Applications of Agent Based Approaches in Business: A Three Essay Dissertation

Shankar Prawesh
University of South Florida, sprawesh@hotmail.com

Applications of Agent Based Approaches in Business

(A Three Essay Dissertation)

by

Shankar Prawesh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Information Systems and Decision Sciences
College of Business
University of South Florida

Co-Major Professor: Balaji Padmanabhan, Ph.D.
Co-Major Professor: Kaushal Chari, Ph.D.
Manish Agrawal, Ph.D.
Wolfgang S. Jank, Ph.D.

Date of Approval:
July 10, 2013

Keywords: Survey, Simulation, News Recommender Systems,
Urn Models, Outsourcing, Imitation

Table of Contents

List of Figures

List of Tables

Abstract

The goal of this dissertation is to investigate the enabling role that agent based simulation plays in business and policy. The aforementioned issue has been addressed in this dissertation through three distinct, but related essays. The first essay is a literature review of different research applications of agent based simulation in various business disciplines, such as finance, economics, information systems, management, marketing and accounting. Various agent based simulation tools to develop computational models are discussed. The second essay uses an agent-based simulation approach to study important properties of the widely used most popular news recommender systems (NRS). This essay highlights the major limitations of most popular NRS in terms of: (i) susceptibility towards manipulation and (ii) unduly penalizing the article which may have "just" missed making the cutoff in most popular list. A probabilistic variant of recommendation has been introduced as an alternative to most popular list. Classical results from urn models are used to derive theoretical results for special cases, and to study specific properties of the probabilistic recommender. In addition to simulations, various statistical methodologies are used, such as regression based methodologies as part of a broader decision analysis tool. The third essay views firms as agents in building regression based empirical models to investigate the impact of outsourcing on firms. Using an economy wide panel data of outsourcing expenses of firms, the third essay first

investigates the value addition by the IT backgrounds of project owners in managing IT

related projects. Then it investigates the impact of peer-pressure on a firm's outsourcing

behavior.

Chapter 1 : Dissertation Overview

Agent based models (ABMs) are used to emulate sophisticated computational and behavioral phenomena. Due to limitations of traditional econometric models and the 'dynamic stochastic general equilibrium' models to capture extreme events such as financial crises, agent based models have been used as a methodology to simulate economic phenomena (Farmer et al. 2009). This dissertation research explores the use of agent based modeling techniques to address various business and policy issues.

Overview of the Three Essays

The first essay positions the dissertation work based on a literature review of agent-based models in business. To understand the current state of the art in agent based simulations in business, articles using agent based simulation as a research methodology, were reviewed. The literature survey presents various research work related to agent based models in finance and economics, information systems, management, marketing and accounting. The computational approaches used to address different research questions are discussed. Developing an agent based model requires representing agent behavior, interaction and its environment through computer programs. Hence a summary of different development platforms and platform specific requirements of programming knowledge are also discussed.

The second essay examines an application of agent-based simulation in news recommender systems (NRS). The motivation for the work in online news recommenders

is to introduce a manipulation resistant NRS. In particular, this essay investigates the different properties of the "most popular" (or most emailed) NRS - widely used by most of the media websites. Through simulation this essay shows that whereas recommendation of the N most read articles is easily susceptible to manipulation, a simple probabilistic variant is more robust to common manipulation strategies.

In the context of NRS, manipulation can be understood as an act of a person (or group of person), when they try to artificially inflate the counts (or clicks) of a target article of their interest. This problem has been well recognized in the case of recommender systems (Weber 2010; Lerman 2007).

To address the main limitations that were identified, the second essay presents a probabilistic NRS. Probabilistic recommendation of articles is based on probabilistic sampling without replacement for N articles. The probability that an article will be recommended at any given time step is proportional to the count it has received thus far.

This research shows that for the "N most popular" recommender, probabilistic selection has many desirable properties. Specifically, the $(N + 1)^{th}$ article, which may have "just" missed making the cutoff, is unduly penalized under common user models. Small differences initially are easily amplified − an observation that can be used by manipulators. Probabilistic selection on the other hand, creates no such artificial penalty. Further, every article will have some chance to appear in the recommended list in the probabilistic NRS. Classical results from urn models have been used to derive theoretical results for special cases and to study specific properties of the probabilistic recommender. The urn models used for analytical derivations are namely, Pólya's and Bernard Friedman's urn models (Freedman 1965). The trade-off between the Top-N NRS and

proposed probabilistic variant has been discussed in terms of count distortion and information quality[1]. Finally, results on manipulation for the probabilistic NRS in comparison with an "adapted" *influence limiter* heuristic (Resnick and Sami, 2007), has been discussed.

Further, data from a local news website DailyMe Inc. has been obtained to determine the popularity distribution of articles. The data provided listed specific articles along with cookie IDs and time stamps read across the five different local news websites. The distribution generated through this dataset has been used to complement the findings from simulation results.

It has also been noted that the probabilistic mechanism proposed in the second essay has one limitation that it sometimes could pick (with low probability) articles that are not popular. It has been shown that a novel solution to this is through a class of probabilistic NRS with *feedback*.

The third essay of dissertation which views firms as agents, investigates the impact of outsourcing contracts as well as the impact of IT background of project leaders on firms through various financial measures such as operating expenses, overhead (selling, general & administrative) expenses and profitability. While the IT backgrounds of project owners result in bringing down expenses, projects managed by executives with non-IT backgrounds improved firm profitability. IT systems integration outsourcing projects during the period of 1995 – 2010 in US market, is used to establish these findings.

---

[1] Counts of articles has been assumed as the surrogate measure of quality

Further, a herding model of outsourcing behavior among firms has been developed. This herding model builds on the Information-based Theory of business imitation (Lieberman et al. 2006), and is operationalized using a two-step regression model. The impact of peer pressure and profit-margin on firms to undertake outsourcing activities is modeled in the first-step of the regression model. The estimated outsourcing decision from the first step is then used to predict overhead expenses of firms in the second step of regression. The effects of imitative behavior and profit margin are mediated through the action taken by a firm, on its overhead expenses. The use of peer-pressure to model outsourcing behaviors of firms, is a unique contribution to the outsourcing literature. Data on outsourcing contracts of firms in the automotive sector signed during the period of: 1995 – 2010 is used for analysis. Finally, this essay concludes with discussing the research opportunities related with representing outsourcing behavior of firms through ABM perspective.

References

Chou, C. L.-y., Du, T., and Lai, V. S. 2007. "Continuous auditing with a multi-agent system," *Decision Support Systems* (42:4), pp 2274-2292.

Farmer, J. D., and Foley, D. 2009. "The economy needs agent-based modelling," Nature (460:7256), pp 685-686.

Lieberman, M. B., and Asaba, S. 2006. "Why do firms imitate each other?," Academy of Management Review (31:2), pp 366-385.

Freedman, D. A. 1965. Bernard Friedman's Urn. *The Annals of Mathematical Statistics,* volume 36(3): 956-970.

Lerman, K., 2007. User participation in social media: Digg study. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology,* pages 255–258.

Lieberman, M. B., and Asaba, S. 2006. "Why do firms imitate each other?," *Academy of Management Review* (31:2), pp 366-385.

Resnick, P., and Sami, R., 2007. The Influence Limiter: Provably Manipulation-Resistant Recommender Systems. In *Proceedings of ACM Conference on Recommender Systems (RecSys07).*

Weber, T. E. 2010. Cracking the New York Times Popularity Code. *The Daily Beast*, December 19, 2010.

Chapter 2 : A Survey of Agent Based Simulation in Business

Introduction

The rapid growth in computing resources has caused a dramatic shift in the way research is conducted in many fields of study. A case in point, analytical modeling has limitations in terms of capturing rich details related to model dynamics. However, simulation modeling can address some of these shortcomings especially in the context of complex behaviors and systems that require analysis of multiple interdependent processes (Harrison, Carroll, & Carley, 2007).

ABM can offer a specific kind of computational simulation for complex business situations. ABM is based on the notion that the whole of many systems is greater or more complex than the simple sum of its constituents called agents (North and Macal 2007). It is a tool to simulate complex systems for the purpose of studying emergent behaviors (Bonabeau 2007). Agent-based models are specified using either equations or rules, or both. It focuses on modeling the behavior of adaptive actors who make up a social (or complex) system. The process of viewing organizations as agent based systems can provide valuable insights into emergent behaviors. It also provides the flexibility to generate and experiment with various 'if-then' conditions.

Agents are autonomous decision making entities that can make assessment of situations in making decisions. Agents may change or evolve, allowing unanticipated behaviors to emerge. Their decision is determined through a given set of rules

(Bonabeau, 2002). At the simplest level, ABM represents agents and the relationships among them. However, sophisticated models may incorporate learning and adaptation behavior of agents through evolutionary techniques (Bonabeau, 2002). The development of agent based models is usually grounded on empirical data that provide broad parameters for the simulation.

Agents operate in an environment where the interactions among them lead to observed phenomena known as the emergent behavior. Usually, the environment and behavior of agents are generated in simulation through statistical distributions observed in data. The emergence of computational social science is based on the use of ABM in economics, sociology, business and political science.

In particular, the focus in this essay is on the use of ABM and its precursor, the computer simulation models in business domains. To understand the current state of the art of ABM research in business, we reviewed the following major representative journals in each of the following fields: Finance and Economics, Accounting, Management Information Systems, Marketing, and Management. The list of sample journals include (but not limited): *MIS Quarterly*, *Information Systems Research*, *Management Science*, *Decision Support System*, *Journal of Management Information System*, *Organization Science*, *Administrative Science Quarterly*, *Academy of Management Journal*, *Academy of Management Review*, *Strategic Management Journal*, *The Accounting Review*, *Journal of Accounting Research*, *Journal of Accounting and Economics*, *Accounting Organizations and Society*, *Journal of Marketing*, *Marketing Science*, *American Economic Review* and *Quantitative Finance*. Some other outlets were also searched especially for ABM in Finance and Economics. In the following sections

7

few ABM research examples in each field are presented in detail. Finally, we conclude with a discussion on the platforms for implementing ABM.

Finance and Economics

In recent years there have been major developments in agent-based simulation to represent various phenomena in economics. An agent-based software platform called EURACE[2] has been developed for the simulation of the European economy for optimizing the impact of regulatory decisions. Another example is Minsky[3], a software program for designing monetary macroeconomic models.

In economics research, the major limitations of efficient market hypotheses and rational expectations assumptions (Lo, 2004; The Economist 2010), can be overcome by ABM. Also, traditional and analytical modeling approaches in economics fail to capture events related to crisis situations. ABM has been suggested as an approach to model crisis and early-warning systems (Farmer et al. 2009; The Economist 2010). Whereas in traditional economic models, interactions among agents take place indirectly through pricing, in ABM direct interactions among agents can be modeled. Further, ABM does not assume equilibrium in an economy. Large scale projects such as Eurace, CRISIS and FuturICT have used ABM to model (Iori & Porter, 2012).

ABM has been extensively used in financial economics. Suggestions have been made regarding modeling financial markets as various possible simulation outcomes similar to traffic forecasting models (Iori & Porter, 2012).

With the use of ABMs, the theoretical assumptions related to equilibrium conditions have been replaced by less restrictive assumptions requiring agents to have

---

[2] www.economist.com/blogs/freeexchange/2013/01/remaking-macro-0
[3] www.sourceforge.net/projects/minsky

bounded rationality, where they adapt to market forces. Agents may also use technical rules such as artificial neural networks to forecast. In the finance literature, agents have been represented in many forms ranging from passive automatons with no intelligence and to data gathering decision-makers with sophisticated learning capabilities (Iori & Porter, 2012).

There are some excellent reviews of research on ABM in finance and economics (Chakraborti et al., 2011; Iori & Porter, 2012). A critical review of ABMs in Economics has been presented by Cristelli et al (2011).

Cristelli et al. (2011) discuss why and how ABM advance the understanding of the dynamics and the statistical properties of financial markets beyond the classical theory of economics. In this review Cristelli et al. (2011) discuss eight different agent based models in the field of economics. These models are as follows:

Kim and Markowitz Model: Kim et al. (1989), who were first to recognize the potential of ABM in finance, presented a model to explain Black Monday. Their findings are based on discrete event simulation of stock market.

Santa Fe Artificial Stock Market (Tesfatsion & Judd, 2006): An artificial stock market developed by researchers at Santa Fe to investigate market trends in the presence of heterogeneous forecasting strategy of agents.

Minority Game (Challet et al. 1998): One of most widely studied strategies in finance in which agents receive payoff if their strategies belong to the minority side in a game.

Caldarelli, Marsili and Zhang: Caldarelli et al. (1997) show that endogenous mechanisms of financial markets are sufficient to obtain a stable and self-organized market.

Lux and Marchesi: Lux et al. (1999) show that the scaling laws observed in financial markets could be generated through agents' mutual interaction.

Giardina and Bouchad: Giardina et al. (2003) have introduced a model on the tractability of Minority Games and the Santa Fe virtual stock market.

The destabilizing effect of leverage: Thurner et al. (2012) show the phenomena of fat tails and volatility clustering as a result of leverage and margin calls.

Credit network and bankrupt avalanches: Gatti et al. (2009) study the property of a credit network and the causes of the emergence of bankruptcy avalanches.

Based on findings from the aforementioned models, Cristelli et al. (2011) have made suggestions regarding open questions, and highlight under-researched issues such as non-stationarity and self-organization in the context of financial markets. Farmer and Foley (2009) have pointed out the limitations of econometric models and 'Dynamic Stochastic General Equilibrium' models in terms of making strong assumptions such as perfect world and minimal deviation in the future from the current state. Hence these models by design, fail to capture the great changes during financial crises (such as the financial crisis that started in the last quarter of 2007). Farmer and Foley (2009) also point out that most mathematical models in practice are used to calculate potential profit and risk of individual trades without assembling the various pieces to understand the whole economic systems. In light of aforementioned limitations of current modeling practices, ABM has been suggested to capture the wider range of non-linear behavior,

and to simulate an artificial economy under different policy scenarios (Farmer & Foley, 2009).

One of the earliest research in finance, related to the use of ABM is in behavioral finance, where traders are represented as agents encompassing trader behavior through psychological or sociological properties. LeBaron (2006) has provided an overview of the use of these classes of models. However, utility functions defined in these models are not necessarily true representation of reality.

Cont and Bouchaud (2000) have introduced the concept of noise traders who form random clusters of traders sharing similar outlook on financial markets (termed as herding). This idea has been further explored in later research. ABMs have been categorized into two categories (Chakraborti et al. 2011): (a) order-driven models and kinetic-theory for wealth distribution and (b) game theoretic modeling. Below we discuss such research in each of these categories (Chakraborti et al., 2011).

Order driven models and the kinetic-theory of wealth distribution. Chiarella and Iori (2002) have built an ABM where different types of traders submit orders based on different strategies: chartist, fundamentalist and noise traders. Orders are considered as particles moving along a price-line, where each collision is a transaction (Chakraborti et al., 2011). In these models, orders are viewed as an arriving flow, whose properties are determined by empirically observing the trading mechanism, thus leading to phenomena called 'Stylized Facts' i.e., empirical properties that could be observed on a large number of market orders (Chakraborti et al., 2011). Findings based on stylized facts lead to the concept of 'zero-intelligence traders' (ZI). The expression of ZI traders is described by Gode and Sunder (1993). These traders are termed as ZI traders because they have no

intelligence, do not seek to maximize profits, and do not have learning capability or memory.

Caldarelli, Marsili, and Zhang (1997) present traders' strategies and speculation on endogenous price fluctuation, using a prototype model of stock market interaction among traders without external influences. The model generates realistic price histories that have statistical properties similar to those observed in the real world. LiCalzi and Pellizzari (2003) present ABM of market dynamics based on structural assumptions that represent trading mechanism, and behavioral assumptions of traders. Findings by LiCalzi and Pellizzari (2003) support the hypotheses that statistical properties of financial time series are due to the microstructure of the stock market.

Lux and Marchesi (1999) have developed a multi-agent model of financial markets involving two groups of traders, namely, fundamentalist and noise traders (or chartist). Fundamentalist follows the strategy based on efficient market hypothesis, and expect price to follow discounted sums of expected future earnings. This strategy consists of buying or selling assets, when prices are perceived to be below or above the fundamental value of assets. Noise traders on the other hand, identify price trends and patterns, and consider behavior of other traders, thereby, gravitating towards herding behavior. Their model supports the notion that the scaling laws in finance arise from mutual interaction of agents with heterogeneous beliefs and strategies, and that alternation between tranquil and turbulent period is a result of changes in membership of groups (Lux & Marchesi, 1999).

LeBaron (2006) has described the development of the Santa Fe Artificial Stock Market used to understand the behavior of environments having evolving trader behavior

(Tesfatsion et al., 2006). Agents use a classifier system to estimate the returns. The classifier is based on a number of properties, which in turn are mapped to different parameters. Periodically, the worst set of classifier rules are removed and replaced with new rules generated by a genetic algorithm (GA). The model has been able to reproduce many empirical phenomena observed in financial returns such as: excess kurtosis, low linear autocorrelation and volatility clustering. Some of the later developments in this stream include: Chiarella & Iori(2002) and LeBaron & Yamamoto(2007).

Challet and Stinchcombe (2001) have reported the statistical analysis of the Island ECN order book. They analyze the static and dynamic properties of the system by treating orders as massive particles, and price as the position of the particles. Cont and Bouchaud (2000) have been able to generate excess kurtosis of financial returns while representing traders, who imitate each other, using agents. The theoretical foundation provided them has guided many subsequent research in ABM in finance. For example, Feng, et al. (2012) have constructed an ABM to quantitatively demonstrate that "fat tails" in return distributions arise when traders share similar technical trading strategies and decisions.

Mike and Farmer (2008) have developed a behavioral model for liquidity and volatility based on empirical regularities in trading flow in the London Stock Exchange. In this empirical study involving a group of stocks the authors observe that large fluctuations of absolute returns behave as per, power law.

Cont (2007) has also proposed a simple ABM that links variations in market activity to threshold behavior of market participants, thereby leading to the phenomena of volatility clustering and investor inertia. They define volatility clustering as large changes

in prices that tend to cluster together, resulting in the persistence of amplitudes of price changes. Due to the simplicity of the model, the origins of volatility clustering are traceable to the agent behaviors.

Using maximum likelihood estimation in the context of ABM, Lux (2012) has produced abrupt changes of mood in short-run sentiment, and slower changes in medium-term sentiment, where the influence of social interaction is less pronounced.

Tedeschi, Iori, and Gallegati (2012) have introduced an order driven model with heterogeneous agents that imitate each other on a dynamic network. They implement an endogenous mechanism of imitation using 'preferential attachment', such that each trader is imitated by others with a probability proportional to its profit. The mechanism of link formation allows the authors to investigate assumptions under which the most successful traders endogenously rise and fall over time, as well as the imitation effects on asset prices and the distribution of agents' wealth.

Treating each agent as a gas molecule, and each trade as an elastic or money-conserving two-body collision, Chatterjee et al. (2004), have simulated an Ideal Gas Model of trading markets. They introduce agent heterogeneity using saving propensity of agents. Chakrabarti and Chakrabarti (2009) have developed a framework based on microeconomic theory from which ideal gas like market models can be represented. They introduce a kinetic exchange to model N-agent exchange economy, where agents have same statistical properties.

Game theoretic models. Challet and Zhang (1997) have introduced and analyzed binary games where N players have to choose one of the two sides independently and those on the minority side win – hence the game is termed as 'minority game'. This

model also incorporates the notion of bounded rationality among players, as they are allowed to make decisions based on finite set of ad-hoc strategies based on the past record. The analyzing power of agents is limited and can adapt when necessary. The approach of Challet and Zhang (1997) has been one of the earliest efforts in modeling emerging intelligence and cooperation among agents in a society through ABM. In an extension, Challet and Zhang (1998) have analyzed the ability of players to learn a given payoff. They introduce the concept of Darwinism to allow worst player to be replaced by a clone of the best.

Vriend (2000) has demonstrated the difference between individual and social learning through an example of a standard Cournot oligopoly game. Each individual firm does not know what the optimal output level is, which, it needs to learn. This problem has been modeled through a GA. The GA has been implemented in two ways. In the first implementation, the GA is used to model social or population learning, in which firms look around and tend to imitate and re-combine ideas of other firms that appear to be successful. The more successful the selection rules are, the more likely they are to be selected for the process of imitation and re-combination, where the measure of success is the profits generated by reach rule. In the second implementation, GA has been used to model individual learning. In individual learning, each individual selects a rule based on its fitness. The rules that had been more successful recently, are more likely to be chosen. Hence instead of looking how well other firms with different rules were doing, in the individual rule, firms check how well it had been doing in past, when these rules were used. Vriend (2000) found that social learning produced a much higher average output than individual learning.

Sysi-Aho et al. (2004) introduced an adaptation mechanism based on genetic algorithms to model minority games. when agents find their performances too low, they modify their strategies to improve their performances and to become more successful. The authors observed that adaptation results in competition among agents that in turn pulled the collective system into a state where the aggregate utility was the largest. Another example of modeling games is the Kolkata Paise Restaurant problem. Which is a repeated game played between a large numbers of agents having no interaction among themselves (Chakraborti et al., 2011). Prospective agents choose from N restaurants simultaneously on a given day. Each restaurant has different rank but the same price for a meal and can serve only one agent. Information regarding distributions of agents on the previous day is available, as each agent try to choose a restaurant with highest possible rank, while avoiding the crowd. If multiple agents arrive at any restaurant on any day, then one among them is chosen randomly, and rests are not served (Chakraborti et al., 2011).

From simple zero intelligence (randomly behaving) agents, ABM of financial markets has evolved to modeling using sophisticated agents with micro foundations. The cases where zero intelligence agents may perform poorly are situations where agents have opportunities for learning, along with feedback loops between agents' action and the state of environment (Ladley, 2012). The research on considering the financial system as network, which is still in early stage, can help address various important issues related to optimal network design of banking systems. (Iori & Porter, 2012).

Chen (2012) discusses the origins of agent-based computational economics from markets, cellular-automata, tournaments and experimental-economics perspectives. The

market origin has its roots in the competitive general equilibrium model proposed by Leon Walras, whose work inspired the Paris Stock Exchange. The notion to displace the Walrasian auctioneer by a decentralized process has been a major motivation for using ABM in economics. Building on the notion of simple agents, autonomous agents have been introduced and ecologically constructed. The tournament origin has been a precursor to autonomous agents. Recent attempts have been made to use human-like agents that have personality, emotions and cultural backgrounds, in economics research studies.

Economic policies. Poledna (2011) has provided recent examples of ABM for economic policies development. A notable work in this field is by Thurner et al. (2012), where they build a simple model of leveraged asset purchases with margin calls. It has been shown that fat tails and clustered volatility are results of leverage limits that cause funds to sell in a falling market instead of "irrational behavior" of traders. Haber (2002) presented a macroeconomic ABM of national economy by simulating both private and public sector. Households, companies and government agencies are treated as separate agents in the formation of fiscal and monetary policy.

