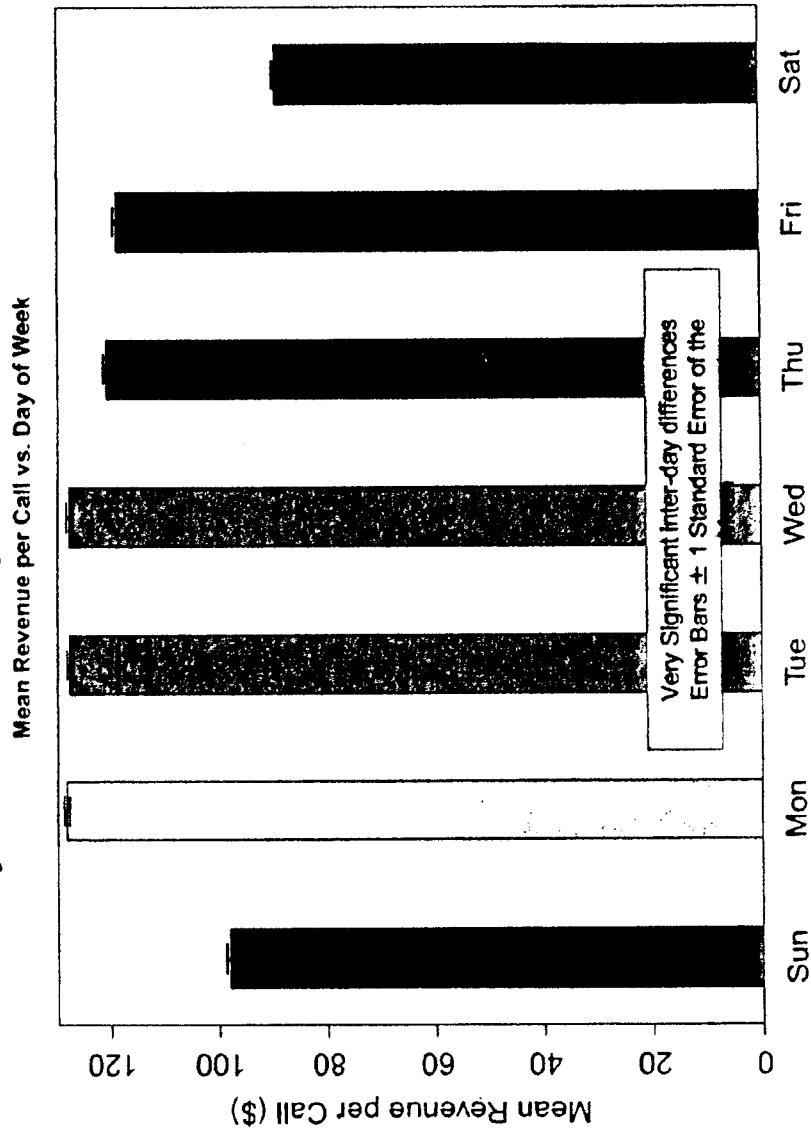


FIG. 26
Hour of Day Cycle – Epoch Averaged

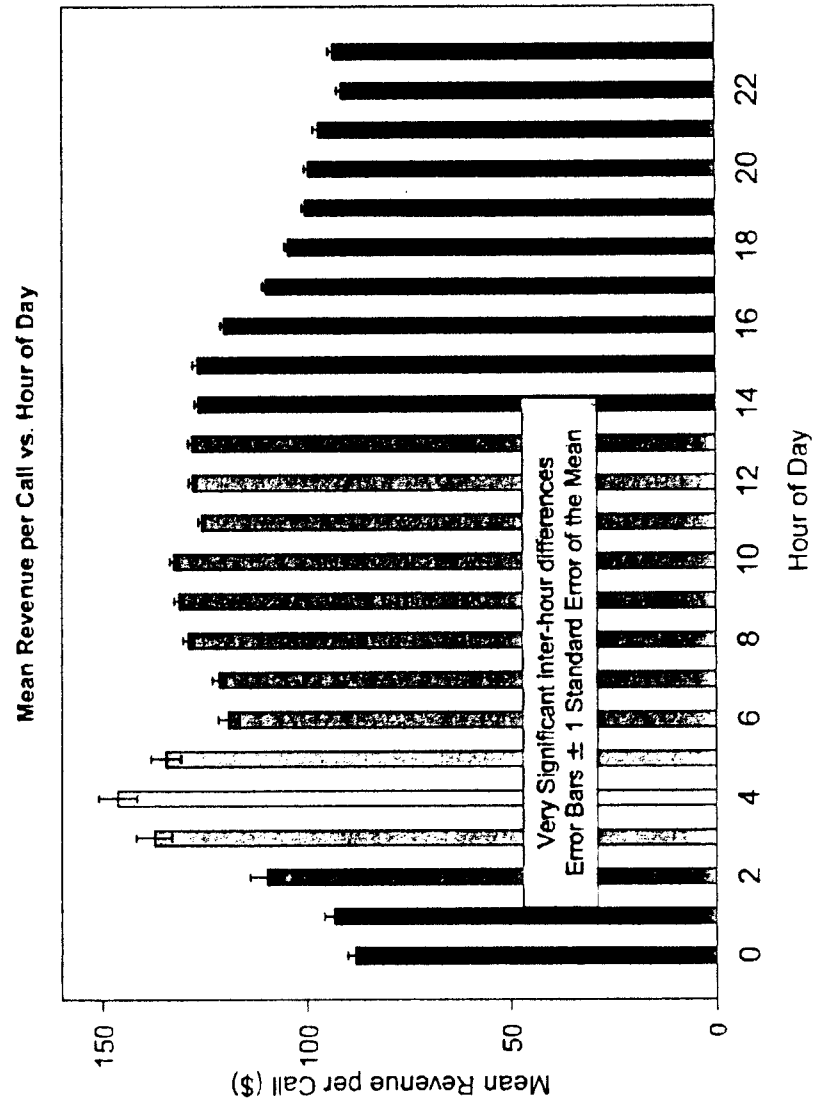


FIG. 27