The Tobin Tax is usually imposed on all foreign exchange transactions which should discourage short term speculation while leaving longer term investors relatively unaffected to reduce market volatility (Iori & Porter, 2012). Mannaro, et al. (2008) have examined the effects of Tobin Tax on foreign exchange and stock markets, using an artificial financial market based on heterogeneous agents. Through simulation findings, authors have found that the tax actually increases volatility and decreases trading volume.

Recently Geanakoplos et al. (2012) have developed an ABM of housing markets with individual data from greater Washington DC area. Their findings suggest that housing boom and bust of 1997-2007 was largely driven by leverage. Their model consists of every household of the economy, with tremendous heterogeneity.

Calibration of agent-based models for financial markets. A critical guide to empirical validation of ABM has been discussed by Fagiolo, Moneta, and Windrum (2007). A detailed survey on major approaches for empirical validation of agent-based models has also been discussed. The artificial data should be compared with real data and the structural parameters of the model should be tuned in such a way that simulated data imitates real data (Iori & Porter, 2012). The method of simulated moments has been suggested as a possible solution of this problem. To deal with validation and estimation of agent based models by means of simulation methods based on actual data, Gilli and Winker (2003) have proposed a continuous global optimization heuristic. The estimation results of some parameters for a standard ABM of the DM/US-$ exchange rate has been also discussed. Using the CEBI database for Italian firms, Bianchi et al. (2007) have discussed validation experiments for ABM. Initial setup and the model parameters have been estimated using actual data. Ex-post validation of simulation results with respect to actual data point to, the success of agent-based models in reproducing the observed reality. In another example, Mike and Farmer (2008) have developed a simple ABM for trading order flow in the London Stock Exchange in which, all components of the model are validated against real data. This model of order flow has been used to simulate price formulation under a continuous double auction. The model is developed based on a single

stock and then tested on 25 stocks. The predictive capability of this model was good for low volatility and small tick size stocks.

Information Systems

Broadly, agents have been used in Information Systems in application areas such as auctions, organizational IT-use, software engineering and network driven phenomena. In one of the earlier works, Chari (2000) presented an overview of software agents, agent applications and research opportunities in the information systems domain, where software agents were defined as a collection of computer programs that act on behalf of some entity and have some intelligence. In Chari (2000), software agents were categorized as automation agents, information agents, transaction agents, workflow management systems and monitoring & control agents. The use of software agents in different contexts has been also discussed by Chou et al. (2007). In the remaining part of this section, specific works in different areas of information systems have been discussed.

Software agents. One context in which agents have been studied in information systems is the concept of software agents that work on behalf of users. While some of the papers in this area focus on intelligent agents that can perform tasks based on learning user preference, others in this area deal with multiple interacting agents, that is closer to traditional agent-based simulations.

Yang et al. (2000) discuss the development of intelligent internet search agents based on hybrid simulated annealing, where an intelligent search agent refers a search engine that has the capability to make adjustments according to the progress in searching and to generate personalized results according to users' preferences.

An agent based recommender system for web search has been proposed by Birukov et al. (2005). Agents use data mining techniques in order to learn and discover user behavior, as they interact with other agents to share knowledge about their users. The improvement in performance of the overall search engine has been demonstrated through experimental results.

Du et al. (2005) have proposed a framework for using mobile agents to highlight the autonomous behavior of firms in the e-marketplace in terms of allowing corporate data to be maintained by local buyers and sellers. They discuss findings based on simulation design to explore the performance of mobile agents in different product-purchasing policies. The advantages of agents have been discussed in terms of aggregating orders and shortening the execution time.

Wainer et al. (2007) present a set of protocols for scheduling a meeting among agents that represent their respective user's interest. Multi-agent scheduling systems have been used to present simulation results for different protocols.

A model for software agents that can automate negotiations by allowing agents to learn from bidding behavior of opponents has been presented by Chari et al. (2007). In this research, agents' behaviors have been modeled through a multi-issue heuristic (MILH). The simulation results indicate that software agents can replicate the behavior of human negotiators.

To analyze and understand a dynamic power change in the US wholesale electricity market, Sueyoshi et al. (2008) have developed an intelligent decision making tool MAIS, based on agent-based systems. The software uses probabilistic reasoning and reinforcement learning to assess different trading strategies in a competitive electricity

trading environment. Market entities such as generators, wholesalers, market administrator, network operator and policy regulator have been represented as software agents in the system.

Lau et al. (2008) have modeled intelligent software agents that use a probabilistic decision making mechanism in a simulated e-market place for negotiation. They show that probabilistic negotiation agents empowered with knowledge discovery mechanisms are more effective and efficient than Pareto optimal negotiation agents in e-market places.

Nunamaker et al. (2011) have proposed an IT artifact called Embodied Conversational Agent – based kiosk for automated interviewing based on detecting changes in arousal, behavior and cognitive effort. Software agents use heterogeneous sensors to detect human physiology and behavior during interactions. They have argued that these agent-based systems make knowledge-based recommendations and exhibit human characteristics such as rationality, intelligence, autonomy and environmental perception (Nunamaker et al. 2011). Recent works in the context of negotiation agents include: Lin et al. (2013); Ren et al. (2013).

Auction. Auction mechanisms where potential buyers place competitive bids on assets and services have been widely studied in economics traditionally, but more recently in information systems, where computational models are used for studying auctions, have been an active area of research. With the advent of internet, bidders often partcipate in auctions online without being physically present at the auction site. Bapna, Goes, and Gupta (2003) used multiagent-based simulations to present a relatively risk-free and cost-effective environment to both bid takers and bid makers in a web based

dynamic price setting process. The optimization of bid taker revenues and welfare implications have been considered in model development. The simulation is based on theoretical revenue generating properties, with simulation parameters instantiated using real online auctions. The authors justify the simulation approach, as a test bed for design choices of auctioneers and bidding strategy of bidders.

Mehta and Bhattacharyya (2006) present design, development and validation methodologies of an agent-based model for B2C electronic auctions. The model involves incorporating the behavior of an auctioneer, consumer and retailer, and the environment in which agents operate. The simulation based approach provides additional insights on market characteristics such as alternative distribution of posted prices, demand for items, and degree of product differentiation in market, as well as consumer characteristics and auction parameters.

Jones et al. (2006) have used agent-based simulation of the market for television advertising slots in order to analyze an e-market design that allows multiple market segments to be served simultaneously with a single rule-based combinatorial auction. Other notable research involving agent-based approaches in auction include Avenali et al. (2007) and Gregg et al. (2006). Descriptions of electronic market simulators have been provided by Fasli et al. (2008).

Adomavicius, Gupta, and Zhdanov (2009) have discussed analytical, computational, and empirical analysis of strategies for intelligent bid formulation that provide opaque feedback information to bidders, and present a challenge in formulating appropriate bids. Software agents have been used in making bids in the presence of limited information provided by the mechanism. In the context of multi-item auction,

Bichler, Shabalin, and Pikovsky (2009) benchmark different iterative combinatorial auctions and design, and analyze new auction rules for auctions and pseudo-dual linear prices using a simulation model.

Recently, Guo, Jank, and Rand (2011) have proposed an ABM that simulates bidders with different bidding strategies, and their interactions with others. The model has been calibrated by matching the emerging simulated price process with that of the observed auction data using a genetic algorithm. The proposed methodology has been applied in the context of eBay auctions for digital cameras.

Greenwald, Kannan, and Krishnan (2010) have developed a partially observable Markov decision process model of supplier bidding behavior, and use a multi-agent e-market simulation to analyze the effects of complete and incomplete information policies on the expected price paid by the procurer. The information revelation policies have been developed using ideas from the multi-agents literature, the machine learning literature, and the economics literature.

Social networks. Chang, Oh, Pinsonneault, and Kwon (2010) have used ABM to investigate the outcome of strategic alliances between two smaller online search engines competing with a dominant market leader in settings where an advertiser's decision is based on the result of network influence, and the advertiser's individual preferences. The ABM consists of modeling agents' (advertisers) behavior in response to the environment (strategic alliances). Influence relationships between online advertisers have been modeled through three types of network: scale free, small world and random. The findings suggest that in the presence of network influence and cascading effects, an

alliance with near half market share could compete with a leader with majority market share.

Organizational use of IT. In one of the earliest research in Information Systems on modeling humans and their interactions in a team as objects in a computerized environment, Rao et al. (1995) have developed computer programs to model decision systems and team processes drawing ideas from team theory, informational processing and social choice paradigms. Extending on the coordination framework provided by this research, Raghu et al. (2004) present an approach to organizational modeling that combines both agent-centric and activity-centric approaches. The agent-centric approach captures specific aspects of the human component.

Nan (2011) has presented a theory-building approach through modeling collective level information technology (IT) use patterns from a bottoms-up approach. ABM has been introduced as a tool for computationally representing IT use process into three interrelated elements: agents in an IT use process, interactions related with mutually adaptive agent-behavior, and the environment of organizational IT use.

Open source software. In the software engineering context, Oh and Jeon (2007) have investigated the basic pattern of interactions among open source software (OSS) community from the Ising theory perspective – widely used in physics. The model has been implemented using simulation, treating OSS community members as agents on empirical data collected from two OSS communities. They conclude that: (1) membership herding is highly present when external influences are weak, but decreases when external influences increase, (2) propensity of membership herding is most likely to be seen in a large network with random connectivity, and (3) in large networks, when

external influences are weak, random connectivity result in higher network strength than scale-free connectivity.

Zaffar et al. (2011) have proposed a framework that investigates a broad range of social and economic factors on the diffusion dynamics of open source software using an agent based computational economics approach. The authors illustrate the impact of key variables such as: license, support and interoperability costs, frequency of upgrades and interactions with firms on the diffusion dynamics.

Simulation as a Decision Support Tool. Valluri et al. (2005) have used an agent-based model to study game theoretic supplier selection, where neither the suppliers nor the buyers possess full information. Agents have been used to model suppliers who learn to produce at optimal levels through a pre-specified system of rewards and punishments administered by the buyer. Supporting their findings both theoretically and through a Japanese automotive-market the authors conclude that it is optimal for buyers to transact with relatively few suppliers.

Wang et al. (2007) discuss a multi-agent simulation method to simulate heterogeneous project-team coordination and argue that it is a valid decision support tool for IT investment decisions. Their simulation model is based on a theoretical framework, and is validated using a real world case from plastic tooling industry.

The phenomena of knowledge sharing in an organization is analyzed by Wang et al. (2009), using an agent based model. In particular, they simulate employee knowledge sharing behaviors by making parametric assumptions on employee decision strategies and organizational interventions. Authors have also argued that agent-based approaches can

be combined with data mining to develop management analysis tool to study organizational knowledge sharing.

Agent based simulation has also been used to assess the relative performance of reinforcement learning systems (Gaines et al. 2013). The classifier systems have been implemented in the context of the Iterated Prisoner's Dilemma.

Agent based modeling in operations management and supply chain. The computational model to capture the components of a supply chain has been implemented by Strader et al. (1998) using multi-agent simulation. This model studies the impact of information sharing on order fulfillment in divergent assembly supply chains. The authors have also argued that simulation is the most appropriate tool to study processes such as supply chains that exhibit decentralized command and control.

Kim (2009) has introduced the notion of modeling supply chain as complex adaptive systems, where firms (or agents) interact with one another and adapt themselves. A social factor, trust, is used in modeling an agents' behavior.

In other notable examples, Liang et al. (2006) have developed a multi-agent system to simulate a supply chain where agents operate with different inventory systems, while Kim et al. (2010) have used agent negotiations to allocate orders to participants for supply chain formation.

The use of simulation based approaches to model games in operations management has been discussed by Van Der Zee et al. (2012). Lovric et al. (2012) have studied revenue management for public transport operators through agent-based modeling. The modeling approach of Lovric et al. (2012) has been evaluated using real world smart card transaction data.

Organization and Management

Though a popular research methodology in physical and social sciences, simulation historically was under-represented in management research, primarily because simulation methods were not well understood. Recently though, about 8% of journal articles have used simulation methodology (Harrison et al. 2007).

In this chapter, the discussion on organizational research is restricted to agent-based models, in particular efforts related to genetic algorithms, NK model, and cellular automata (Harrison et al., 2007). In management research, ABM can also be framed using tools such as longitudinal social network analysis and game theory (Fioretti & Lomi, 2011).

Simulation allows management theorists to make realistic assumptions that may in turn lead researchers to generate hypotheses that are integrated and consistent (Harrison et al., 2007). Simulation based approaches (such as agent-based simulation) can contribute to organization research in following ways (Axelrod, 1997; Harrison et al., 2007):

Prediction: Analysis of simulation output may reveal relationships among variables that can be viewed as prediction of the simulation model.

Proofs: Some kinds of existential simulation can show that it is possible for the modeled processes to produce certain kinds of behavior.

Discovery: Simulations can be used to discover unexpected patterns due to interaction among agents.

Explanation: If simulation outcomes are close to observed behaviors, then postulated processes can be possible explanations for the behaviors.

Critique: Simulation can be used to examine the theoretical explanation for the observed phenomena, and also to explore alternative explanation for the observed phenomena.

Prescription: A simulation model can provide a better way of organizing or performing a certain task.

Empirical Guidance: Development of theories and models using simulation may guide us towards the possibility of uncovering systematic connections among previously unconnected variables.

Simulation in organizational research. Organizational decision making is a combination and alignment of individual visions and desires through cognitive and political processes (Fioretti & Lomi, 2011). For analyzing processes that lead individual agent's goals and decisions to a macro level organization behavior, ABM is a promising approach.

An ABM views group members as agents who receive and categorize information, carry out a few actions that are relevant to their specialization, and pass on tasks to other team members. This kind of modeling approach is able to distinguish between the mental model of a single agent and those who act in presence of other team members. For the problems studied by ABM, analysis of decision process is considered more important than final outcomes and equilibriums of the model.

The organization learning curve can also be modeled using an ABM where agents represent workers exploring different possibilities of production, until stable routine emerges for cumulative production. In this model, the overall behavior – i.e., the shape of

learning curve depends on individual interactions and also on the speed at which equilibrium production time is achieved.

One of the earliest examples of simulation in organizational research is by Cohen, March, and Olsen (1972). Here organized anarchies of organizations such as problematic preferences, unclear technology and fluid participation are modeled through computer simulations. Modeling cultural heterogeneity as an emergent organizational property, Harrison and Carroll (1991) present agents as members of an organization who influence each other's enculturation and turnover behavior through social influence. March (1991) has studied the problem of allocation of resources between the exploration of new possibilities and exploitation of old certainties. March (1991) drew ideas on complexity theory from biological sciences to examine the trade-offs. In their model, the key elements are: external-reality, individual beliefs, belief update and organizational code. Organization learning process depends on the interaction between groups, the complexity of task environment and the balance of exploration/exploitation within and between groups. Their simulation study has been able to explain the widespread use of exploitation, arguing that if task environment is complex then exploitation cumulating in the knowledge pools of different groups may provide sufficiently good results.

Simulation is increasingly becoming a significant methodological approach for theory development in strategy and organization (Davis et al., 2007), where theory is defined as consisting of constructs linked together by propositions that have an underlying, coherent logic and related assumptions. Simulation enables the exploration, elaboration and extension of simple theories into logically precise and comprehensive theory (Davis et al., 2007).

Building on the approach of (March, 1991), Fang, Lee, and Schilling (2010) have explored how the degree of subgroup isolation and intergroup connectivity influences organizational learning. The interaction pattern among the members of an organization is modeled through a "connected cavemen" model.

There has been consistent increase in use of ABMs to study innovation networks, especially in biotech firms in order to undertake joint research on specific products. Innovation networks evolve out of decisions made by its component firms (Gilbert et al., 2001).

Rivkin and Siggelkow (2003) have used agent-based simulation to examine how and why elements of organizational design depend on one another. They identify sets of design elements that encourage broad search and others that promote stability. Rudolph and Repenning (2002) use previous case study as the basis of simple theory describing how minor events could lead to catastrophes. Simulation has been very effective to study other basic organization processes such as competition and legitimation (Lomi & Larsen, 1996), and imitation and experimentation (Zott, 2003).

Coen and Mritan (2011) examine the dynamic capability of resource allocation to invest in operational capabilities. Using ABM, they model a process of firms competing in factor markets for opportunities to invest in existing capabilities and acquire new ones. The authors conclude that, endowment and search ability both matter, and that in many circumstances, the effects of possessing a superior endowment dominate the effects of superior search ability. Cardinal et al. (2011) examine how new product development performance is affected by product design and the technological environment. Agent-based simulation based on information processing theory has been used to specify and

refine an initial set of theoretical propositions. The authors find the existence of performance trade-offs in product development as well as the importance of performance priorities in influencing project design.

Recently agent-based simulation has been used by Baumann and Stieglitz (2013) to show how firms can improve performance by offering low-powered rewards for the selection and implementation of employee ideas. In purely low-powered ideas, an employee receives no incentives if his ideas are implemented, whereas, in purely high-powered ideas, an employee will accrue all benefits if his ideas are implemented.

To get better understanding of agent based approaches in management research, few simulation approaches are discussed that formulate the basis of ABM, with notable research examples related with them in the management literature.

NK fitness landscapes. This approach focuses on how rapidly and effectively modular systems reach to an optimal point, especially when interactions among system components ("agents") are important (Davis et al., 2007). The system is conceptualized as a set of N nodes and K interactions among the nodes. The system is assumed to use adaptation or search to find the optimal point. For example, Rivkin (2000) has addressed the issue of replication and imitation from NK fitness landscape perspective. N is defined as the elements of strategy and K as the degree of interaction among the elements. Rivkin (2001) used agent-based simulation based on NK models to investigate the structural reasons for the ease of replication and the difficulty of imitation for moderately complex strategies. The decision problem is modeled such that the number of decisions which together constitute a strategy, and the degree to which those decisions interact with one

another, determine firm performance. Simulation findings reveal the value of superior but imperfect information on good solutions to hard problems.

Gavetti and Levinthal (2000) have examined how experience and cognition affected the time needed to find an optimal policy for an organization. In their context, organizational policy is represented by N and K interactions among them. A fitness landscape is created by assigning performance values to every combination of values. When there is little interaction, there are few optimal combinations and as interaction increases, more combinations become locally optimal. The fitness landscape is "rugged" and it is hard to traverse to find the optimal point.

Siggelkow and Levinthal (2003) have used ABM to study the value of three different organization structures: a centralized organization, a decentralized organization and a temporarily decentralized organization to maintain exploration and exploitation strategies for firm performance. Firms with different organizational structures are "released" on performance landscape, over a number of periods, firm search over this landscape for high performing activity configuration. Here, performance landscape is defined as a mapping of all possible sets of a firm's choices onto performance values. By comparing the performance of firms with different organizational structures over a large number of landscapes, performances of different organizational structures were examined.

Gavetti, Levinthal, and Rivkin (2005) have used ABM to examine how firms discover effective competitive positions in worlds that are both novel and complex. They argue that analogical reasoning may be helpful, allowing managers to transfer useful

wisdom from similar settings they have experienced in the past. To generate family of landscapes, the authors have used an adaptation of NK model.

Lazer and Friedman (2007) examine how the structure of communication networks can affect the system level performance. Using ABM, the authors have presented a model of information sharing in which the less successful emulate the more successful. Simulation results suggest that when agents are dealing with a complex problem, the more efficient the network at disseminating information, the better the short-run, but lower the long-run performance of the system. NK problem space specification has been used to model numerical problem spaces.

Levinthal and Posen (2007) develop and test a model on the effectiveness of selection processes in eliminating less fit organizations from a population when organizations are undergoing adaptive changes.

To gain insights into the effects of differential reliability on the efficacy of selection, ABM using NK methodology has been implemented. The authors conclude that selection may be systematically prone to errors and that selection errors are endogenous to, and differ markedly across firms' search strategies.

A coupled search process in an organization is described as managers' activity to search for high-level choices that shape the search for low-level, operational choices, which in turn determine performance. Using ABM, Siggelkow and Rivkin (2009) show that coupled search processes obscure the performance impact of high-level choices through two mechanisms namely: survivor effect and a wanderer effect. A performance landscape based on NK model has been used to implement the simulation model.

To study inter-organizational alliance relationship, Aggarwal et al. (2011) use an agent-based simulation of inter-firm decision making. Simulation results point to complex interplay between interdependencies, governance structure, and firms' search capabilities. A firm's decision making process has been modeled using NK model. The performance landscape in modeled in 'N+1' dimensions, where N horizontal dimensions represent the space of all possible alternatives for each of the N policy choices, and one vertical dimension representing the performance level resulting from each overall choice configuration. Two major components of NK model implemented are: (i) mapping of choices to performance and (ii) generating and assessing alternative choice configuration.

Using ABM based on the idea of performance landscape, Siggelkow & Rivkin (2006) show that in multilevel organizations, increased exploration at lower levels can backfire, reducing overall exploration and diminishing performance in environments that require broad search.

NKC model, which is an extension of NK model, was developed to model co-evolution of species. It has been adapted in strategy research (Ganco & Agarwal, 2009). In NKC model, there are N elements of a decision vector of each firm. The parameter K measures the degree of interdependence or intra-firm coupling among the N elements of the decision vector. The parameter C specifies the extent to which individual firms' "sublandscapes" are tied together – i.e., inter-firm coupling.

NKC framework has been used by Ganco et al. (2009) to model an industry with differentiated products, where each firm occupies a certain exogenous niche. The model also incorporates inter-firm interaction i.e., each firm's choices have an impact on the payoffs of the choices of the other firms. Using NKC model, the authors have

investigated how entrant characteristics interact with environmental characteristics to explain differences in firm performance.

In another example, Siggelkow and Rivkin (2005) have used ABM to examine the effects of environmental turbulence, and complexity on formal design of organizations. The authors account for differences between simulation results and conventional wisdom due to powers of department heads to withhold information about departmental options, to control decision-making agendas, to veto firm-wide alternatives, and to take unilateral actions.

Porter et al. (2008) provide an introduction to complementarity framework and the NK-model for agent-based simulation studies, in the context of interactions among activities, and the consequences of these interactions on the creation and sustainability of competitive advantage. Noting that neither NK-model simulation approach nor the complementarity frameworks are suitable to study contextual interactions (i.e., interactions that are influenced by other activity choices made by a firm), future research directions for contextual interactions have been provided.

Genetic algorithms. Genetic algorithms focus on how rapidly and effectively a population of heterogeneous agents, represented as genes, adaptively learns. Adaptation occurs through a stochastic evolutionary process that includes: mutation, selection, crossover, and reproduction from one generation to next that eventually leads to gradual improvement. Eventually, only high-performing agents remain in the population. Thus can be used to examine the evolution of specific types of strategies within an agent.

In an example related to organizational research, Bruderer and Singh (1996) used genetic algorithm to examine organizational evolution within a population of

organizations. They found that learning accelerated the discovery of an effective organizational form.

Cellular automata. Cellular automata assume a system of agents that are spatially related. Spatial relatedness implies that the degree to which agents influence each other is dependent upon distance between them (Harrison et al., 2007). Agents behave according to some simple rules. Usually the rules that relate to spatial processes are uniform and deterministic. Some rules govern how neighbors affect an agent's behavior.

Lomi and Larsen (1996) used cellular automata to examine the tension between competition and legitimation process. It was observed how competitive and legitimating behavior among agents affected population density, founding rates and failure rates (macro level patterns).

The use of agent-based simulation based on cellular automata to study the participation of firms in online communities as a means to enhance demand for their products, has been discussed by Miller, Fabian, and Lin (2009). Using a simulation model, the authors demonstrate, how demand evolves as a function of interpersonal communication a firm's chosen strategy. They have also identified key variables affecting the diffusion of product preferences and assess the effectiveness under different conditions.

Reinforcement learning. Fang and Levinthal (2009) have studied the merits and disadvantages of exploitive behavior in the context of multi-stage decision making. To examine the trade-off between exploration and exploitation in multi-stage settings, a mechanism based on Q learning has been developed. Q learning explicitly models the evolution of an actors' existing representation of task environment by updating rules

through dynamic programming and reinforcing successful strategies over the unsuccessful ones. Fang et al. (2009) find that in a multistage problem; exploitation can lead to decline in both long-run and immediate decline in payoffs. Further, a decision policy that is mildly exploitive is superior to an explicit maximization of perceived payoffs.

Fang (2012) outline a theoretical model of organizational learning to account for empirical regularities based on credit assignment, where sequentially interdependent activities are termed as credit assignment problem. Using simulation and human subjects to validate credit assignment problem, Fang (2012) has provided a baseline model for the future development of an ABM consisting of heterogeneous agents that could have better fit with data.

Organizational simulation as research methodology. Harrison et al. (2007) and Davis et al. (2007) provide suggestions for developing simulation models for organization research. This section is based on these suggestions. Constructing simulation models involve identifying the underlying processes that govern the behavior of an agent (or actor) and formalizing them as a set of mathematical equations or computational rules. Transformation rules also need to be specified for determining the evolution of system over time. The resulting model embodies theoretical development and ideas. Hypotheses are normally not offered in simulation research; instead a model's consequences are determined computationally, which then lead to development of hypotheses or theoretical conclusions. Theoretical rigor introduced by formal modeling is considered one of the main strengths of simulation.

A simulation developer needs to consider the following five factors: (a) initial conditions, (b) time structure, (c) outcome determination, (d) number of simulation runs (iterations) and (e) sensitivity analysis with respect to different variables. These concepts for developing simulation models have been incorporated in the example of coin toss by Harrison and Carroll (1991).

Consider a problem of coin toss in which we want to get the probability of getting a first head and then a tail in two independent coin tosses. The processes of computational model are coin tosses. Parameter p can be defined as the probability of getting head (not necessarily fair). The faces of coin can be simulated through generating random numbers uniformly between 0 and 1. The initial condition does not need to be specified, as outcome depends on generating the random number p. The time structure is two periods. The run can be repeated many times with different random numbers to determine the percentage of heads and tails. Sensitivity analysis can be performed through changing the values of p.

Internal validity of computation model is established through verification. The verification of computational model could be done in the following ways:

Comparing simulation results with the propositions of the simple theory. If simulation confirms the propositions, then the theoretical logic and its computational representation are likely to be correct.

Computational findings should also be verified through robustness check (sensitivity analysis) and extreme value of constructs.

Mismatch between the propositions from simple theories and the simulation results can addressed by eliminating coding errors and shortcoming in the theoretical logic.

External validity of theory (i.e. generalizability and predictability) could be established through comparing simulation results with empirical data. Validation is especially necessary for non-empirical arguments (formal analytic modeling) and theory based on other scientific disciplines (Davis et al., 2007).

Challenges in organizational simulation research.

Model complexity: For realistic and elaborate models it often becomes problematic to determine what drives the results. From the view point of theory development, model should be presented as a simplified abstraction of the system – that retains the key elements of the relevant processes without unduly complicating the model (Harrison et al., 2007). Another approach can be to start with a simple model, and then elaborate it with adding complexity stepwise.

Model grounding: To incorporate the real-world behavior in simulation, models processes could be based on empirical work. Ungrounded parameters and issues often require thorough sensitivity analysis. When grounding is not possible, simulations can be used to explore consequences of theoretically derived processes.

Problems and limitations: Apart from common pitfalls seen in other research methods, presentation and details are one of main issues with simulation results. In the absence of sufficient details, it becomes difficult to develop any level of confidence on the findings. Bugs in computer programs can also produce spurious results. Translation of formal models in computer codes also poses a threat, in which different order of

execution of codes may produce different results. Replicating simulation results is often an issue, for example, independent attempts to replicate Garbage Can Model has produced mixed results (Harrison et al., 2007). As a major improvement, the findings of agent-based garbage can model has been presented by Fioretti and Lomi (2010) – eliminating several flaws from the original model. Generalizing simulation results is also considered problematic, and generalization beyond the range of simulation parameters should be treated as conjecture.

ABM offers new venues for management and organizational research and is also capable of generating non-linear behaviors of complex systems that are, often difficult to model through traditional research methodologies. Computational models can be treated as larger laboratories that allow management researchers to experiment possibilities (Burton & Obel, 2011).

Other Business Disciplines

In this section we discuss agent-based models in marketing and accounting applications. Given that there are recent surveys related to this, we keep this discussion relatively short, and refer readers to the relevant work.

Marketing phenomena such as product diffusion, is often categorized as complex processes that involves interaction among various agents e.g., consumers, sellers, distributors (Rand & Rust, 2011). Use of ABM to represent complex marketing phenomena, difficult to model through analytical or empirical approaches, is proposed by Rand and Rust (2011), who also provide a detailed review in marketing.

In the marketing literature ABM mostly appears in the study of diffusion of innovations and new product adoption (Rand & Rust, 2011). An excellent overview of

the use of ABM in innovation (or new product development) research has been provided by Garcia (2005). Here, the focus is on notable examples of ABM in the marketing research literature.

The effects of individual and network level, and negative word of mouth on a firm's profits have been explored using ABM by Goldenberg et al. (2007). The effect of negative word-of-mouth on the net present value (NPV) of a firm was found to be substantial, even when initial numbers of dissatisfied customers were relatively small. Weak ties of a given network help to spread harmful information through networks.

Goldenberg et al. (2009) have examined the role of hubs in diffusion and adoption. The dynamics of system is modeled through ABM where each agent is a potential adopter of innovation. Hubs are identified as people with large number of ties with other people. When the numbers of adopters in a neighborhood exceeded the threshold, an agent adopted the product. Stephen et al. (2010) have used ABM of online social networks to study information dissemination in online social networks.

Garber et al. (2004) have used Cellular Automata in presence of "small-world" network – a variant of ABM, as a predictive tool to assess the success of a new product shortly after launch time. Spatial divergence approach based on cross-entropy divergence measures have been used to determine the distance between simulated and real-life data of the adoption process. In another example, Goldenberg et al. (2002) have used Cellular Automata for generating and analyzing data, to investigate conditions under which a "saddle" occurs. Saddle in the context of sales is defined as an initial peak, then a trough of sufficient depth, and duration to exclude random fluctuations, and eventually the sales levels that exceed the initial peak.

When is it better to use ABM, and when should differential equations be used? This question has been addressed by Rahmandad and Sterman (2010) in the context of the diffusion of contagious disease. Examining the effects of individual heterogeneity on different network topologies, they conclude that differential equation and mean agent-based dynamics differ in several metrics.

Feng et al. (2012) use ABM to calculate aggregate diffusion dynamics for the adoption of new products without the mean-field approximation. Clusters-dynamics have been used to derive analytic approximation of the aggregate diffusion dynamics in multidimensional ABM. They conclude that the one-dimensional model and the Bass model provide a lower bound and an upper bound, respectively, for the aggregation of diffusion dynamics in ABM with any spatial structures.

Chou et al. (2007) have proposed an agent-based continuous auditing model (ABCAM). The basic premise of ABCAM has been that the various tasks performed by human auditors can be performed through software agents. The system uses mobile and intelligent agents to help human auditors perform various accounting tasks. In their context, mobility refers to the agents' (software object) ability to travel from one platform to another, and intelligence refers to the deployment of different degrees of artificial intelligence. The system is able to undertake automatic auditing in real time, and is easily adaptable to changes in auditing requirements.

Davis & Pesch (2012) have developed an ABM model to examine the emergent characteristics of fraud in organizations. Heterogeneous agents, with different motives and opportunity to commit fraud and pro-fraud attitude interact with each other that lead to attitude formation towards fraud. The model allows an evaluation of the relative

42

efficacy of mechanism designed to prevent fraud. The use of ABM provides insights into fraud even when data in organizations are censored.

The model consists of organizations, where agent representation of these organizations interact repeatedly (Davis et al., 2012). A model to investigate the impact of mechanisms to prevent or detect fraud is compared against a benchmark model in which all agents have the opportunity and motive to commit a fraud and pro-fraud attitude. Broadly two patterns emerge from the analysis of the benchmark model, depending upon how susceptible individual agents are to social influence. When average susceptibility is low, the number of fraudsters converges toward a specific level over time. When average susceptibility is moderate to high, the number of fraudsters vacillates over time between extremes: either almost everybody is fraudster or nobody.

Simulation Platforms

For the correct implementation of ABM, we need to choose an appropriate framework for its development. The implementation of a large scale ABM largely has been based on Object-Oriented (OO) programming. OO programming provides a framework to implement agents, their behavior and the environment surrounding them (North & Macal, 2007). The complete model is built of objects. The modular nature of OO programming provides flexibility in developing, maintaining and enhancing a complex model.

UML is often considered a standardized way of OO software design. Due to the graphical visualization, UML presents a higher level of abstraction than that of OO programming (Bersini, 2012), which is then easier to produce and communicate. The advantage of using UML diagrams in terms of: class, sequence, state and activity

diagrams has been discussed by Bersini (2012). The remaining part of this section presents a, discussion on modeling platforms that provide either or both of the aforementioned techniques (OO or UML) to build ABMs.

A comparison of various ABM tools have been presented in prior research, e.g., (Gatchell, 2008; Lytinen & Railsback, 2012; Nikolai & Madey, 2009). Also, a summary of different simulation tools can be accessed online: www2.econ.iastate.edu/tesfatsi/acecode.htm, www.openabm.org/page/modeling-platforms (Nikolai & Madey, 2009; Tobias & Hofmann, 2004). Some of the platforms mentioned in aforementioned sources do not have active support.

In the present work, instead of producing the comprehensive list of ABM platforms, major current development environments with active developer support have been listed (Table-2.1) and their salient features are discussed. Repast[4] is considered the most popular java based programming library for developing models (Nikolai et al., 2009; Tobias et al., 2004). Repast is developed by Social Science Research Computing, University of Chicago for social scientific use. It is supported in Java and Groovy programming languages. Repast also provides support to use other computation packages such as, MATLAB, R, JUNG etc. Repast HPC provides support for high performance distributed computing platforms. Another development framework, MASON, has similar features to Repast (Tobias et al., 2004).

NetLogo[5] is considered well documented and easy to learn tool for ABMs (Lytinen & Railsback, 2012). It is a popular tool among social scientists, those are new to

---

[4] http://repast.sourceforge.net
[5] http://ccl.northwestern.edu/netlogo

programming. NetLogo was developed at the Center for Connected Learning (CCL) at
Northwestern University and its development was influenced by its precursor StarLogo.

Table 2.1: ABM Development Platforms

Package	Language	Brief Summary	License
Repast	Java, Groovy	ABM toolkit with various built-in features	Open source
AnyLogic	Java	Supports system dynamics, discrete event and ABM. Widely used in modeling of manufacturing and logistics, business processes, human resources, consumer and patient behavior	Proprietary
NetLogo	Logo dialect	Multi-agent programmable modeling environment	Open source
Breve	Python	3D simulation environments for multi-agent systems and artificial life	Open source
Cougaar	Java	Especially designed for large-scale distributed agent-based applications	Open source
TNG	C++	TNG[6] is a framework for studying the formation and evolution of trade networks among strategically interacting agents	Open source
Altreva Adaptive Modeler	GUI interface	An application for creating agent-based market simulation based on evolutionary computing for stocks, forex currencies, exchange traded funds (ETFs)	Proprietary
Cormas	Smalltalk	Models concerned with the management of renewable resources, economic exchanges of agricultural products, and natural resources and land-use dynamics	Open source
Gambit	C++	Library of game theory software and tools for the construction and analysis of finite extensive and strategic games	Open source
JASA	Java	High-performance auction simulator using different auction mechanism	Open source
MATSim	Java	Java based platform to implement agent-based transport simulation	Open source

---

[6] http://www2.econ.iastate.edu/tesfatsi/tnghome.htm

The model execution time of NetLogo is considerably faster than ReLogo – a dialect of Logo, based on NetLogo. ReLogo is embedded in Eclipse development environment, and provides access to RePast libraries (Lytinen et al., 2012). Most of the platforms have specialized purposes. Their specialized support is discussed in the Table-2.1.

References

Adomavicius, G., Gupta, A., & Zhdanov, D. 2009. Designing intelligent software agents for auctions with limited information feedback. *Information Systems Research*, 20(4): 507.

Aggarwal, V. A., Siggelkow, N., & Singh, H. 2011. Governing collaborative activity: interdependence and the impact of coordination and exploration. *Strategic management journal*, 32(7): 705-730.

Avenali, A., and Bassanini, A. 2007. "Simulating combinatorial auctions with dominance requirement and loll bids through automated agents," *Decision Support Systems* (43:1), pp 211-228.

Axelrod, R. M. 1997. *The complexity of cooperation: Agent-based models of competition and collaboration*: Princeton Univ Pr.

Bapna, R., Goes, P., & Gupta, A. 2003. Replicating online Yankee auctions to analyze auctioneers' and bidders' strategies. *Information Systems Research*, 14(3): 244-268.

Baumann, O., & Stieglitz, N. 2013. Rewarding value-creating ideas in organizations: The power of low-powered incentives*. *Strategic management journal*.

Bersini, H. 2012. UML for ABM. *Journal of Artificial Societies and Social Simulation*, 15(1): 9.

Bianchi, C., Cirillo, P., Gallegati, M., & Vagliasindi, P. A. 2007. Validating and calibrating agent-based models: a case study. *Computational Economics*, 30(3): 245-264.

Bichler, M., Shabalin, P., & Pikovsky, A. 2009. A computational analysis of linear price iterative combinatorial auction formats. *Information Systems Research*, 20(1): 33-59.

Birukov, A., Blanzieri, E., and Giorgini, P. Year. "Implicit: An agent-based recommendation system for web search," *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, ACM2005, pp. 618-624.

Bonabeau, E. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 3): 7280.

Bonabeau, E. 2007. Understanding and managing complexity risk. *MIT Sloan management review*, 48(4): 62-68.

Bruderer, E., & Singh, J. V. 1996. Organizational evolution, learning, and selection: A genetic-algorithm-based model. *Academy of Management Journal*: 1322-1349.

Burton, R. M., & Obel, B. 2011. Computational modeling for what-is, what-might-be, and what-should-be studies—And triangulation. *Organization Science*, 22(5): 1195-1202.

Caldarelli, G., Marsili, M., & Zhang, Y.-C. 1997. A prototype model of stock exchange. *EPL (Europhysics Letters)*, 40(5): 479.

Cardinal, B. L., Turner, S. F., Fern, M. J., and Burton, M. R. 2011. "Organizing for Product Development Across Technological Environments: Performance Trade-offs and Priorities," *Organization Science* (22:4).

Chakrabarti, A. S., & Chakrabarti, B. K. 2009. Microeconomics of the ideal gas like market models. *Physica A: Statistical Mechanics and its Applications*, 388(19): 4151-4158.

Chakraborti, A., Toke, I. M., Patriarca, M., & Abergel, F. 2011. Econophysics review: II. Agent-based models. *Quantitative Finance*, 11(7): 1013-1041.

Challet, D., & Stinchcombe, R. 2001. Analyzing and modeling 1+ 1< i> d</i> markets. *Physica A: Statistical Mechanics and its Applications*, 300(1): 285-299.

Challet, D., & Zhang, Y.-C. 1997. Emergence of cooperation and organization in an evolutionary game. *Physica A: Statistical Mechanics and its Applications*, 246(3): 407-418.

Challet, D., & Zhang, Y.-C. 1998. On the minority game: Analytical and numerical studies. *Physica A: Statistical Mechanics and its Applications*, 256(3): 514-532.

Chang, R. M., Oh, W., Pinsonneault, A., & Kwon, D. 2010. A Network Perspective of Digital Competition in Online Advertising Industries: A Simulation-Based Approach. *Information Systems Research*, 21(3): 571-593.

Chari, K. 2000. Intelligent Agents: Overview, Applications and Research Directions. *INFORMS Computing Society Newsletter* 21(2).

Chari, K., and Agrawal, M. 2007. "Multi-issue automated negotiations using agents," *INFORMS Journal on Computing* (19:4), pp 588-595.

Chatterjee, A., K Chakrabarti, B., & Manna, S. 2004. Pareto law in a kinetic model of market with random saving propensity. *Physica A: Statistical Mechanics and its Applications*, 335(1): 155-163.

Chen, S.-H. 2012. Varieties of agents in agent-based computational economics: A historical and an interdisciplinary perspective. *Journal of Economic Dynamics and Control*, 36(1): 1-25.

Chiarella, C., & Iori, G. 2002. A simulation analysis of the microstructure of double auction markets. *Quantitative Finance*, 2(5): 346-353.

Chou, C. L.-y., Du, T., and Lai, V. S. 2007. "Continuous auditing with a multi-agent system," *Decision Support Systems* (42:4), pp 2274-2292.

Coen, C. A., and Maritan, C. A. 2011. "Investing in capabilities: The dynamics of resource allocation," *Organization Science* (22:1), pp 99-117.

Cohen, M. D., March, J. G., & Olsen, J. P. 1972. A garbage can model of organizational choice. *Administrative science quarterly*: 1-25.

Cont, R. 2007. Volatility clustering in financial markets: empirical facts and agent-based models, *Long memory in economics*: 289-309: Springer.

Cont, R., & Bouchaud, J.-P. 2000. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(2): 170-196.

Cristelli, M., Pietronero, L., & Zaccaria, A. 2011. Critical overview of agent-based models for economics. *arXiv preprint arXiv:1101.1847*.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. 2007. Developing Theory Through Simulation Methods. *Academy of Management Review*, 32(2): 480-499.

Davis, J. S., & Pesch, H. L. 2012. Fraud dynamics and controls in organizations. *Accounting, Organizations and Society*.

Du, T. C., Li, E. Y., and Wei, E. 2005. "Mobile agents for a brokering service in the electronic marketplace," *Decision Support Systems* (39:3), pp 371-383.

The Economist, 2010. Agents of Change. *http://www.economist.com/node/16636121*.

Fagiolo, G., Moneta, A., & Windrum, P. 2007. A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems. *Comput. Econ.*, 30(3): 195-226.

Fang, C., and Levinthal, D. 2009. "Near-term liability of exploitation: exploration and exploitation in multistage problems," *Organization Science* (20:3), pp 538-551.

Fang, C., Lee, J., & Schilling, M. A. 2010. Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science*, 21(3): 625-642.

Fang, C. 2012. "Organizational learning as credit assignment: A model and two experiments," *Organization Science* (23:6), pp 1717-1732.

Farmer, J. D., & Foley, D. 2009. The economy needs agent-based modelling. *Nature*, 460(7256): 685-686.

Fasli, M., and Michalakopoulos, M. 2008. "e-Game: A platform for developing auction-based market simulations," *Decision Support Systems* (44:2), pp 469-481.

Feng, L., Li, B., Podobnik, B., Preis, T., & Stanley, H. E. 2012. Linking agent-based models and stochastic models of financial markets. *Proceedings of the National Academy of Sciences*, 109(22): 8388-8393.

Fioretti, G., & Lomi, A. 2010. Passing the buck in the garbage can model of organizational choice. *Computational and Mathematical Organization Theory*, 16(2): 113-143.

Fioretti, G., & Lomi, A. 2011. Agent-based simulation models in organization science. *Available at SSRN 1874885*.

Gaines, D. A., & Pakath, R. 2013. An Examination of Evolved Behavior in Two Reinforcement Learning Systems. *Decision Support Systems*.

Ganco, M., & Agarwal, R. 2009. Performance differentials between diversifying entrants and entrepreneurial start-ups: A complexity approach. *Academy of Management Review*, 34(2): 228-252.

Garber, T., Goldenberg, J., Libai, B., & Muller, E. 2004. From density to destiny: Using spatial dimension of sales data for early prediction of new product success. *Marketing Science*, 23(3): 419-428.

Garcia, R. 2005. Uses of Agent-Based Modeling in Innovation/New Product Development Research*. *Journal of Product Innovation Management*, 22(5): 380-398.

Gatchell, A. 2008. Agent-Based Modeling. *Working Paper*.

Gatti, D. D., Gallegati, M., Greenwald, B. C., Russo, A., and Stiglitz, J. E. 2009. "Business fluctuations and bankruptcy avalanches in an evolving network economy," *Journal of economic interaction and coordination* (4:2), pp 195-212.

Gavetti, G., & Levinthal, D. 2000. Looking forward and looking backward: Cognitive and experiential search. *Administrative science quarterly*, 45(1): 113-137.

Gavetti, G., Levinthal, D. A., & Rivkin, J. W. 2005. Strategy making in novel and complex worlds: the power of analogy. *Strategic management journal*, 26(8): 691-712.

Geanakoplos, J., Axtell, R., Farmer, D. J., Howitt, P., Conlee, B., Goldstein, J., Hendrey, M., Palmer, N. M., & Yang, C.-Y. 2012. Getting at Systemic Risk via an Agent-Based Model of the Housing Market. *American Economic Review*, 102(3): 53-58.

Giardina, I., and Bouchaud, J.-P. 2003. "Bubbles, crashes and intermittency in agent based market models," *The European Physical Journal B-Condensed Matter and Complex Systems* (31:3), pp 421-437.

Gilbert, N., Pyka, A., & Ahrweiler, P. 2001. Innovation networks-a simulation approach. *Journal of Artificial Societies and Social Simulation*, 4(3): 1-13.

Gilli, M., & Winker, P. 2003. A global optimization heuristic for estimating agent based models. *Computational Statistics & Data Analysis*, 42(3): 299-312.

Gode, D. K., & Sunder, S. 1993. Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality. *Journal of Political Economy*, 101(1): 119-137.

Goldenberg, J., Han, S., Lehmann, D., & Hong, J. 2009. The role of hubs in the adoption processes. *Journal of Marketing*, 73(2).

Goldenberg, J., Libai, B., Moldovan, S., & Muller, E. 2007. The NPV of bad news. *International Journal of Research in Marketing*, 24(3): 186-200.

Goldenberg, J., Libai, B., & Muller, E. 2002. Riding the saddle: How cross-market communications can create a major slump in sales. *The Journal of Marketing*: 1-16.

Greenwald, A., Kannan, K., & Krishnan, R. 2010. On evaluating information revelation policies in procurement auctions: A Markov decision process approach. *Information Systems Research*, 21(1): 15-36.

Gregg, D. G., and Walczak, S. 2006. "Auction Advisor: an agent-based online-auction decision support system," *Decision Support Systems* (41:2), pp 449-471.

Guo, W., Jank, W., & Rand, W. 2011. Estimating functional agent-based models: an application to bid shading in online markets format. *GECCO 2011, Dublin, Ireland, July 12-16, 2011*.

Haber, G. 2002. Monetary and fiscal policy analysis with an agent-based macroeconomic model. *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)*, 228(2+ 3): 276-295.

Harrison, J. R., & Carroll, G. R. 1991. Keeping the faith: A model of cultural transmission in formal organizations. *Administrative science quarterly*: 552-582.

Harrison, J. R., Carroll, G. R., & Carley, K. M. 2007. Simulation modeling in organizational and management research. *Academy of Management Review*, 32(4): 1229-1245.

Iori, G., & Porter, J. 2012. *Agent-Based Modelling for Financial Markets*.

Jones, J. L., Easley, R. F., and Koehler, G. J. 2006. "Market segmentation within consolidated e-markets: A generalized combinatorial auction approach," *Journal of Management Information Systems* (23:1), pp 161-182.

Kim, G.-r., and Markowitz, H. M. 1989. "Investment rules, margin, and market volatility," *The Journal of Portfolio Management* (16:1), pp 45-52.

Kim, H. S., & Cho, J. H. 2010. Supply chain formation using agent negotiation. *Decision Support Systems*, 49(1): 77-90.

Kim, W.-S. 2009. "Effects of a trust mechanism on complex adaptive supply networks: An agent-based social simulation study," *Journal of Artificial Societies and Social Simulation* (12:3), p 4.

Ladley, D. 2012. Zero intelligence in economics and finance. *Knowledge Engineering Review*, 27(2): 273.

Lau, R. Y., Li, Y., Song, D., and Kwok, R. C. W. 2008. "Knowledge discovery for adaptive negotiation agents in e-marketplaces," *Decision Support Systems* (45:2), pp 310-323.

Lazer, D., & Friedman, A. 2007. The network structure of exploration and exploitation. *Administrative science quarterly*, 52(4): 667-694.

LeBaron, B. 2006. Agent-based computational finance. *Handbook of computational economics*, 2: 1187-1233.

LeBaron, B., & Yamamoto, R. 2007. Long-memory in an order-driven market. *Physica A: Statistical Mechanics and its Applications*, 383(1): 85-89.

Levinthal, D., & Posen, H. E. 2007. Myopia of selection: Does organizational adaptation limit the efficacy of population selection? *Administrative science quarterly*, 52(4): 586-620.

Liang, W.-Y., & Huang, C.-C. 2006. Agent-based demand forecast in multi-echelon supply chain. *Decision Support Systems*, 42(1): 390-407.

LiCalzi, M., & Pellizzari, P. 2003. Fundamentalists clashing over the book: a study of order-driven stock markets. *Quantitative Finance*, 3(6): 470-480.

Lin, R., Gal, Y. a. K., Kraus, S., and Mazliah, Y. 2013. "Training with Automated Agents Improves People's Behavior in Negotiation and Coordination Tasks," *Decision Support Systems*).

Lo, A. 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management, 30, 15-29*.

Lomi, A., & Larsen, E. R. 1996. Interacting locally and evolving globally: A computational approach to the dynamics of organizational populations. *Academy of Management Journal*: 1287-1321.

Lovric, M., Li, T., & Vervest, P. 2012. Sustainable revenue management: A smart card enabled agent-based modeling approach. *Decision Support Systems*, 2012: 1-15.

Lux, T. 2012. Estimation of an agent-based model of investor sentiment formation in financial markets. *Journal of Economic Dynamics and Control*.

Lux, T., & Marchesi, M. 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature*, 397(6719): 498-500.

Lytinen, S. L., & Railsback, S. F. 2012. *The Evolution of Agent-based Simulation Platforms: A Review of NetLogo 5.0 and ReLogo*. Paper presented at the Proceedings of the Fourth International Symposium on Agent-Based Modeling and Simulation.

Mannaro, K., Marchesi, M., & Setzu, A. 2008. Using an artificial financial market for assessing the impact of Tobin-like transaction taxes. *Journal of Economic Behavior & Organization*, 67(2): 445-462.

March, J. G. 1991. Exploration and exploitation in organizational learning. *Organization Science*, 2(1): 71-87.

Mehta, K., & Bhattacharyya, S. 2006. Design, development and validation of an agent-based model of electronic auctions. *Information Technology and Management*, 7(3): 191-212.

Mike, S., & Farmer, J. D. 2008. An empirical behavioral model of liquidity and volatility. *Journal of Economic Dynamics and Control*, 32(1): 200-234.

Miller, K. D., Fabian, F., & Lin, S. J. 2009. Strategies for online communities. *Strategic management journal*, 30(3): 305-322.

Nan, N. 2011. Capturing bottom-up information technology use processes: a complex adaptive systems model. *MIS Quarterly*, 35(2): 505-532.

Nikolai, C., & Madey, G. 2009. Tools of the trade: A survey of various agent based modeling platforms. *Journal of Artificial Societies and Social Simulation*, 12(2): 2.

North, M. J., & Macal, C. M. 2007. *Managing business complexity: discovering strategic solutions with agent-based modeling and simulation*: Oxford university press, USA.

Nunamaker, J. F., DErrICk, D. C., Elkins, A. C., Burgoon, J. K., and Patton, M. W. 2011. "Embodied Conversational Agent-Based Kiosk for Automated Interviewing," *Journal of Management Information Systems* (28:1), pp 17-48.

Oh, W., & Jeon, S. 2007. Membership herding and network stability in the open source community: The Ising perspective. *Management Science*, 53(7): 1086-1101.

Poledna, S. 2011. *Agent-based models in econophysics*. uniwien.

Porter, M., and Siggelkow, N. 2008. "Contextuality within activity systems and sustainability of competitive advantage," *The Academy of Management Perspectives* (22:2), pp 34-56.

Raghu, T., Jayaraman, B., & Rao, H. 2004. Toward an integration of agent-and activity-centric approaches in organizational process modeling: Incorporating incentive mechanisms. *Information Systems Research*, 15(4): 316-335.

Rahmandad, H., & Sterman, J. 2010. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management Science*, 54(5): 998.

Rand, W., & Rust, R. T. 2011. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, 28(3): 181-193.

Rao, H. R., Chaudhury, A., & Chakka, M. 1995. Modeling Team Processes: Issues and a Specific Example. *Information Systems Research*, 6(3): 255-285.

Ren, F., and Zhang, M. 2013. "A Singe Issue Negotiation Model for Agents Bargaining in Dynamic Electronic Markets," *Decision Support Systems (To appear)*.

Rivkin, J. W. 2000. Imitation of complex strategies. *Management Science*, 46(6): 824-844.

Rivkin, J. W. 2001. Reproducing Knowledge: Replication Without Imitation at Moderate Complexity. *Organization Science*, 12(3): 274-293.

Rivkin, J. W., & Siggelkow, N. 2003. Balancing search and stability: Interdependencies among elements of organizational design. *Management Science*, 49(3): 290-311.

Rudolph, J. W., & Repenning, N. P. 2002. Disaster dynamics: Understanding the role of quantity in organizational collapse. *Administrative science quarterly*: 1-30.

Siggelkow, N., & Levinthal, D. A. 2003. Temporarily divide to conquer: Centralized, decentralized, and reintegrated organizational approaches to exploration and adaptation. *Organization Science*, 14(6): 650-669.

Siggelkow, N., & Rivkin, J. W. 2005. Speed and search: Designing organizations for turbulence and complexity. *Organization Science*, 16(2): 101-122.

Siggelkow, N., & Rivkin, J. W. 2006. When exploration backfires: Unintended consequences of multilevel organizational search. *Academy of Management Journal*, 49(4): 779-795.

Siggelkow, N., & Rivkin, J. W. 2009. Hiding the evidence of valid theories: How coupled search processes obscure performance differences among organizations. *Administrative science quarterly*, 54(4): 602-634.

Stephen, A., Dover, Y., & Goldenberg, J. 2010. A comparison of the effects of transmitter activity and connectivity on the diffusion of information over online social networks.

Strader, T. J., Lin, F.-R., and Shaw, M. J. 1998. "Simulation of order fulfillment in divergent assembly supply chains," *Journal of Artificial Societies and Social Simulation* (1:2), pp 36-37.

Sueyoshi, T., and Tadiparthi, G. R. 2008. "An agent-based decision support system for wholesale electricity market," *Decision Support Systems* (44:2), pp 425-446.

Sysi-Aho, M., Chakraborti, A., & Kaski, K. 2004. Searching for good strategies in adaptive minority games. *Physical Review E*, 69(3): 036125.

Tedeschi, G., Iori, G., & Gallegati, M. 2012. Herding effects in order driven markets: The rise and fall of gurus. *Journal of Economic Behavior & Organization*, 81(1): 82-96.

Tesfatsion, L., & Judd, K. L. 2006. *Handbook of computational economics, Volume 2: Agent-based computational economics*: North-Holland.

Thurner, S., Farmer, J. D., & Geanakoplos, J. 2012. Leverage causes fat tails and clustered volatility. *Quantitative Finance*, 12(5): 695-707.

Tobias, R., & Hofmann, C. 2004. Evaluation of free Java-libraries for social-scientific agent based simulation. *Journal of Artificial Societies and Social Simulation*, 7(1).

Van Der Zee, D.-J., Holkenborg, B., & Robinson, S. 2012. Conceptual modeling for simulation-based serious gaming. *Decision Support Systems*.

Valluri, A., and Croson, D. C. 2005. "Agent learning in supplier selection models," *Decision Support Systems* (39:2), pp 219-240.

Van Der Zee, D.-J., Holkenborg, B., and Robinson, S. 2012. "Conceptual modeling for simulation-based serious gaming," *Decision Support Systems*).

Vriend, N. J. 2000. An illustration of the essential difference between individual and social learning, and its consequences for computational analyses. *Journal of Economic Dynamics and Control*, 24(1): 1-19.

Wainer, J., Ferreira, P. R., and Constantino, E. R. 2007. "Scheduling meetings through multi-agent negotiations," *Decision Support Systems* (44:1), pp 285-297.

Wang, J., Gwebu, K., Shanker, M., and Troutt, M. D. 2009. "An application of agent-based simulation to knowledge sharing," *Decision Support Systems* (46:2), pp 532-541.

Wang, T.-W., and Tadisina, S. K. 2007. "Simulating Internet-based collaboration: A cost-benefit case study using a multi-agent model," *Decision Support Systems* (43:2), pp 645-662.

Yang, C. C., Yen, J., and Chen, H. 2000. "Intelligent internet searching agent based on hybrid simulated annealing," *Decision Support Systems* (28:3), pp 269-277.

Zaffar, M. A., Kumar, R. L., & Zhao, K. 2011. Diffusion dynamics of open source software: An agent-based computational economics (ACE) approach. *Decision Support Systems*, 51(3): 597-608.

Zott, C. 2003. Dynamic capabilities and the emergence of intraindustry differential firm performance: insights from a simulation study. *Strategic management journal*, 24(2): 97-125.

Chapter 3 : Count Amplification and Manipulation

Resistance in Top-N News Recommender

Introduction

Historically, mass media has played an important role in creating and sustaining mass opinion and behavior in society on issues ranging from policy, violence, new product adoption, family and health related issues (Myers 2000, Rogers 1976). Traditionally, editorial perspectives have driven the decisions of what news to present to readers, and media editors have therefore been in positions to form and shape opinion. However, that trend is changing with technology-driven decisions that are being used instead, or in conjunction.

In the last ten years, the Web has grown to become the primary news source for many users. At the same time there has been bigger penetration of social media such as tweets, Facebook posts and online videos (Economist-b 2011). The Economist has noted that this change in news consumption behavior has "*turned the news industry upside down, making it more participatory, social, diverse and partisan*" (Economist-b 2011). Readers often volunteer to submit, share and comment on news articles. Referrals from social networks are the fastest growing source of traffic for some news websites and, as The Economist writes, "*the most popular stories cause a flood of traffic as recommendations ripple across social networks*" (Economist-b 2011). Hence, once an article makes it into a "most popular" list, there can be a self-reinforcing effect that can

further impact its ultimate readership or influence. Figure 3.1 presents some variants of most popular list displayed by popular media sites.

Figure 3.1. Variants of "most-popular" recommender

The focus of this research is to investigate the phenomena emerging through reader interaction with News Recommendation Systems (NRS hereafter), and to address the issue of manipulation in NRS. While there is little work that has addressed the issue of recommender manipulation for news, this topic is important because significant public opinion in the society is known to be influenced by user exposure to news. For example, Phillips (1974) studied the effect of publicity given to suicide stories and found that there was an immediate increase in suicide cases after such news was publicized.

To distort opinion, recommender systems are an easy target for manipulators. For example, Lerman (2007-a) describes a Digg controversy in which a user posted an analysis proving that the top 30 users of Digg were responsible for a disproportionate fraction of the front page. The allegation was that the top users conspired to promote their own articles at the expense of other articles, leading to such an increased concentration. In response, Digg modified the algorithm to devalue votes from friends. Also, there are

special group of online users known to be in existence, such as the "*Internet Water Army*" (Chen, et al. 2011) who get paid for posting comments, threads and news articles. These groups are known to "flood" the internet with purposeful comments and articles. Chen, et al. (2011) discuss techniques to identify such manipulators from behavioral and semantic data. The implication is that by identifying and removing such users, manipulation might decrease.

The susceptibility of most popular lists towards manipulation has been demonstrated by Weber (December 19, 2010) – Managing Editor of NewsWeek. To demonstrate that these systems can be easily gamed, he used a group of people to place a relatively old science story in the most emailed list.

These examples highlight the context of our research agenda. Further, NRS in comparison to other recommender systems, operate in a fundamentally different environment due to a constant stream of news. Such an environment places a greater need for effective recommender systems, yet suffers from potentially easier manipulation due to several factors such as the greater use of implicit feedback mechanisms where clicks are counted as votes, sparseness in various topic categories and incentive mechanisms currently in place that encourage greater clicks for higher advertising revenue.

In this research, we study two very different selection mechanisms and discuss the trade-off between them. One of selection mechanisms, called "most popular" (or Top-N) list is widely used in current practice. We note that the term Top-N is also used in the context of personalized recommender systems (Deshpande and Karypis 2004). But, in the present research, we use it to refer to the "most popular" news recommender and its

variants such as most e-mailed or most viewed. The other selection mechanism, probabilistic selection, has been introduced in this research.

Our findings have been presented in two ways. First using simulation, we show that the most popular recommender is prone to artificially amplifying small differences. The $(N+1)^{th}$ article, which may have "just" missed the cutoff, is often unduly penalized in terms of readership counts in the long run. The probabilistic variant is shown instead to be robust. This weakness of the most popular recommender can be exploited by manipulators who seek to gain popularity for their articles. In this context we also show that the probabilistic mechanism is again more robust. Building on statistical results on classical urn models, namely Pólya's and Bernard Friedman's urn models, (Freedman 1965) we derive some theoretical insights for special cases. The trade-off between the Top-N NRS and the proposed probabilistic variant is discussed in terms of count distortion and information quality[7]. Whereas we do observe some loss of information quality in probabilistic NRS, it is highly robust towards minimizing artificial amplification in the counts of the recommended articles in comparison with the Top-N NRS. We present results on manipulation for the probabilistic NRS in comparison with an "adapted" *influence limiter* heuristic (Resnick and Sami 2007). We have also discussed our key findings in a more realistic setting, with data collected from five different local news websites. Finally, an extension of probabilistic selection has been introduced and we demonstrate that this extended model can be used to address an interesting issue of social desirability between the Top-N and probabilistic selection mechanisms. To our knowledge these are all unique contributions of this research.

---

[7] Counts of articles has been assumed as the surrogate measure of quality

Related Work

In one of the earliest research in online manipulation, Dellarocas (2006) presented theoretical analysis of manipulation strategies and its impact on the firm and consumer assuming that the main source of quality information for consumers is an online product review forum. This work has established various results on effects of online forum manipulation in a simple monopoly setting. The analysis of results shows the existence of a setting where forum manipulation is equivalent to a form of quality signaling that benefits consumers. Also, if consumers expect that firms will manipulate, as the volume and quality of user-generated online content increases, then there will be a certain threshold beyond which firms will have to engage in profit-reducing online manipulation practices. The findings from closed-form solutions have been also generalized in a wide range of multi-firm settings and for a broad class of consumer utilities, firm payoff functions and signal distributions. Finally, the author has proposed an idea of filtering technologies that make it costlier for firms to manipulate. We take a similar approach to study NRS through simulation and develop analytical results.

Manipulation resistant recommender systems discussed in the literature is also related to our work (Resnick and Sami 2007, Resnick and Sami 2008). Resnick and Sami (2007) introduced the *Influence Limiter* algorithm for items recommendation, controlling rater's influence on recommender systems through reputation acquired over time. The authors show that the optimal strategy of a rater is to induce predictions that accurately reveal the rater's information about the item. Using an information-theoretic measure, the authors establish that the negative impact of any rater is bounded by a given limit. In their subsequent work Resnick and Sami (2008) establish the tradeoffs between

resistance to manipulation by an attacker and the optimal use of genuine ratings in recommender systems. A lower bound on how much information must be discarded is also provided. Lee and Zhu (2012) have studied *shilling attacks* detection on recommender systems and have proposed a two phase procedure. First, a multidimensional scaling has been used to identify distinct behavior and to narrow down the detection space by filtering out noise profiles. In the second phase, a clustering based method has been used to discriminate the attackers.

Van Roy and Yan (2010) have studied linear collaborative filtering (CF) algorithms and have shown it to be robust in comparison to nearest neighbor algorithms widely used in commercial systems. This analysis of linear CF algorithms shows that as a user rates an increasing number of products, the average accuracy becomes insensitive to manipulated data. The authors have established bounds on distortion as a function of percentage of manipulated data and number of products rated by a user whose future rating will be predicted.

In particular for NRS, Largillier, et al. (2010) have discussed a robust voting system for social news websites based on SpotRank. Considering voting as a recommendation, Lergillier et al. present a set of heuristics that demotes the effects of manipulation. SpotRank is built over $ad - hoc$ statistical filters, a collusion detection mechanism and also the reputation of users and proposed news. In their work, they discuss several issues of social NRS, such as the existence of cabals (collusion of large group of users that vote for each other), those who try to manipulate the system using daily mailing lists, some users posting many links to flood the system, and using several

61

IP addresses to vote for themselves. Lerman (2007-b) has discussed a model for the news aggregation process by Digg for news recommendation and ratings.

In the context of social influence Salganik, et al. (2006) found that the presence of social influence leads to greater inequality and unpredictability in the popularity of songs. In a broader context, the issue of popularity has been addressed by Easley and Kleinberg (2010), in which they have argued that the power law seems to dominate in cases where quantity being measured can be viewed as any kind of popularity.

Model

We present the main findings of our study using the approach of a thought experiment implemented as a simulation. This has been a powerful tool to address various issues related with social sciences and public policy (Maroulis, et al. 2010, Schelling 1971). For instance, using a thought experiment, Schelling (1971), showed that a small preference for one's neighbors to be of the same color could lead to total segregation of society, and using a similar methodology Maroulis, et al. (2010) studied the survival of public schools based on individual choices.

Model description. We set up the simulation model as follows. We maintain a Comprehensive List (*CL*) of articles and their corresponding counts (or clicks). From CL, N articles are selected for display as "recommendations". Before the simulation starts articles are assigned random counts in some range (e.g. between 0 and 1000). Articles are sorted in decreasing order of their counts, and the articles with high counts are selected for the Display List (*DL*). Further, the $(N + 1)^{th}$ article was deliberately assigned a count of exactly one less than the count of Nth article.

The selection of articles in the DL is updated at a pre-selected time step, and this selection of articles is based on two different selection processes namely, the $Top - N$ and *probabilistic selection*. The Top-N selection is a "hard cutoff", which selects N articles for display corresponding to the highest counts. This is typically how most online news sites display the most popular or viewed articles, typically in a prominent box or sidebar. Probabilistic selection on the other hand, is a mechanism proposed here, where articles are selected probabilistically based on their counts thus far. In this mechanism, every article in CL will have some probability, based on its count, to appear in DL.

Probabilistic selection of articles is based on probabilistic sampling without replacement for N articles. The probability that an article will be selected in DL is given by $prob(a) = \frac{count_a}{\sum_j count_j}$, where $count_a$ represents the count of an article 'a' at a given time step and $\sum_j count_j$ represents the total counts of articles not yet selected for DL. This sampling process is repeated N times to generate the N recommendations in DL. Pseudo code for the implementation of these selection processes is discussed later in this section.

Two different reader models were also implemented. In both models, a user is assumed to select an article either from DL with some probability p or from the remaining list $RL (= CL - DL)$ with probability $1 - p$. In the first model, a reader selects an article from DL randomly. Whereas, in the second model, the top-most article in the DL has the highest probability of being selected, and the bottom-most has the lowest probability, with a linear decrease in the selection probability between top-most and bottom-most articles.

For the second reader model, the probability of a particular article with rank $i, i \in \{1,2,..N\}$ in DL being read (selected) is given by $r_i = \frac{N+1-i}{\sum_{i=1}^{N} i}$. Here, we define rank as the order in which articles are displayed in the recommended list. For ease of exposition, the present model intentionally leaves out other complicated factors of news arrival and reader behavior based on front-page display of news websites. However, in the sensitivity analysis section we have presented findings based on some real world distribution of article counts (and reader behavior that gives rise to the power law distribution in popularity).

Implementation of NRS. Pseudo code for the simulation and *probabilistic selection* is presented below. ("*Select*" can be count-based or probabilistic; while "*Choose*" can be based on either of the two reader models described above).

For each reader
    Sort the updated count and **select** N articles for DL
    If selected article is from DL(i.e with probability p)
       Choose an article from DL and increase its count by 1
    Else
       Randomly choose an article from RL; $(RL = CL - DL)$ and increase its count
    by 1
End for.

Probabilistic selection.

1. The count of articles are $c[1], c[2], ... c[n]$
2. $count[1] = 0$
3. for x = 2 to $n + 1$

$$count[x] = count[x - 1] + c[x - 1]$$

4. end for
5. for y = 1 to N
   a. generate a random integer (R) between 0 and $count[n + 1]$
   b. determine the indices between which R lies, as $(i, i + 1)$
   c. select article corresponding to the count $c[i]$ for DL
   d. j ← $c[i]$
   e. While $(i < k \leq n)$

   $$count[k] = count[k + 1] - j$$

   f. end while
   g. n=n-1

   end for

Measures. In order to compare different user models and selection mechanisms we introduce two specific measures here. Both of these measures are based on the counts of Nth and $(N + 1)^{th}$ articles over the complete simulation. Both N^{th} and $(N + 1)^{th}$ articles selected here are based on the initial counts of articles before the simulation starts.

Measure M1. This is defined as the logarithmic-ratio of the counts of Nth and $(N + 1)^{th}$ articles at each time-step as follows:

$$M1(i) = \ln(count_{Ni}) - \ln(count_{(N+1)i}) = \ln\frac{count_{Ni}}{count_{(N+1)i}} \text{ at the } i^{th} \text{ iteration of}$$

the simulation. This measures the relative change in counts of Nth and $(N + 1)^{th}$ article, hence count amplification between the articles. This measure has been chosen to demonstrate the fact that even if N^{th} article makes it into DL by a hair, in the count based selection, in long run, it will have significantly higher popularity than the $(N + 1)^{th}$

article simply by virtue of being in such a prominent list. We also use this measure to demonstrate how a manipulator can exploit the self-reinforcing nature of top-N lists. At the start of the simulation $count(N) \sim count(N+1)$, hence $M1(0) \sim 0$.

Measure M2. This is defined as the count (hits) of the jth article divided by the total number of count (hits) at a given time. We denote it as M2 and at the ith iteration it will be $M2(i) = \frac{count_{ji}}{\sum_{k=1}^{n} count_{ki}}$. It represents the share of the counts for any particular article j in the NRS, over iterations and can be understood as a success measure of an article in a given selection mechanism. Other things being equal, articles with higher market shares can be considered more "successful" than others.

Update rule. At each time period the model proceeds as follows. One reader arrives at each time step. Upon arrival reader selects probabilistically to read an article either from displayed list (DL) or the remaining list (RL) of articles. The probability of selection of an article either from DL or RL is controlled in the simulation. If a reader selects an article from RL, then random selection of an article is performed. The count of the selected article is increased by 1.

If a reader selects an article from DL then random selection of an article is performed for Reader Model 1 and selection of an article is performed according to probability r_i for the Reader Model 2. The count of the selected article is increased by 1.

For the two different NRS, count-based and probabilistic, the selection of N articles is made for $DL,$ and DL is updated at each time step.

Manipulation. To study manipulation, we assume that a manipulator can create artificial clicks to raise the counts of a selected article (such as by creating fake IDs for instance). These fake counts are randomly distributed over the given interval. These fake

counts are created by malicious readers who upon arrival increase the count of a particular article by 1. The particular article selected for manipulation in the present model is the $(N + 1)^{th}$, since this is the article that would have just missed the hard "top $- N$" cutoff. Also, we study two types of manipulation – early and uniform – to examine what impact each might have. In "early" manipulation, the fake clicks are assumed to be distributed in some early part of the time period; in "uniform" manipulation the fake clicks are uniformly distributed over the entire time interval. We also examine the extent of manipulation (high and low, based on how many fake counts are generated) and the impact it can have.

Simulation Results

The analyses of our results are based on two sections: (1) without manipulation and (2) with manipulation. The simulation results for "without manipulation" explain the phenomenon that emerges using different NRS based on different selection mechanisms. In, particular we compare the two measures M1 and M2 for N^{th} and $(N + 1)^{th}$ articles and discuss findings based on them. Manipulation has been introduced to demonstrate the susceptibility of the Top-N NRS and the robustness of the proposed probabilistic NRS as an alternative. Manipulation has been introduced in two stages to study the effects of early manipulation and manipulation over large interval of time. In the first case the manipulated counts are distributed uniformly between 0 and 100 and in the second case manipulated counts are distributed uniformly between 0 and 1500. We consider different scenarios based on (a) the reader models (two), (b) the existence of manipulation (two) and (c) the selection mechanism (two – count-based and probabilistic) as described in the tree in Figure 3.2. The leaves of the tree correspond to specific simulation scenarios. As

Figure 3.2 shows, there are 12 leaves for some specific choice of global simulation parameters.

Table 3.1.The model parameters used in the simulation

Parameter	Value
No. of Readers	1500
No. of articles in DL	10
No. of articles in CL	200
Initial counts of articles[8]	Random Integer between 0 and1000
Manipulation Counts	10 and 50
Probability of selection of an article from DL (p)	0.9, 0.5, 0.25, 0.1
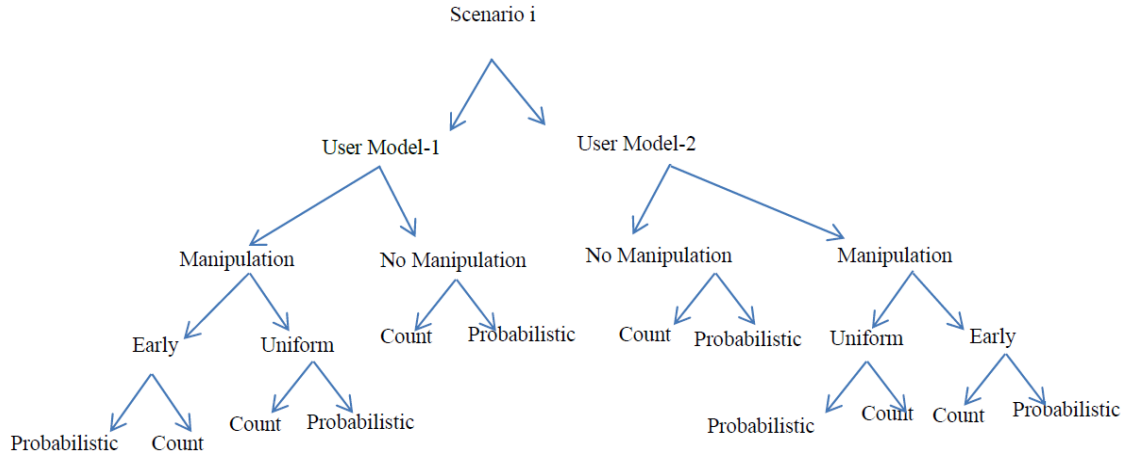
Figure 3.2. Graph for specific selection of global parameters

Two of the global simulation parameters are (1) probability of a reader selecting an article from DL instead of from RL (varied as 0.9, 0.5, 0.25, 0.1), and (2) the extent of manipulation (high or low, implemented in the simulation as manipulated counts). The

---

[8] Except N^{th} and $(N + 1)^{th}$ articles. Counts for these articles were assigned such that $count(N + 1) = count(N) - 1$. This was done deliberately to test how the hard cutoff treats very small initial differences in quality between articles.

value of simulation parameters used is listed in Table 3.1. For any specific choice of these two parameters we have 12 graphs in the results (corresponding to the 12 leaves of the tree).

Table 3.2. Abbreviations used in the figures

Abbreviation	Definition
M2_d	M2 for the Nth article in Top-N NRS
M2_u	M2 for the $(N + 1)^{th}$ article in Top-N NRS
p_M2_d	M2 for the Nth article in probabilistic NRS
p_M2_u	M2 for the $(N + 1)^{th}$ article in probabilistic NRS
M1_count	M1 for N^{th} and $(N + 1)^{th}$ article in Top-N NRS
M1_p	M1 for N^{th} and $(N + 1)^{th}$ article in probabilistic NRS
p	Represents the probability that an article will be read from the DL

Though we considered different selection probabilities from DL, in the context of the present research we have developed our discussion for a case of influential ($p = 0.9$) NRS (the different simulation paths in the graphs are better seen in color). While the probabilities of article selection from such recommended lists are not known in general, this special case is interesting since it captures a setting in which NRS particularly influence readership.

Results without manipulation. We summarize our findings based on the measures M1 and M2 through selected simulation scenarios. The selected simulation results are presented in figure 3.3-3.8, where left panels are for the M2 measure while the

rights panels are for the M1 measure. The list of various abbreviations used in these figures is given in Table 3.2.
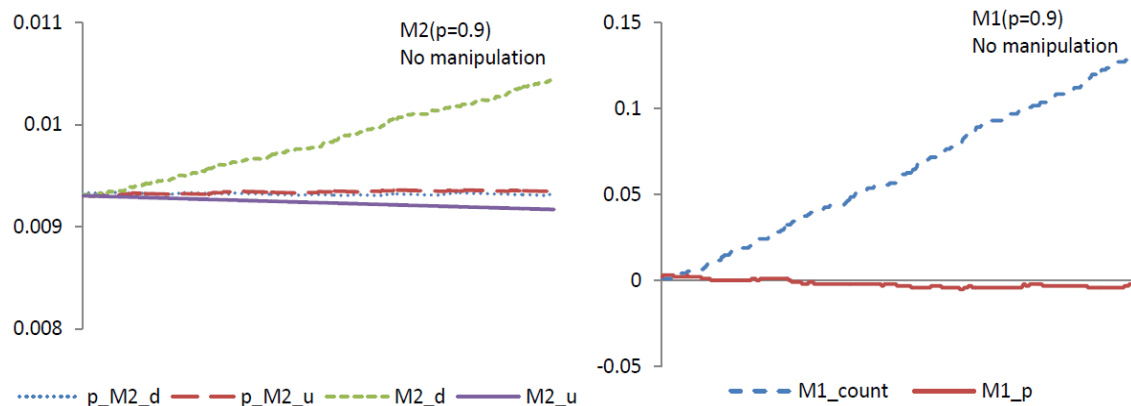
Figure 3.3. Simulation results for the user-model 1 without manipulation (P=0.9)

When there is a high probability that a reader will click on the article recommended by NRS (or *DL*), even negligible initial difference between the counts of Nth and $(N + 1)^{th}$ article gets amplified heavily in the count-based NRS, as it is evident from the consistent increasing pattern of *M1_count* in figures 3.3 and 3.4 for both reader models. For the probabilistic selection mechanism the value of *M1_p* remains close to its initial value (figures 3.3 & 3.4).

The path followed by M2 for N^{th} and $(N + 1)^{th}$ articles in probabilistic NRS (for both reader models) is bounded above and below by the hard cutoff counterpart. In other words, for the count-based NRS, the difference of share between the displayed (*M2_d*) and non-displayed (*M2_u*) article shows a consistent increasing pattern even though initial difference between displayed and non-displayed article was negligible (recall that the only difference between the Nth and $(N + 1)^{th}$ article was a single count/click). This observation highlights the issue of inequality in success of articles created due to presence of hard cutoff NRS.
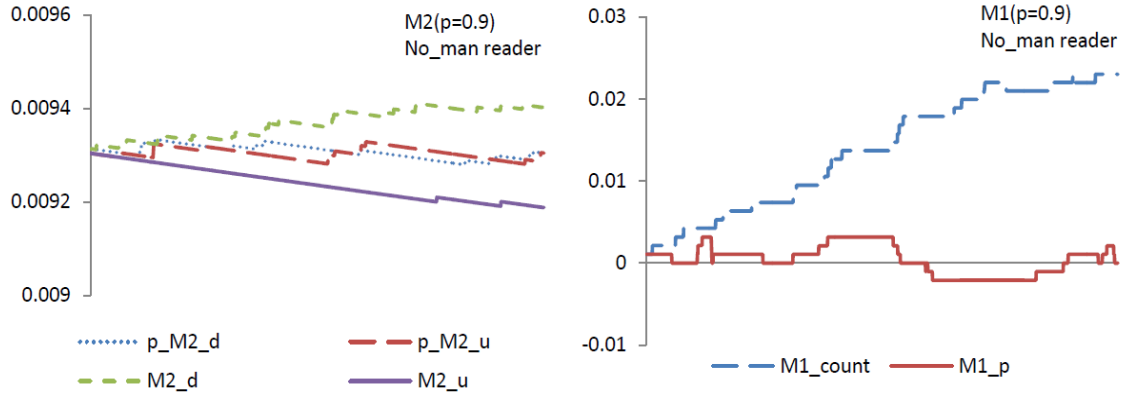
Figure 3.4. Simulation results for the user-model 2 without manipulation (P=0.9)

In a natural system we expect that the share of counts for articles that are almost identical will not vary much. Hence these findings suggest that popular mechanisms using hard cutoffs may be susceptible to fundamentally creating, or amplifying, differences that may not be desirable. Probabilistic selection on the other hand is a more robust mechanism from this perspective.

Results with manipulation. In this section we will discuss the effects of different manipulation scenarios on both NRS. Manipulation counts are uniformly distributed over initial 100 ("early manipulation") and over the entire 1500 article counts ("uniform manipulation"). Two manipulation counts considered are 10 ("low") and 50 ("high") when the system is slightly and heavily manipulated. In total we have four different scenarios of manipulation.

Low fake counts uniformly distributed early.

Low fake counts uniformly distributed over the entire process.

High fake counts uniformly distributed early.

High fake counts uniformly distributed over the entire process.

71

First we will discuss the findings of low manipulated counts. For the $(N + 1)^{th}$ article in RL, its count was increased by 10 randomly, but early in the process. However, findings in this case were completely reversed from the findings in non-manipulated systems, as the reversal of the "M1 paths" in the right panel of figure 3.5 and 3.6 clearly shows.
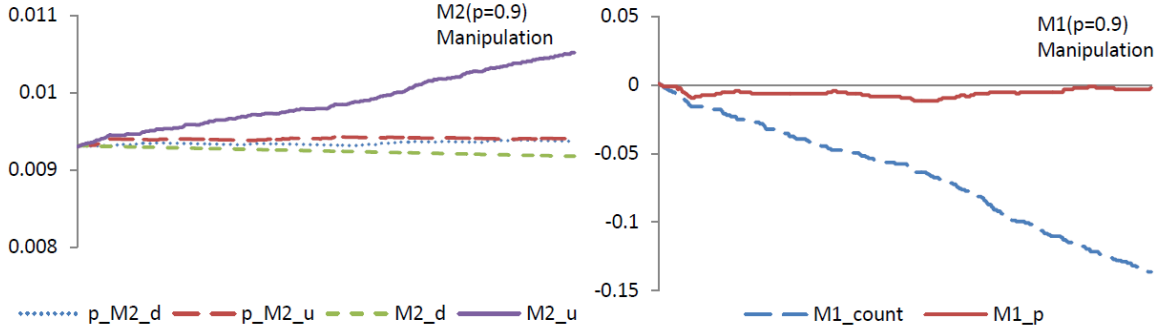
Figure 3.5. Simulation results for the user-model 1 with little early manipulation (P=0.9)

Figure 3.5 presents the case of user-model with random selection of articles from the top-10 list, with reading probability $p = 0.9$. Both measures M1 and M2 (Figure 3.5) suggest that the differences in counts for the manipulated $((N + 1)^{th})$ and the non-manipulated article (N^{th}) gets amplified even if genuine readers arrive in the system. For the second user-model, in which selection of an article is based on r_i, similar phenomena are observed (Figure 3.6).

This suggests that once a manipulator is successful in making his article appear in the DL, the implicit feedback mechanism of count-based NRS will help the manipulated article gain more counts as more readers arrive. This characteristic of the Top-N NRS invites manipulators to put little investment initially to increase the counts of a particular article to make it appear in the DL, after which no further manipulation may be required.

However, for the probabilistic NRS, manipulation seems to have little or no effect (Figure 3.5, 3.6). For low fake counts distributed uniformly, the findings are similar to the case of non-manipulated count-based and probabilistic NRS. Hence, it suggests that a manipulation strategy may not be successful if the effort of a manipulator is distributed over large period of time.
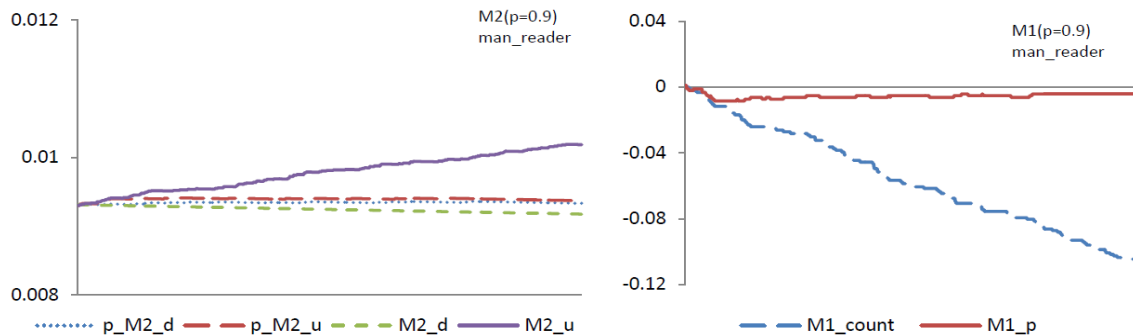
Figure 3.6. Simulation results for user-model 2 with little early manipulation (P=0.9)

We used the second manipulation strategy with high fake counts to compare the performance of both NRS, when the system is heavily attacked by manipulators. In the first case, when 50 counts are randomly distributed over first 100 counts, i.e., system is heavily manipulated in the early stage. The major benefit of probabilistic NRS appears. In all cases, probabilistic NRS produced stable results in which M1 and M2 are not amplified after the manipulation, whereas the performance of count-based NRS is highly distorted for high probability of selection of articles from DL, as seen by declining M1_count trajectory(figure 3.7, 3.8). Also, as expected, the manipulator gets higher benefits in the second reader model with heavy early manipulation strategy (compare right panels of Figure 3.7, 3.8). Finally, for the 50 fake counts distributed over 1500 counts no clear pattern emerges, however both NRS are similar for low selection probability from DL.
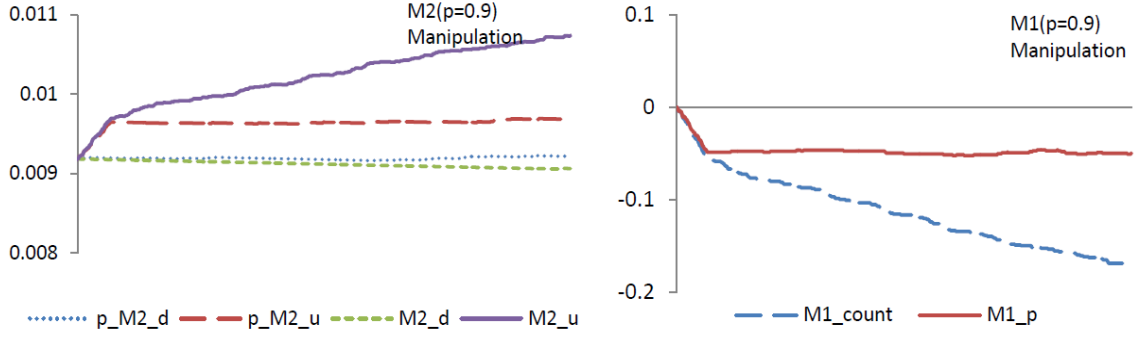
Figure 3.7. Simulation results for the user-model 1 with heavy early manipulation (P=0.9)
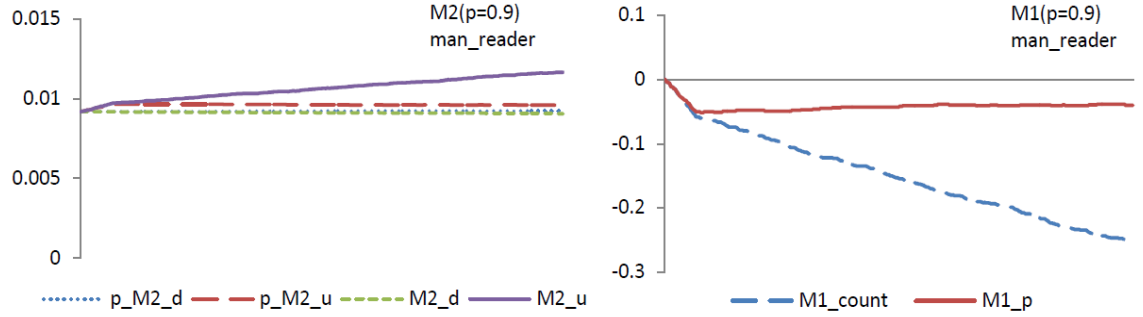
Figure 3.8. Simulation results for the user-model 2 with heavy early manipulation (P=0.9)

Analytical Results

To understand how easily amplification can happen for the hard cutoff NRS, and the robustness of probabilistic NRS toward amplification, we present insights of processes generated through both NRS in a simple setting of a two article case. The discussion that follows provides an intuitive explanation of the phenomenon for a single time step, which is just for illustrative purposes. The complete proofs have been provided in appendix 1.

Assumptions. 1. Two articles are available for recommendation for readers, (article-*a* and article-*b*).

This assumption helps us to establish the analogy between NRS and urn models.

2. Reader upon arrival reads the recommended article with probability p or reads the other with probability $1 - p$.

3. The natural counts for article-a and article-b at time $t = 0$ are given by n_0 and m_0 respectively.

The "natural counts" can be interpreted as the overall preferences of readers for these two articles before any recommender was put in place. Further, without loss of generality we assume $n_0 > m_0$.

Illustration. Let us denote the initial share of article-a and article-b by p_a and p_b respectively, and it is given by $\frac{n_0}{n_0+m_0}$ and $\frac{m_0}{n_0+m_0}$. In this simple one time period model the NRS results in the amplification of the count of recommended article, if at the next step due to recommendation $E(p_a) > \frac{n_0}{n_0+m_0}$.

Count Based NRS

The probability of the recommended article being read is given by p. In the hard cutoff NRS, article-a is always recommended since it has the higher count. Hence, any reading probability $p > \frac{n_0}{n_0+m_0}$ will result in amplification of the counts for the recommended article. Consider a case when $n_0 \sim m_0$ $e.g.$ $n_0 = m_0 + 1$, then hard cutoff NRS will be susceptible to amplification if $p > 0.5$. Given the two article case here, we expect p to be greater than 0.5 for the recommended article.

Probabilistic NRS

In probabilistic NRS, article-a can be read in two ways. The article is in the recommended list (with probability p_a), and the reader chooses to read the recommended

article (with probability p). Or, article-a can be in the other list RL (with probability $1 - p_a$), and the reader chooses to read the un-recommended article (with probability $1 - p$). The total probability that an article-a will be read is therefore given by

$$p(read) = p * p_a + (1 - p) * (1 - p_a)$$

So, in the case of probabilistic NRS, the amplification will happen for the recommended article if

$$p\left(\frac{n_0}{n_0 + m_0}\right) + (1 - p)\left(\frac{m_0}{n_0 + m_0}\right) > \frac{n_0}{n_0 + m_0}$$

$$\Leftrightarrow m_0 + p(n_0 - m_0) > n_0$$

$$\Leftrightarrow p(n_0 - m_0) > n_0 - m_0$$

The above condition will never be true for any probability p. It is easy to see that when the counts are similar, probabilistic NRS does not create amplification (reading probabilities will both be 0.5).

Building on this, below we present results for the more general case where we examine counts at the end of n iterations.

Proposition 1. In the Top-N NRS total expected count ($denoted\ as\ E(A_n^h)$) for article-a after $'n'$ iterations is given by $(n_0 + np)$.

Proposition 2. In the probabilistic NRS total expected count ($denoted\ as\ E(A_n^p)$) for article-a after $'n'$ iterations is bounded by the interval (I_1, I_2).

Where $I_1 = \left(\frac{n_0 + m_0 - 1}{n_0 + m_0 + n - 1}\right)\left(\frac{n_0 - m_0}{2}\right) + \frac{n_0 + m_0 + n}{2}$ and

$I_2 = \frac{n_0}{n_0 + m_0}(n_0 + m_0 + n)$.

We discuss the implications of these propositions shortly. Before doing so, we briefly comment on the proofs (presented in the appendix). Proposition 1 has been

76

established through a simple binomial process. Whereas for Proposition 2, modeling based on an urn framework from probability theory has been used. To the best of our knowledge, the only prior work that has used urn models in the context of recommender systems is Fleder and Hosanagar (2009) where they study the impact of recommender systems on sales diversity. However, our use of Pólya's and Bernard Friedman's urn models to derive analytical results is novel and our analytical results have been established in a substantially different manner. Below we discuss our use of these urn models in the proofs.

The probability of article-a being recommended in probabilistic NRS is given by p_{at}, where p_{at} represents the share of the article-a at any given time t; initially we have $p_{a0} > p_{b0}$ (assumption 3).

For $t > 0$ the total probability that the article-a being read at time t in probabilistic NRS is

$$p_t(read) = p * p_{at} + (1 - p) * (1 - p_{at}) \tag{1}$$

Each time an article is read, its count is increased by 1. We also define two parallel processes that start with the same initial condition. However, for these processes reading probabilities (i.e., p) for the recommended article is given by 0 and 1 respectively, at each time step. We denote reading probabilities for these processes at each time step as,

$$p_{tl}(read) = 1 - p_{at} \tag{2}$$

$$\text{and } p_{tu}(read) = p_{at} \tag{3}$$

Let us denote the count of article-a being A_n^p, A_{nl}^p and A_{nu}^p after n time steps for the processes defined by equations (1), (2) and (3) respectively. Say, τ_n denotes the total

counts of articles in the system at any given time n. The value of τ_n at a given time n is known *a priori* in the present framework and is equal to $n_0 + m_0 + n$.

Since, $p_{tl}(read) \leq p_t(read) \leq p_{tu}(read)$, the following relation holds for the processes defined by equations (1), (2) and (3)

$$E\left(A_{nl}^p\right) \leq E(A_n^p) \leq E(A_{nu}^p) \tag{4}$$

$E\left(A_{nl}^p\right), E(A_{nu}^p)$ are the values of I_1 and I_2 respectively, mentioned in Proposition 2; that will be derived in this section based on the urn formulation.

But, before that, we present the urn problem as described by Bernard Friedman (Freedman 1965). An urn contains W_n white balls and B_n black balls at time n. One ball is drawn at random and then replaced, while α balls of the same color as the ball drawn and the β balls of the opposite color are added to the urn. Now, let us consider two cases that will be used in the present research.

Case1: $\beta = 0$ describes the Pólya Urn mechanism in the above section where selection probability of a white ball (and vice versa for a black ball) at each time step is given by its share – which is a characteristic of the problem proposed by Pólya to model contagion (Eggenberger and Pólya 1923). When $p_t(read) = p_{at}$ (i.e., share of the article 'a'), the path followed by A_n^p is obtained through the Pólya Urn mechanism with $\alpha = 1$.

Case2: The special case of Friedman's Urn with $\alpha = 0, \beta = 1$ helps us to establish lower bound for $E(A_n^p)$. In this case the selection probability of a ball is given by $1 -$ (share of a ball in the urn). When $p_t(read) = (1 - p_{at})$ (i.e. $1 - \{$share of the article $'a'\})$, the path followed by A_n^p is obtained through special case of Friedman's Urn formulation described here.

78

Figure 3.9 pictorially depicts the urn processes. The proofs of the propositions are completed in the appendix.
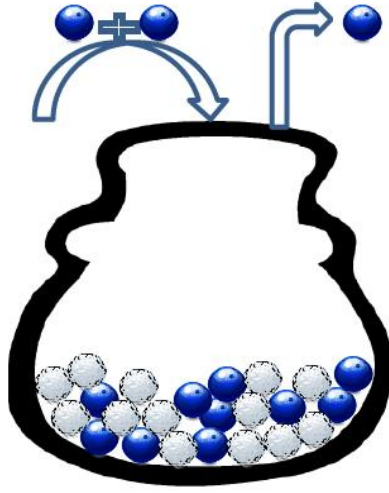
Figure 3.9. The Pólya urn   A Bernard Friedman urn
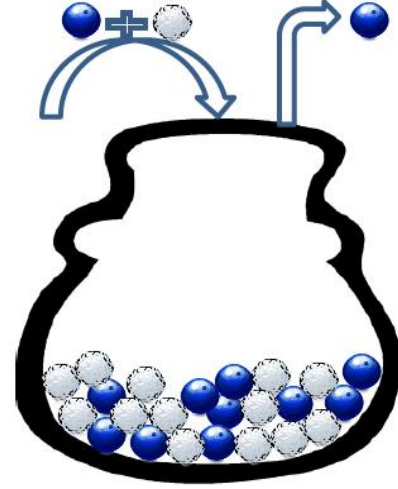
Implications. From propositions 1 and 2 we have $E(A_n^h) = n_0 + n * p$ and

$$\left(\frac{n_0 + m_0 - 1}{n_0 + m_0 + n - 1}\right)\left(\frac{n_0 - m_0}{2}\right) + \frac{n_0 + m_0 + n}{2} \leq E(A_n^p) \leq n_0 + \frac{n_0}{n_0 + m_0} n$$

Now consider a case where NRS has fairly strong influence on reading behavior i.e., $p{\sim}1$ and the difference in the sufficiently large natural counts after which articles $'a'$ and $'b'$ make into NRS is negligible i.e., $n_0 - m_0 {\sim} 0$, in particular let us assume $n_0 = m_0 + 1$. So, the approximate value of expected count of article-a in hard cutoff NRS and probabilistic NRS is given by

$$E(A_n^h) = m_0 + 1 + n \tag{5}$$

and

$$\frac{m_0}{2m_0 + n} + \frac{2m_0 + n + 1}{2} \leq E(A_n^p) \leq m_0 + 1 + \left(\frac{m_0 + 1}{2m_0 + 1}\right) n \tag{6}$$

Increase in the counts of Top-N selection and probabilistic selection NRS due to recommendation can be obtained through subtracting the initial count of article-a in expressions (5) and (6). So, we have

$$E(A_n^h) - (m_0 + 1) = n \qquad \text{(7) and}$$

$$\frac{m_0}{2m_0 + n} + \frac{n-1}{2} \leq E(A_n^p) - (m_0 + 1) \leq \left(\frac{m_0+1}{2m_0+1}\right)n \quad \text{(8)}$$

Using approximation $\frac{m_0+1}{2m_0+1} \sim \frac{1}{2}$ in expression (8) for sufficiently large m gives us following condition

$$\frac{m_0}{2m_0+n} + \frac{n}{2} - \frac{1}{2} \leq E(A_n^p) - (m_0+1) \leq \frac{n}{2} \qquad \text{(9)}$$

For large n, from (7) and (9)

$$E(A_n^h) - (m_0+1) \rightarrow n \quad \text{and} \quad E(A_n^p) - (m_0+1) \rightarrow \frac{n}{2}$$

So, from the above expressions we conclude that for two equally good articles, probabilistic NRS is less susceptible to artificial amplification in counts for the recommended article, whereas hard cutoff NRS generates processes that leads to highly amplified counts for the recommended article when the NRS is fairly influential (p is very high). This is the case since two articles with the same counts initially should increase their respective counts by $\sim n/2$ at the end of n iterations, which happens with the probabilistic mechanism only.

NRS Manipulation

Proposition 3. (Effectiveness of Early Manipulation). Consider two scenarios in which an article is manipulated once at two different time steps t_1 and t_2 such that $t_1 < t_2 (t_1, t_2 < n, t_0 = 0)$ for any NRS; where n represents the total number of new counts

for both articles over the entire time. We call these manipulation strategies $Ma1$ and $Ma2$. Then for any NRS (Top-N or probabilistic), $Ma1$ will be more beneficial for a manipulator than $Ma2$.

Proof:
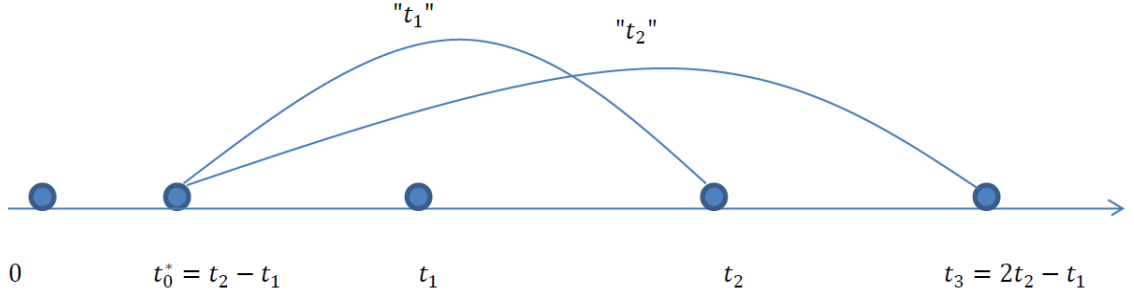
Figure 3.10. Illustration for proposition 3

On the contrary, let us assume that (late) manipulation at t_2 will be more beneficial for a manipulator than t_1. Then we can find a new time point $t_0^* = t_2 - t_1$ such that $t_3 > t_2$ and $t_3 = t_2 + (t_2 - t_1) = 2t_2 - t_1$ which will be more beneficial for manipulation than implementing manipulation at t_2 and hence also from t_1 (difference between t_0^* and t_2 is t_1). Applying the same argument, again we can obtain a time point $t_0^* = 2t_2 - t_1 - t_1 = 2t_2 - 2t_1$ such that $t_4 > t_3$ and $t_4 = 3t_2 - 2t_1$ will be more beneficial for manipulation than implementing manipulation at t_3 and hence from t_1. Repeatedly applying the same argument we can find an integer $m > \frac{n - t_2}{t_2 - t_1}$ such that applying the above argument m times gives us a time point $t_m > n$ will be beneficial for manipulation. Hence under the above assumptions, a manipulator will get maximum benefit without introducing any manipulated count in the system. This cannot be the case since the act of manipulation in this model provides a strictly higher count for the article

being manipulated and is assumed to have no additional cost to the manipulator. Hence by contradiction the proposition holds.

To examine the bounds for both Top-N and probabilistic NRS, we consider a case in which manipulation is introduced in both NRS at a very early stage. Let both NRS operate until n total new counts are received in the system. At extreme if all ϵ fake counts are introduced consecutively at the very early stage, then the total count of article$-b$ after manipulation will be $m_0 + \epsilon$. Further assume that manipulation ϵ introduced in the system is such that $(m_0 + \epsilon) > n_0$. After manipulation, both NRS can be viewed to operate as genuine NRS, but with distorted initial counts $n_0, (m_0 + \epsilon)$ for article-a and article$-b$ respectively.

Proposition 4. For a manipulator who injects ϵ fake counts in the probabilistic NRS, the increase in counts of article$-b$ after manipulation is bounded by $\frac{m_0+\epsilon}{n_0+m_0+\epsilon}(n - \epsilon)$ where n is the total number of new counts for both articles over the entire time.

Proof: We denote the distorted share of article a and b at time t after injection of manipulation as p'_{at} and p'_{bt} respectively. Clearly $p'_{bt} > p'_{at}$ and probability that the article b will be read at time t will have following property:

$$p'_{bt}(read) = p * p'_{bt} + (1 - p) * (1 - p'_{bt}) \qquad (10)$$

$$\text{and } p'_{bt}(read) \le p'_{bt}$$

Let us denote the count of article b being $B_n^{,p}$ and $B_{nu}^{,p}$ after n time steps (i.e., $(n - \epsilon)$ time steps after manipulation) for the processes where $p'_{bt}(read)$ is given by $\{p * p'_{bt} + (1 - p) * (1 - p'_{bt})\}$ and p'_{bt} respectively. For these processes the expected count of article b after n time steps satisfies the following relation.

$$E(B_{n-\epsilon}^{;p}) \le E(B_{n-\epsilon u}^{;p}) \tag{11}$$

For the random processes when $p'_{bt}(read) = p'_{bt}$ (i.e. $p = 1$), the path followed by the count of article$-b$ is similar to the Pólya's urn mechanism as discussed earlier. But, in the present case, initial count of articles 'a' and 'b' has been changed to n_0 and $m_0 + \epsilon$. The expression for $E(B_{n-\epsilon u}^{;p})$ can be obtained in the similar way as discussed in appendix (for *proposition 2*) to obtain the upper bound (i.e. I_2). So,

$$E(B_{n-\epsilon u}^{;p}) = (m_0 + \epsilon) + (n - \epsilon)\frac{m_0 + \epsilon}{n_0 + m_0 + \epsilon} \tag{12}$$

Using the inequality (11) and the result (12)

$$E(B_{n-\epsilon}^{;p}) - (m_0 + \epsilon) \le \frac{m_0 + \epsilon}{n_0 + m_0 + \epsilon}(n - \epsilon) \tag{13}$$

Corollary. When the distorted initial counts of articles a and b is n_0 , $(m_0 + \epsilon)$ respectively, the increase in the expected count of article$-b$ after manipulation in hard cutoff NRS is equal to $(n - \epsilon) * p$ where n is the total number of new counts of both articles over the entire time.

The above result can be established with simple binomial model used in *proposition 1* over $n - \epsilon$ time steps with $B_n^{;h} = (m_0 + \epsilon)$ initally. Let $B_n^{;h}$ represents the total count of article$-b$ after n^{th} iteration in the hard cutoff NRS. Then

$$E(B_n^{;h}) = m_0 + \epsilon + (n - \epsilon) * p \tag{14}$$

Implications. When NRS has fairly strong influence on the reading behavior, i.e., $p\sim 1$ a manipulator can drive the majority of the reader's attention towards the manipulated article as illustrated in expression (14) $E(B_n^{;h}) - (m_0 + \epsilon) \to n - \epsilon$ for any ϵ that satisfies the condition $(m_0 + \epsilon) > n_0$.

For illustration consider a special case when $n_0 = m_0 + 1$. For this condition by injecting any fake count $\epsilon > 1, e.g. \ \epsilon = 2$ initially, the hard cutoff NRS can be completely rigged. Whereas, in case of probabilistic NRS $E(B_{n-\epsilon}^p) - (m_0 + \epsilon) \leq \frac{m_0+\epsilon}{n_0+m_0+\epsilon}(n - \epsilon)$, and hence its performance is not disturbed by small manipulation efforts, as for small value of ϵ the expression in (13) can be approximated as $\sim \frac{n}{2}$ for large n.

Hence in a special case, we show analytically that (1) early manipulation can pay off well for a manipulator, and (2) that this is true only for the Top-N recommender, since the probabilistic mechanism is shown to be robust against such manipulation.

Analysis of Probabilistic NRS

In this section we further analyze probabilistic NRS in two ways. First, we present and discuss an accuracy-distortion tradeoff. Then, we compare it against a novel adaptation of the Influence Limiter algorithm.

The accuracy/distortion tradeoff.

Accuracy (MAE). One drawback of the probabilistic recommendations is that it potentially chooses articles to recommend that might not be in the current "best" list. To quantify that loss in the recommendation process, the Top-N and probabilistic NRS are compared based on the "quality" (measured as popularity) of the articles appearing in the recommended list. A widely used measure for this purpose is mean absolute error (MAE). It represents an efficient means to measure the statistical accuracy of predictions of articles appearing in the Top-N recommendation (Ziegler, et al. 2005). Let us denote

the count of ith article appearing in count-based NRS at j^{th} time step as N_{ij}^h and in probabilistic NRS as N_{ij}^p. The MAE metric denoted as $|E|$ is defined as,

$$|E|_j = \frac{\left(\sum_i N_{ij}^h - \sum_i N_{ij}^p\right)}{\sum_i N_{ij}^h}$$

In the above expression, $\sum_i N_{ij}^h$ and $\sum_i N_{ij}^p$ represent the sum of counts of all articles that appear in count-based and probabilistic NRS respectively, at the jth time step. The MAE metric has been averaged over the number of iterations, as the simulation progresses.

$$|\bar{E}| = \frac{1}{|t|} \sum_{j=1}^{t} \frac{\left(\sum_i N_{ij}^h - \sum_i N_{ij}^p\right)}{\sum_i N_{ij}^h} \qquad (15)$$

This metric presents accuracy loss in terms of "high" ranked articles assuming that users will have little or no interest in the "low" ranked articles, averaged over the complete simulation.

Distortion (KL). Assuming that the initial share of articles represents the "true" preference of readers, the distortion created by each NRS in comparison with their initial share is given by $Kullback - Leibler$ (KL) distortion measure (Kullback and Leibler 1951). Let us denote the probability distribution for articles in each NRS (probabilistic and Top-N NRS) at the iteration t as $q_t(x_i)$. Then the KL distortion for the articles $\{x_1, x_2, \ldots \ldots x_n\}$ is given by.

$$D_{KL}(p||q_t) = \sum_{i=1}^{n} p(x_i) \ln\left(\frac{p(x_i)}{q_t(x_i)}\right) \qquad (16)$$

In other words, the above expression represents the inefficiency of the distribution q, when the true distribution of articles is p (given initially).

Since, the emergence of counts of the articles in a given NRS is a probabilistic process, the data was generated through fifteen replications of the complete simulation for the different values of reading probability for both reader models. The results discussed below are based on the mean value of metric over fifteen replications, plotted against different choice of reading probabilities.

Considering the performance based on MAE (equation 15), we observe that Top-N seems to perform better than probabilistic NRS (Figure 3.11), as the findings are established from both reader models in the simulation. However, under the second metric (KL, equation 16) clearly probabilistic NRS outperforms Top-N NRS for both reader models (Figure 3.12). These findings present the tradeoff between the two NRS. While
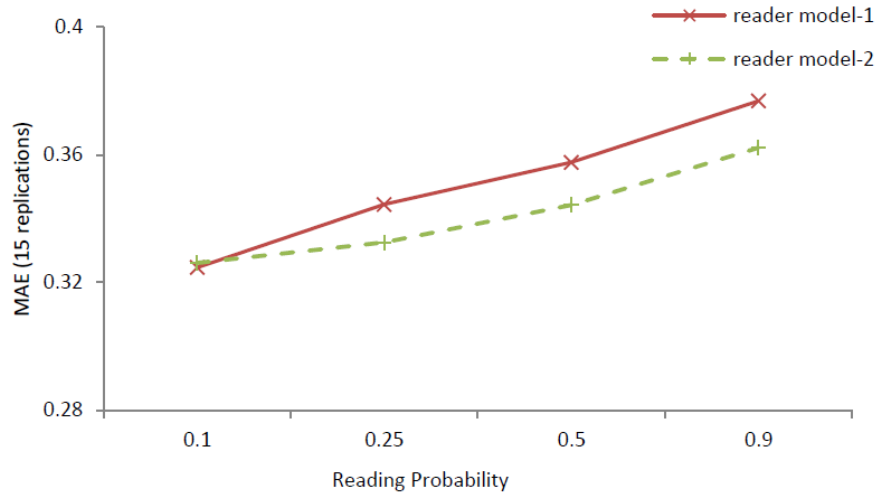
Figure 3.11. Mean Absolute Error vs. Reading Probability

probabilistic NRS seems to have a small accuracy loss (in terms of counts of articles it recommends) it is more true to the natural shares of the articles and does not create distortions which otherwise can occur.
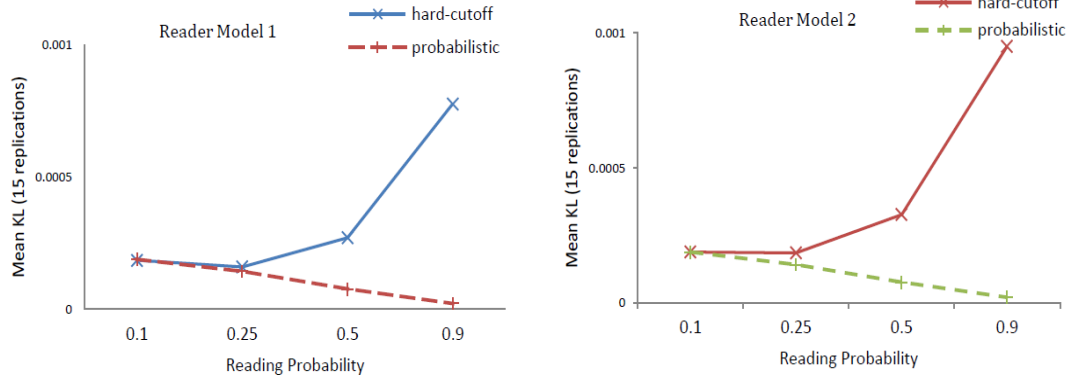
Figure 3.12. Mean KL distortion vs. reading probability (both Reader Models)

After the counts of articles have achieved a steady state in a natural system, we expect that the share of articles will not deviate much. This behavior of system is achieved through probabilistic NRS, with a slight loss in recommendation accuracy.

Comparison to an "adapted" influence limiter heuristic. We have discussed the advantage of probabilistic mechanism in terms of robustness towards manipulability. However, one limitation in the news recommender research is the lack of benchmark to which the performance of probabilistic mechanism can be compared towards manipulation. So the approach of Resnick et al. (2007) has been adapted in our context to compare the effects of manipulation in NRS.

As mentioned earlier, the Influence Limiter algorithm (Resnick and Sami 2007) generates item recommendations controlling rater's influence on recommender systems through reputation acquired over time. The reputation of a rater is updated based on rating provided by him to an item and the *loss function* determined through the prediction made to a target user compared to the actual preference of the target user.

In this research our focus has been on the counts of articles, and the reader's individual behavior (or reading pattern) has been left out for ease of exposition. Hence,

the approach of Resnick and Sami (2007) cannot directly be used. Instead, we limit the influence of fake counts to generate article recommendations. In a similar vein, it should be also noted that in the present analysis counts of articles is updated, instead of rater's reputation.

In our approach, we assign reputation for each article based on prior information about the average inter arrival time of two consecutive clicks for the recommended article, the total counts received by the article, and the time period of observation in which influence limiting process operates. The influence limiting process operates in a pre-defined time interval. An article is assigned a reputation based on observation during this period. After this time interval new counts received by an article are updated based on its reputation score.

We assume that the average time interval of two consecutive counts received by a recommended article is less than the average time interval of two consecutive counts received by the other articles in the system. A measure β_{ij} has been introduced that limits the influence of a manipulator in the top$-N$ NRS. For any article j at time t_i it is defined as,

$$\beta_{ij} = \min(1, R_{ij}) = \min\left(1, \frac{t_i - t_0}{\alpha * (c_{ij} - 1)}\right) \tag{17}$$

Influence limiting process operates between a pre-selected time intervals (t_0, t_n), and can be determined through the designer's experience or other appropriate choice can be the time interval when manipulation activity is most observed. For every $t_0 \leq t_i \leq t_n$, an article $j's$ reputation is updated as given in equation 17. In the expression α represents the average time interval that is "reasonable" between two consecutive counts received by a recommended article in the top$-N$ NRS (this can be determined through

the arrival distribution of counts of the recommended articles), and c_{ij} is the number of counts received by the article j in the time interval given by (t_0, t_i). After t_0, at any given time point t_i the new count received by the article j (denoted as c_{ij}') passes through an influence limiting process to generate a modified count given by $\tilde{c}_{ij}$ as described below in the pseudo code. ($\tilde{c}_{0j}$ represents count received before t_0). After t_n, each new count received by any of the articles, is modified through is reputation β_{nj} at time t_n. When $R_{ij} \geq 1$, all weight is on c_{ij}' i.e., article j has full credibility.

An Adapted Influence Limiter Heuristic:

1. Get $\tilde{c}_{0j}$ for each article at $t = t_0$
2. For each article j, $c_{0j} \leftarrow 0$
3. For each t_i, when $1 \leq t_i \leq t_n$ and $c_{ij} \geq 2$
   a. For each article j
      i. $\beta_{ij} \leftarrow \min(1, R_{ij})$
      ii. $\tilde{c}_{ij} \leftarrow \tilde{c}_{i-1j} + \beta_{ij} * c_{ij}'$
      iii. $c_{ij} \leftarrow c_{i-1j} + c_{ij}'$
   b. End for
4. End for

Let us consider the first user model in our simulation (when the reader performs random selection of an article from the recommended list). Also it should be noted that in the context of manipulation, we are concerned about articles appearing in the recommended list. The initial 100 time steps have been selected as the observation period before implementing the modified count (the influence limiting heuristic) for each article. The selection of an article from the recommended list is performed randomly, hence the expected count that an article will receive over initial 100 time steps will be$\left(\frac{100}{10}\right) * p = 10 * p$, where p is the selected reading probability in the simulation.

89

Hence, the expected time interval between two consecutive counts received by an article in Top$-N$ NRS is given by $\alpha = \left(\frac{100}{10*p}\right) = \frac{10}{p}$. Based on our choice of time period for the observation $t_i - t_0$ will be t_i. Hence, the reputation of an article j at time t_i will be given by (equation (17))

$$\beta_{ij} = \min(1, \frac{t_i * p}{10 * (c_{ij} - 1)})$$

As established earlier, the major issue of interest is manipulation at the early stage (which was shown to be more effective for the manipulator). Hence, variants of manipulation examined are heavy and low early manipulation. As before the articles of interest in this are also the Nth and $(N + 1)^{th}$ articles in the list.
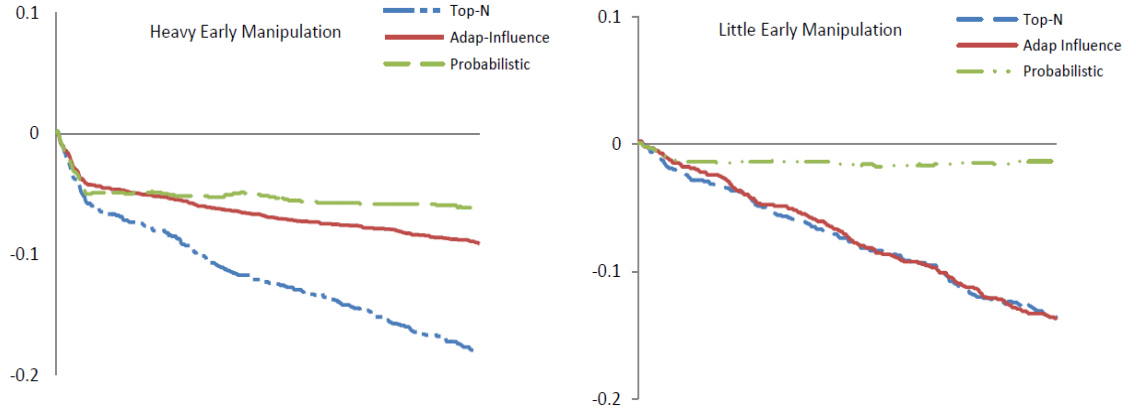
Figure 3.13. Comparison of Manipulation based on M1

The results suggest that in the case of extreme manipulations, the proposed adapted influence limiter heuristic performs similar to probabilistic NRS (figure 3.13, left panel). This seems to be by the design of the adapted influence limiter heuristic - as the manipulator injects more fake counts for the target article, this leads to less reputation for it (β_{ij}). In turn, new counts received by the manipulated article cause less cumulative

increase in its count. However, small manipulation effort (especially if an article has just missed the cutoff for Top-N and the manipulator is in a position to determine this) may go undetected in case of the adapted influence limiter (figure 3.13, right panel). Here, probabilistic NRS is still robust.

Sensitivity Analysis

In case of news articles, where majority of queries are driven by front page display or recommended articles, we expect popularity to exhibit some kind of power law distribution. The rationale for power-law distribution of popularity, especially in web-based systems, has been suggested by Easley and Kleinberg (2010). This assumption of popularity is also consistent with the effect of social influence discussed by Salganik, et al. (2006). In their experiment for artificial music market, they found that in the presence of social influence, such as media sites, we observe greater inequality – popular entities are more popular and unpopular entities are less popular. From a given power-law distribution its corresponding Zipf distribution can also be obtained (Adamic 2000).

To validate the popularity distribution of articles, we obtained data on popularity of articles from DailyMe Inc., a company that provides news personalization technology to a large number of media sites. There are five datasets from five different local news websites serving markets in Connecticut, Pennsylvania, New York, Colorado and Massachusetts, collected during the period of February 2012 to April 2012. The data provided listed specific articles along with cookie IDs and time stamps read across the five different local news websites.

Figure 3.14. log-log plot for popularity of articles at five different sites

Figure 3.14 shows the normalized frequency distribution on a log-log scale using the *logarithmic binning* with multiplier of 2 – similar to the procedure described by Newman (2005). The X-axis corresponds to natural log value of bins and Y-axis corresponds to the natural log value of normalized frequencies. Data from these five real local news websites show the pattern of power-law in popularity. Based on the findings,

the power-law exponent used to discuss results in the sensitivity analysis is given by $c = 1.7$. We used this exponent value in the modified simulation model, where the initial distribution of article counts was generated using power-law distribution with exponent 1.7. We note that these are relatively smaller local news Web sites and we do not therefore make broader generalizations about the power law based on these alone. However, it provides valuable insights for this sensitivity analysis.

The detailed information about data analysis, model implementation and findings in this case has been provided in appendix 2. Our main findings from the sensitivity analyses were that probabilistic selection continues to offer significant benefits compared to Top-N when it comes to mitigating count amplification as well as offering resistance to manipulation. However, the continued benefit from manipulation (i.e., after manipulation activity stops) is lower in this case than was the case under the uniform distribution.

Social Desirability

The analysis presented in this paper demonstrates that the probabilistic selection mechanism is effective in addressing some of the key limitations of the Top-N NRS. These limitations were (1) amplifying the negligible initial difference in the counts of Nth and $(N + 1)^{th}$ articles (2) less choice of articles offered to readers' by top-N list and (3) susceptibility to manipulation by artificially inflating the count of a target article.

Still, it is difficult to argue universal superiority of one of the two selection processes for recommendations. For example, when an implementer is facing a situation with suspected manipulation activities or she wants to create a set of diverse recommendations in the recommended list, then surely, the probabilistic mechanism will be more desirable. However, in a situation where an implementer wants to maximize the

short-term revenue or to allow a genuine article to (perhaps deservedly) become more popular, then the Top-N NRS may be the appropriate choice.

One approach in framing this issue is to ask which mechanism is socially more desirable. Clearly the choice of a particular mechanism depends on the goal of an implementer and desired effect he wants to create through those recommendations. Researches in other contexts have also framed it in this lens. For instance, Salganik, et al. (2006) highlight the problem related with the measure of quality, through an experimental approach, in the presence of social influence.

In this research we do not answer this issue directly. Instead we view this as a *control* that the media owners can exercise by their choice of parameters. If viewed in this manner, the natural question is to ask if there can be some continuous spectrum of control that can be used (possibly fine-tuned) by managers to achieve any outcome or behavior that they may desire. Below we show that this is possible. By introducing a feedback parameter, we can offer managers an elegant approach to control the behavior of the system such that it can operate in the entire spectrum. We provide details below.

We extend the approach of probabilistic selection to provide greater flexibility for an implementer. In this modified approach, the selection probability of an article −'a' having count $c_a(t)$ at time t, is given by

$$p_a(t) = \frac{c_a{}^{\gamma}(t)}{\sum_j c_j{}^{\gamma}(t)}, \gamma \geq 0 \qquad (18)$$

One advantage of this modified probabilistic approach (equation 18) is that we can generate different known selection processes through tuning the parameter γ in a single unified equation. To understand the behavior of systems for the modified selection

process, we consider different values that γ can take, and briefly explain the selection

processes corresponding to those values.

$\gamma = 0$: In this case all articles have the same probability of being selected in the

display list, which essentially simulates a random recommender. Also, in this case we do

not incorporate any information generated through user's interaction with NRS. This

selection process can be desirable when an implementer wants to completely eliminate

the effect of social influence from NRS.

$0 < \gamma < 1$: For the given probability function, the number of times an article

appears in DL will tend to be in equal proportion for all articles after very large time

interval. In practice, values of the feedback parameter in this range may have very limited

application.

$\gamma = 1$: In this case, the count evolution process can be analyzed as a combination

of Pólya and Bernard Freedman urn problems in a special case. This particular case has

been discussed in detail throughout this paper as the main probabilistic selection

mechanism. As mentioned earlier, the selection mechanism in this case is desirable to

generate an even distribution in popularity, to generate diverse recommendations and to

thwart manipulation efforts.

$1 < \gamma < \infty$: For $\gamma > 1$, we will have a system with positive feedback for the

articles with high counts. In other words, the NRS generates recommendations such that

the articles with high counts will have an even higher probability of being selected in the

display list (DL) at the next time step. In this case, after a finite time, the probabilistic

NRS will behave similar to Top-N NRS i.e., N articles with high counts will always be

selected for recommendation. The feedback based approach in this range is desirable to

mitigate the issue of penalizing the marginal next article (the case with count based selection process), and at the same time maintaining the effect of Top-N selection.

$\gamma \to \infty$: In this case, the probabilistic recommendation generated by equation 18, is essentially a replication of most popular NRS (i.e., identical to the Top-N mechanism considered earlier in this paper). To understand this, let us consider the expression given by equation 18,

$$p_a(t) = \frac{c_a{}^\gamma(t)}{\sum_j c_j{}^\gamma(t)} = \frac{1}{1 + \sum_{j \neq a}\left(\frac{c_j(t)}{c_a(t)}\right)^\gamma}$$

We assume that all articles have different counts[9]. Then, $\forall j$ such that $\frac{c_j(t)}{c_a(t)} < 1$; $\lim_{\gamma \to \infty}\left(\frac{c_j(t)}{c_a(t)}\right)^\gamma \to 0$. So, for the article with highest count (among those which are not yet selected for DL), the selection probability in DL will be 1. Hence, N probabilistic selections in this case correspond to selection of N articles with decreasing order of their counts.

The proposed feedback mechanism in this section provides implementers flexibility in selecting of articles and also allows users to process the recommended information in different ways. Depending on various situations, we can reduce the rich get richer effects for articles, or amplify them, or steer them in different directions (with articles with low counts becoming more popular for $\gamma < 0$) by help of the parameter γ. Therefore, the use of the above feedback model, provides a broader range of control that can be exerted to optimize the behavior of the system for a particular manager. We defer analytical results and a more detailed study of this to future work.

---

[9] Without loss of generality.

Discussion and Conclusion

There has been growing evidence of the influence of News Recommendation Systems on users. A recent article in the Wall Street Journal (WSJ) (Warren and Jurgensen 2007) had noted that the influence of NRS is sparking a new form of "payola" as marketers try to get more votes and allow users to vote for their favorite submissions. This phenomenon has been further propelled by social networking applications such as Facebook and Twitter, as noted by The Economist in recent review of news industry (Economist-b 2011). As per the article published in WSJ, the aggregation process of news through NRS is also giving rise to an "obsessive sub-culture of a few active users who just purely for the thrill of it, are trolling the web-space for news and ideas to share with others". For example, a Reddit user is known for "scoping" drove about 100,000 visitors to one amateur photographer's website (Warren and Jurgensen 2007). There are also some marketing companies in existence who promise clients that they can get a client front-page exposure in exchange for a fee (Warren and Jurgensen 2007). In other cases users can also buy Facebook fans (likes) ("http://socialnetworksolutionz.com/index.html) or tweets ("http://pay4tweet.com), to gain popularity.

In light of all this, news recommender systems should be particularly careful to avoid common manipulative strategies. At present, the articles with highest count or popularity are displayed on the front page prominently on most news sites and these are seen by millions of people. It is evident from the findings we present in this research that the practice of using a "hard cutoff" is in particular a potentially troublesome one. In addition to unduly penalizing the possibly equally good next article that missed this

cutoff, this system is quite vulnerable to manipulation. A simple probabilistic mechanism can instead be used to present popular articles and has some desirable properties as we show and study in this paper.

Practically, implementers may instead choose a more flexible mechanism that may offer the benefits of both Top-N and probabilistic selection. This can be done using a parameterized extension of the probabilistic selection mechanism, as noted in the chapter. We defer analytical and empirical treatment of the parameterized extension to future work.

In the present research we have established our main results based on simulation and theoretical results using widely studied urn models. The performance of the common Top-N recommender and the probabilistic counterpart proposed here has been analyzed based on two different metrics. Further the tradeoff from using the probabilistic recommender is also shown. Finally, an adapted influence limiter algorithm has been introduced, and its performance has been compared with its probabilistic counterpart. We have also, in a sensitivity analysis driven by real data from local news Web sites, shown the robustness of our main results to distributional assumptions. To our knowledge, the problem studied here is novel and these are all unique contributions of our research.

The probabilistic NRS has practical implications in terms of providing a better way of utilizing information generated through users in comparison to the current Top-N NRS in the recommendation process. The present research also has policy implications, as government and policy think-tanks are increasingly concerned about the entire process of news generation, curation and distribution (Economist-a 2011, Loretta and Brian 2011). In an only somewhat light vein, Burt Herman writes in a prediction for the

Niemann Journalism Lab that "in the coming year, social media journalists will "#Occupythenews".

It should be also noted that although we have derived our results in a framework of discrete time steps, statistical distribution of urn functions has been widely studied in continuous case. For example, (Freedman 1965), discusses the asymptotic behavior urn of functions. In future research, these functions can be explored to address other issues related to recommender systems research.

Other possible extensions of the present research could be to study the impact of hard cutoff in personalized recommended systems. Although for a given user, even in the context personalized recommendation, the issue of hard cutoff still exists, it would be interesting to investigate, to what extent count amplification can be mitigated at aggregate level.

On a broader level, algorithms are increasingly in control of what news articles get shown to which user. Some, such as Eli Pariser, the author of the popular book "The Filter Bubble", believes this to be a potential problem. The popular argument here is that algorithms will influence thought by controlling news, and that such algorithms tend to become hyper-personalized, creating "bubbles", where each user is in a possibly independent bubble. Others, including many academics in a panel at the recent 2011ACM Conference on Recommender Systems in Chicago, believe this problem to be overblown, and that algorithms can both personalize as well as provide adequate diversity to limit such problems. In one of the earliest works in the IS area for instance, Adomavicius and Kwon (2011) present methods to enhance diversity. It is in this context though that some important research problems emerge. Studying the specific characteristics of news

recommendation algorithms is an important area of research, given the fact that news indeed shapes public opinion on a variety of topics and that algorithms are increasingly influencing its distribution.

As a last thought, while comparing probabilistic approaches for selection in recommender systems, we note that a similar argument can be made not just for "news" recommendations, but for any recommender that uses a hard cutoff. For instance, Amazon.com's product recommendations most likely use hard cutoffs based on results generated from collaborative filtering (Linden, et al. 2003), and can perhaps therefore benefit from using probabilistic variants such as described in this paper. Currently, the "Customers who also bought this item also bought" features a list of specific recommendations on each page − the fate of the "next" product in that list that misses such a cutoff is similar to the question studied in this research. However we leave the treatment of this for future work since other types of products (e.g., movies, consumer products) may have other unique characteristics or constraints.

References

"http://pay4tweet.com,

"http://socialnetworksolutionz.com/index.html,

Adamic, L.A., "Zipf, power-laws, and pareto-a ranking tutorial," *Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html*, (2000),

Adomavicius, G. and Y. Kwon, "Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques," *IEEE Transactions on Knowledge and Data Engineering*, (2011), 1-15.

Chen, Cheng, Kui Wu, Venkatesh Srinivasanand Xudong Zhang, "Battling the Internet Water Army: Detection of Hidden Paid Posters," *CoRR*, abs/1111.4297, (2011),

Dellarocas, C., "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science*, 52, 10, (2006), 1577-1593.

Deshpande, Mukund and George Karypis, "Item-based top-<i>N</i> recommendation algorithms," *ACM Trans. Inf. Syst.*, 22, 1, (2004), 143-177.

Easley, D. and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge Univ Pr, 2010.

Economist-a, The, "Mind for Netiquette, or we'll Mind it for you," *The Economist*, Issue Number, December 7, 2011 2011,

Economist-b, The, "Bulletins from the future," *The Economist (US)*, 400, Issue Number, 2011/07/09/ 2011, 3(US).

Eggenberger, F. and G. Pólya, "Über die Statistik verketteter Vorgänge," *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 3, 4, (1923), 279-289.

Fleder, Daniel and Kartik Hosanagar, "Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity," *Management Science*, 55, 5, (2009), 697-712.

Freedman, D.A., "Bernard Friedman's urn," *The Annals of Mathematical Statistics*, 36, 3, (1965), 956-970.

Kullback, S. and R.A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, 22, 1, (1951), 79-86.

Largillier, T., G. Peyronnetand S. Peyronnet, "SpotRank: a robust voting system for social news websites," *Proceedings of the The 4th workshop on Information Credibility*, 2010, 59-66.

Lee, J.S. and D. Zhu, "Shilling Attack Detection—A New Approach for a Trustworthy Recommender System," *INFORMS Journal on Computing*, 24, 1, (2012), 117-131.

Lerman, K., "User Participation in Social Media: Digg Study," *CoRR*, abs/0708.2414, (2007-a),

Lerman, K., "Social Information Processing in News Aggregation," *Internet Computing, IEEE*, 11, 6, (2007-b), 16-28.

Linden, G., B. Smithand J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *Internet Computing, IEEE*, 7, 1, (2003), 76-80.

Loretta, Chao and Spegele Brian, "Beijing Tightens Cyber Control," *Wall Street Journal*, December 17, 2011, 2011,

Maroulis, S., R. Guimerà, H. Petry, MJ Stringer, LM Gomez, LAN Amaraland U. Wilensky, "Complex systems view of educational policy research," *Science*, 330, 6000, (2010), 38-39.

Myers, D.J., "The Diffusion of Collective Violence: Infectiousness, Susceptibility, and Mass Media Networks1," *American Journal of Sociology*, 106, 1, (2000), 173-208.

Newman, M.E.J., "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, 46, 5, (2005), 323-351.

Phillips, DP, "The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect," *American Sociological Review*, 39, 3, (1974), 340.

Resnick, Paul and Rahul Sami, "The influence limiter: provably manipulation-resistant recommender systems," Minneapolis, MN, USA, (2007), 25-32.

Resnick, Paul and Rahul Sami, "The information cost of manipulation-resistance in recommender systems," Lausanne, Switzerland, (2008), 147-154.

Rogers, E.M., "New Product Adoption and Diffusion," *Journal of Consumer Research*, 2, 4, (1976), 290-301.

Salganik, Matthew J., Peter Sheridan Doddsand Duncan J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, 311, 5762, (2006), 854-856.

Schelling, T.C., "Dynamic models of segregation†," *Journal of mathematical sociology*, 1, 2, (1971), 143-186.

Van Roy, B. and X. Yan, "Manipulation robustness of collaborative filtering," *Management Science*, 56, 11, (2010), 1911.

Warren, J. and J. Jurgensen, "The Wizards of Buzz," *Wall Street Journal*, February 10, 2007, 2007,

Weber, Thomas E., "Cracking the New York Times Popularity Code," *The Daily Beast*, December 19, 2010,

Ziegler, Cai-Nicolas, Sean M. McNee, Joseph A. Konstanand Georg Lausen, "Improving recommendation lists through topic diversification," Chiba, Japan, (2005), 22-32.

Chapter 4 : Empirical Analysis of Outsourcing

Effects of IT Backgrounds of Project Owners on the Organizational Impacts of IT Outsourcing Projects

      Introduction. There has been a long standing debate on the issue of the value added by CIOs in a firm. For example, in one of the earliest articles on this issue, Earl and Feeny (1994) discuss both cases where IT is considered an "asset" or a "liability" in a firm. Further, they observe that "*the CIO's ability to add value is the biggest factor in determining whether the organization views IT as an asset or liability*". Subsequently more attention has been paid to investigate the role played by CIOs, as both the growing and the shrinking status of CIOs have been observed (Mateyaschuk, 1999; Overby, 2003). More recently, Luftman and Kempaiah (2008) have noted the large IT budgets managed by CIOs, and their contributions in shaping a firm's strategy.

      To identify the mechanisms by which CIOs (project owners with the tile of CIO or an executive with IT background) add value to firms, we address the following question in this chapter: Does the IT background of project owners who manage system integration outsourcing projects, affect firm performance in terms of cost savings, revenue, and profitability, through the projects they manage?

      System integration (SI) is a process of interlinking different software applications running on different hardware platforms. SI is common in the context of blended non-standard systems – the case often encountered in IT outsourcing.

Although firms enter into different kinds of outsourcing contracts such as application management, business process outsourcing, SI outsourcing, and network and desktop maintenance, SI outsourcing is most widespread both in terms of the volume and the number of contracts. Further, blended non-standard systems that require SI, call for special skills with deep technical background to be successful. This environment therefore provides great opportunity for the IT background of executives to impact project outcomes.

Related work. The importance of CIOs in an organization has been examined in the information systems (IS) literature through various approaches. Feeny, Edwards, and Simpson (1992) have used explanatory framework to improve the quality of CEO/CIO relationship with emphasis on the extensive IT background of CIOs. Armstrong and Sambamurthy (1999) have examined the influence of (i) quality of senior leadership, (ii) sophistication of IT infrastructures, and (iii) organizational size on IT assimilation. Their findings provide robust evidence on the impacts of CIOs' business and IT knowledge on IT assimilation.

Chatterjee, Richardson, and Zmud (2001) have used an event study methodology with capital asset pricing model to examine market reactions to the announcements of newly created CIO positions. They find strong support for positive reactions from the market place for the announcements of newly created CIO positions. Using knowledge-based and resource-based views, Armstrong and Sambamurthy (1999) argue that IT knowledge of senior leadership teams significantly enhance firm's IT assimilation. Based on the organizational studies literature, Bassellier, Benbasat, and Reich (2003) have

identified knowledge and experience both as important factors in determining the competency of business managers in IT.

The threats and opportunities related to the increasing involvement of non-IT managers in IT outsourcing process has been discussed by Gefen, Ragowsky, Licker, and Stern (2011). In the similar context, Westerman and Hunter (2007) have noted that non-IT managers are mostly unaware of the need to manage IT risks.

Banker, Hu, Pavlou, and Luftman (2011) have empirically examined the CIO reporting structure and suggest that this structure not be used as a standard for the strategic role of IT in a firm. They also observe the role of CIOs gradually becoming more influential as IT increasingly plays a pivotal role in a firm's success (Banker et al., 2011). Aral and Weill (2007) argue that different types of IT investments may impact different aspects of firm performance.

Research hypotheses. While prior research highlighted above has reported on the aggregate impacts of the CIO role, in this chapter, we look at the mechanisms by which CIOs contribute to organizations. As mentioned earlier, we do this by investigating the contributions of CIOs in managing system integration IT outsourcing projects. We hypothesize that in managing systems integration IT outsourcing projects, executives who have a background in IT, are better positioned to manage the complexities, and identify constraints and opportunities to meet various project goals. This knowledge improves their negotiation stance since technical knowledge is an important ingredient for effective monitoring,(Gore, Matsunaga, & Eric Yeung, 2011; Keen, 1991).

Cost reduction. Cost reduction is one of the main reasons for IT outsourcing. This is driven by the common belief that an outside vendor can provide better or same level of

service at lower costs (Smith, Mitra, & Narasimhan, 1998). Since vendors serve many clients, they incur lower per unit cost due to economies of scale, the benefits of which are often passed on to client firms in the case of competitive bidding.   For firms, cost efficiency is defined both in terms of operating expenses and overhead expenses. From empirical evidence in other domains suggesting that technical expertise is an important input for effective monitoring, it should follow that, the technical expertise of project owner in the relevant domain (IT skills for systems integration outsourcing) would impact the monitoring and negotiation skills of the project owner. The direct impact of such improved monitoring and negotiations is likely to be lower costs. Since IT costs are largely included as part of SG&A (selling, general and administrative) expenses, project ownership by IT executives should lead to a reduction in a firm's SG&A expenses compared to project ownership by executives with other backgrounds. Further, superior implementation is expected to simplify operations and improve operational efficiency, reducing operating costs. This leads to our first set of hypotheses:

H1a: Reduction in a firm's operating expenses will be greater for system integration outsourcing projects when a project owner is an executive with an IT background in comparison to an executive with a non-IT background.

H1b: Reduction in a firm's selling, general & administrative expenses will be greater for system integration outsourcing projects when the project owner is an executive with an IT background in comparison to an executive with a non-IT background.

Revenue. The contributions of IT outsourcing investments have been evaluated in terms of firm productivity (Han, Kauffman, & Nault, 2011; Loh & Venkatraman, 1992). Further, it has been established that IT outsourcing has positive contributions to the industrial output (Han et al., 2011). Project ownership of system integration IT outsourcing projects by IT executives can impact revenues in various ways. First, the deeper IT knowledge of the project owner is likely to lead to better project selection based on alignment with the organization's capabilities, superior requirements gathering and vendor selection, leading to better implementations, which could improve the organization's order-processing capabilities. Another channel for revenue enhancement comes from the deployment of cost savings achieved through superior project monitoring to other revenue enhancing IT projects. Hence, we propose our next hypothesis

H2: A firm's revenue will be greater when system integration outsourcing project owners are executives with an IT background in comparison to executives with a non-IT background.

Profitability. The first two hypotheses lead to our third hypothesis relating to firm profits. Well positioned outsourcing helps a firm to improve its profitability by staffing, capabilities, facilities and payroll (Jiang, Frazier, & Prater, 2006). By simultaneously lowering costs through superior monitoring of SI outsourcing projects and increasing revenues from deploying cost savings, CIOs can also improve firm profitability. Hence we propose the following research hypothesis for the profitability of a firm from systems integration outsourcing:

H3: A firm will experience higher profits when IT system-integration outsourcing projects are managed by executives with an IT background in comparison to executives with a non-IT background.

Data collection. The economic structure of a firm integrates the knowledge and skill sets of a variety of individuals in the production of value-added products and services. Following this view, we examine the changes in the financial characteristics of a firm before and after the IT systems integration outsourcing contract is signed. Consistent with prior research we use financial measures based on cost efficiency, revenues and profitability (Jiang et al., 2006; Smith et al., 1998).

We use data from various sources to test our hypotheses. Information about the size of SI outsourcing contract, customer (client) name, project owner title, customer industry, contract start date and contract duration are obtained from IDC BuyerPulse Deals Database. The IDC database, which is one of the largest repositories for outsourcing contracts signed in the US, maintains records of different outsourcing contracts announced by firms as well as contract related data. It is used by IT firms to generate leads by identifying expiring contracts.

We obtain financial measures for each publicly traded firm whose contract information is contained in the IDC database from COMPUSTAT. All measures are adjusted to 2005 constant dollars using implicit price deflators. The data for implicit price deflators and the economy wide gross domestic product (GDP) is obtained from the Bureau of Economic Analysis (BEA). All quantities are expressed in million dollars.

The total volume of system integration outsourcing projects (including governmental and privately held firms) signed since 1995 is seen in Figure 4.1.
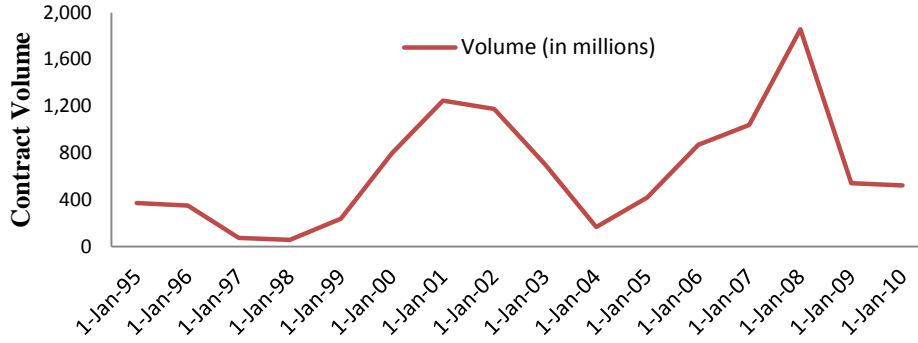
Figure 4.1. Total of Volume of System Integration Contracts Signed During 1995-2010

Clearly, it can be observed that the volume of SI outsourcing contract increased significantly over the last decade. Two troughs since 2000 correspond to the time periods immediately after the tech-bubble and the 2008 financial crisis.

Explanatory variables. Descriptions on the operationalization of explanatory variables follow in this section. For all the measures described below, subscript t denotes the year in which an outsourcing contract was signed.

Contract Value: The dollar amount of a specific contract is divided by the contract length and then summed over all the active contracts a firm had in a given year. Hence, the total contract value for a firm in a year is given by:

$$contract\ value = \sum_{j=all\ contract} \left(\frac{Value\ of\ outsourcing\ contract}{Length\ of\ contract\ in\ years} \right)_j \qquad (1)$$

Project Owner: To examine the contribution of executives with IT-background in SI project management, we divide project owners in two groups: (a) executives with IT background and (b) executives with non-IT background. Project owners with title such as CIO, CTO, IT-manager, IS/IT director are identified as executives with IT background.

Whereas project owners with titles such as chief financial officer (CFO), human resource (HR), chief marketing officer (CMO) are identified as executives with non-IT background (Zhu, Kraemer, & Dedrick, 2004).

Control variables. The following control variables are used.

Firm Size**:** Firm size is usually operationalized using the number of employees or firm revenue. In this chapter we used natural log of annual firm revenue (Whitaker, Mithas, & Krishnan, 2010).

Industry Sector: Following the approach of Whitaker et al. (2010) different industry sector considered are finance, services, trade and logistics and other industrial. Sectors are coded using indicator variables in the regression equation with services-sector being treated as the base category.

Change in GDP: The economy wide exogenous factors such as government policies, and state of the economy, recession are one of the strongest determinants of firm performance in our sample. To control for this effect we use the natural log of the relative change in GDP. This quantity is obtained from BEA in chained 2005 dollars.

Table 4.1. Variable Description and Data Sources

Variable	Source	Operationalization	Deflator
Change in operating-expenses	COMPUSTAT	Natural log of relative change in operating expenses	2005 implicit price deflators from BEA
Change in overhead-expenses	COMPUSTAT	Natural log of relative change in overhead expenses	2005 implicit price deflators from BEA
Change in revenue	COMPUSTAT	Natural log of relative change in sales	
Sales	COMPUSTAT	Natural log of sales	2005 implicit price deflators from BEA
Change in profitability	COMPUSTAT	Natural log of relative change in net income	2005 implicit price deflators from BEA
Contract value	IDC	Natural log of contract value-defined in the equation 1	2005 implicit price deflators from BEA
Contract owner	IDC	Binary variable	
Contract duration	IDC	Months converted in years	2005 implicit price deflators from BEA
Contract size	IDC	In 2005 constant dollars (in millions)	

Analysis and results. We used log-transformation of variables for the regression

analysis. The regression equation for the hypothesis H1a is given by the equation:

$$\ln\left(\frac{operating\ expense_{t+1}}{operating\ expense_{t-1}}\right)$$

$$= \beta_{00} + \beta_{01} * \ln(contract\ value_t) + \beta_{02} * \ln(sales_t) + \beta_{03} * \ln\left(\frac{GDP_{t+1}}{GDP_{t-1}}\right)$$

$$+ \ \beta_{04} * Y_1 + \beta_{05} * sector + \epsilon_0 \quad (2)$$

In the above regression equations Y_1 takes following binary values:

$$Y_1 = \begin{cases} 1, & if\ project\ owner\ has\ an\ IT\ background \\ 0, & if\ project\ owner\ has\ a\ non-IT\ backgound \end{cases}$$

Similarly, for the hypotheses H1b, H2 and H3 the expression of the dependent variables take the form as in equation 2, replacing operating expenses by SG&A, sales and profit respectively.

Table 4.2. Summary Statistics

Variable	Sample Size	Mean	Standard Deviation
ln(Op. Expense Ratio)	112	0.028	0.296
ln(SG&A Expense Ratio)	74	0.047	0.275
ln(Contract Value)	112	1.41	1.241
ln(sale)	112	8.144	1.695
ln(GDP Ratio)	112	0.006	0.034
ln(Profit Ratio)	83	0.067	0.828
ln(Sales Ratio)	112	0.005	0.311

Descriptive statistics. The dataset goes back to 1994, but is comprehensive beginning 1995. We therefore start with a sample of all 1317 system integration contracts signed between the time periods 1995-2010 in the IDC database. Out of 1317 data points, 298 observations are usable based on accounting measures obtained from COMPUSTAT. The sample size is further reduced to 112 due to incomplete project owner title data. Descriptive statistics of the quantitative variables used in the regression model are given below:

In the case of overhead expenses (SG&A), our sample size reduced to 74 observations, as some of the firms (especially in finance) do not report SG&A separately from the overall expenses (operating expenses).

In the regression model for profit (Table 4.7), it should be noted that the dependent variable is not defined when numerator and denominator have opposite signs (positive or negative). Hence, the sample is reduced to include only those cases where the dependent variable is defined. This reduced the sample size from 112 to 83 for H3.

The frequency distribution of the binary variable for the project owner is also examined for each level of sector. The frequency table can be found in the appendix 3.

Table 4.3. Correlation Matrix

	Ln (SG&A r)	ln(Profit r)	ln(Op. Exp.r)	ln(Sales r)	ln(Contract Val)	ln(sale)
ln(SG&A r)						
ln(Profit r)	0.159 (0.238)					
ln(Op. Exp. r)	0.795 (<.001)	0.213 (0.053)				
ln(Sales r)	0.726 (<.001)	0.324 (0.003)	0.898 (<.001)			
ln(Contract Val)	-0.117 (0.32)	-0.104 (0.351)	0.026 (0.784)	-0.045 (0.634)		
ln(sale)	0.007 (0.951)	-0.078 (0.484)	0.120 (0.207)	0.016 (0.867)	0.494 (<.0001)	
ln(GDP r)	0.337 (0.003)	0.281 (0.01)	0.268 (0.004)	0.323 (0.0005)	-0.01843 (0.847)	0.184 (0.051)

The parameters of individual equations are initially estimated using ordinary least squares (OLS). Standard assumptions of OLS are examined for each of the regression models.

The analysis is based on economy wide data with high degree of variation in contracts and firm sizes. Hence, OLS estimates could yield inaccurate estimates of the regression coefficients due to the influence of low-probability events when sample size was less than 400 (Starbuck, 2006).

To address this issue, robust MM regression has been used to limit the effects of extreme outliers (Yohai, 1987). In absence of extreme outliers robust MM regression produces the same coefficients as OLS. Thus, we also use robust MM estimates to correct any possible inconsistencies of OLS estimates. Robust MM regression uses an M-scale estimate to scale the regression residuals. In most cases, both OLS and robust MM estimates are similar in sign and magnitude.

Table 4.4. Parameter Estimates for H1a (n=111)

Variable	Parameter	OLS Estimate	Robust-MM Estimate
Intercept	β_{00}	-0.037	0.067
		(0.771)	(0.526)
ln(Contract Value)	β_{01}	0.004	0.002
		(0.865)	(0.919)
Project owner (Y_1)	**β_{04}**	**-0.110****	**-0.075***
		(0.033)	**(0.079)**
Control variable			
ln(sale)	β_{02}	0.033*	0.016
		(0.054)	(0.275)
ln(GDP)	β_{03}	2.013***	2.123***
		(0.008)	(0.0005)
Financial		-0.18**	-0.127**
		(0.016)	(0.043)
Trade and Logistics		-0.287***	-0.205***
		(0.0001)	(0.0009)
Other industrial		-0.172**	-0.186***
		(0.027)	(0.004)
R^2		0.23	0.13
R^2(adj)		0.17	
F (model)		4.30	
		(0.0003)	

$$\ln\left(\frac{operating\ expense_{t+1}}{operating\ expense_{t-1}}\right)$$

$$= \beta_{00} + \beta_{01} * \ln(contract\ value_t) + \beta_{02} * \ln(sales_t) + \beta_{03} * \ln\left(\frac{GDP_{t+1}}{GDP_{t-1}}\right) + \beta_{04} * Y_1 + \beta_{05}$$

$$* sector + \epsilon_0$$

Note: For all the regression results discussed in this paper we used these notations: $*** = p < 0.01, ** = p < 0.05, * = p < 0.1$

The assumption of normality is checked using the Shapiro-Wilk's test for residuals (Shapiro & Wilk, 1965). In some cases this is rejected at the 5% significance level. No evidence of heteroscedasticity is found in all the models using White's test (White, 1980). The effect of multicollinearity is examined using variance inflation factor (VIF). In all cases, VIF is well within the suggested the limit of 5. Influential outliers are detected using Cook's distance statistic (Cook & Weisberg, 1982). In cases, where the removal of an outlier does not bring significant changes in the regression estimate, the estimates are determined using the complete sample. In some cases, outlier indication is

complemented with exogenous factors; wherever necessary, these cases are discussed in detail for each regression model.

Discussion.

Cost efficiency model (operating expenses). We started with the complete sample to examine the impact of outsourcing contracts and project owners on a firm's operating expenses (equation 2). During OLS analysis, two observations were marked as outliers due to high cook's distance. The closer examination of these two observations revealed that they had significant decreases in the operating expenses in the corresponding year.

For one of the observations– in the year SI outsourcing contract was signed, a "significant" decision on corporate spin-off was also taken. So, it was not included for further analysis. For the other observation, no such "significant" event was found for the corresponding year, so it was retained in the sample for the analysis. This resulted in a sample of size 111 (Table 4.4).

Project owners with IT background were found to play a significant role in reducing the operating expenses in comparison to project owners without any IT background. Thus we could infer that project owners with IT background were better positioned than project owners with non-IT background to meet their firm's goal of lowering expenses while managing SI outsourcing contracts. However, we did not find significant results for the impact of contract value on operating expenses.

Table 4.5. Parameter Estimates for H1b (n=73)

Variable	Parameter	OLS Estimate	Robust-MM Estimate
Intercept	β_{10}	0.151 (0.256)	0.186 (0.054)*
ln(Contract Value)	β_{11}	-0.026 (0.264)	-0.014 (0.41)
Project owner (Y_1)	$\boldsymbol{\beta_{14}}$	**-0.110** (0.045)	**-0.08** (0.045)
Control variable			
ln(sale)	β_{12}	0.016 (0.361)	-0.001 (0.919)
ln(GDP)	β_{13}	3.006*** (0.0005)	1.320** (0.031)
Financial		-0.118 (0.134)	-0.043 (0.481)
Trade and Logistics		-0.269*** (0.001)	-0.171*** (0.007)
Other industrial		-0.218*** (0.005)	-0.117** (0.048)
R^2		0.32	0.14
R^2 (adj)		0.25	
F (model)		4.36 (0.0005)	

$$\ln\left(\frac{SG\&A_{t+1}}{SG\&A_{t-1}}\right) = \beta_{10} + \beta_{11} * \ln(contract\ value_t) + \beta_{12} * \ln(sales_t) + \beta_{13} * \ln\left(\frac{GDP_{t+1}}{GDP_{t-1}}\right) + \beta_{14} * Y_1 + \beta_{15} * sector + \epsilon_1$$

Overhead expenses (SG&A expenses). For the test of hypothesis H1b, during OLS analysis, again the same observation dropped above was marked as an outlier due to significant decrease in SG&A expenses. Because of the reason mentioned earlier, it was not included for analysis. This led to a reduced sample size of 73 (Table 4.5).

Similar to the case earlier, we found that the variable project-owner, was significant (having stronger significance than for H1a). Thereby suggesting that the IT background of project owners could be very effective in managing and monitoring SI projects, thus bringing down the overhead expenses of firms.

Sales. To test hypothesis H2, we examined the impact of project owner on sales (Table 4.6). The residual analysis of OLS regression indicated high Cook's distance

statistics for one particular observation, which was then marked as an outlier. We performed regression analysis excluding this particular observation, but the overall statistical results were similar to the complete sample. Therefore, we reported results for hypothesis H2 based on the complete sample. In this case, effects of both explanatory variables (contract value and project owner) were not consistently significant for both OLS and robust MM regression. This suggested that overall firm revenue depended on various factors apart from outsourcing contracts. The IT background of project owner did help to bring down expenses, but their contribution to increase their firm revenues was not significant.

Profitability. For the OLS regression estimates two observations were marked with very high Cook's distance statistics. The year in which outsourcing contract was signed, the firm corresponding to one of the observations, had gone through a merger which was completed by 2007 (contract year - 2007). So, we excluded this observation from our data analysis.

Table 4.6. Parameter Estimates for H2 (n=112)

Variable	Parameter	OLS Estimate	Robust-MM Estimate
Intercept	β_{20}	0.124	0.121
		(0.395)	(0.266)
ln(Contract Value)	β_{21}	-0.073	-0.0008
		(0.21)	(0.97)
Project owner (Y_1)	β_{24}	**-0.073***	**-0.03**
		(0.058)	**(0.47)**
Control variable			
ln(sale)	β_{22}	0.007	0.003
		(0.726)	(0.86)
ln(GDP)	β_{23}	2.81***	2.6***
		(0.0013)	(<.0001)
Financial		-0.236***	-0.181***
		(0.006)	(0.005)
Trade and Logistics		-0.25***	-0.172***
		(0.003)	(0.006)
Other industrial		-0.093	-0.1428**
		(0.294)	(0.0334)
R^2		0.21	0.14
R^2 (adj)		0.15	
F (model)		3.86	
		(0.0009)	

$$\ln\left(\frac{Sales_{t+1}}{Sales_{t-1}}\right) = \beta_{20} + \beta_{21} * \ln(contract\ value_t) + \beta_{22} * \ln(sales_t) + \beta_{23} * \ln\left(\frac{GDP_{t+1}}{GDP_{t-1}}\right) + \beta_{24} * Y_1 + \beta_{25} * sector + \epsilon_2$$

The second observation was in a quarter when the corresponding firm took a big write-down. Hence, this observation was also excluded from further analysis. The following results were observed after eliminating these two observations from regression analysis (a) a significant increase in global F-value (significant at p-value=0.05; Table 4.7) and (b) improvement in adj-R^2 without any threat of multi-collinearity. Results based on this modified sample have been presented in Table 4.7. Again the contributions of contract value was not significant. However, surprisingly we found negative significant coefficient corresponding to project owner. Project owners of IT systems integration projects with non-IT backgrounds significantly improved firm profitability relative to project managers with IT backgrounds.

Table 4.7. Parameter Estimates for H3 (n=81)

Variable	Parameter	OLS Estimate	Robust-MM Estimate
Intercept	β_{30}	0.476	0.49
		(0.238)	(0.15)
ln(Contract Value)	β_{31}	-0.087	-0.043
		(0.21)	(0.453)
Project owner (Y$_1$)	$\boldsymbol{\beta_{34}}$	**-0.304***	**-0.34****
		(0.069)	**(0.018)**
Control variable			
ln(sale)	β_{32}	-0.03	-0.029
		(0.592)	(0.537)
ln(GDP)	β_{33}	5.85**	7.24***
		(0.021)	(0.0007)
Financial		-0.317	-0.31
		(0.169)	(0.11)
Trade and Logistics		0.098	-0.035
		(0.67)	(0.86)
Other industrial		0.25	0.019
		(0.31)	(0.93)
R^2		0.19	0.17
R^2 (adj)		0.12	
F (model)		2.48	
		(0.02)	

$$\ln\left(\frac{profit_{t+1}}{profit_{t-1}}\right) = \beta_{30} + \beta_{31} * \ln(contract\ value) + \beta_{32} * \ln(sales) + \beta_{33} * \ln\left(\frac{GDP_{t+1}}{GDP_{t-1}}\right) + \beta_{34} * Y_1 + \beta_{35} * sector + \epsilon_3$$

Regarding control variables, change in GDP ratio was found significant in almost all cases with effects of sector being mixed and firm size not being significant in most cases.

Conclusion and discussion. The present research has been led by the long standing question that was posed around two decades before, i.e., does CIO add value (Earl & Feeny, 1994)? Although our results were established in a slightly broader context to include project owners with IT background; we believe the present research makes valuable contribution related to the mechanisms by which the IT backgrounds of project owners impact firm performance metrics.

We found that the IT background of project owner of IT systems integration projects did play an important role in reducing costs. But, no significant results were found for revenues, and adversely affected the profitability of firms relative to executives with non-IT backgrounds.

Our findings resonates with prior suggestions made by IS researchers regarding the increased importance of IT in organizations (Bassellier et al., 2003). In support of IT (or technical) background of project owners Rockart, Earl, and Ross (1996) state:

"The success or failure of an organization's use of IT, however, is only partially dependent on the effectiveness of the IT organization. It is even more dependent on the capability of line managers at all levels to understand the capabilities of the IT resource and to use it effectively."

The management of IT is a well-recognized challenge. CEOs often face decisions as to how to structure the IT function, including the new role for a CIO (Kambil & Lucas, 2002). The present research addresses the IT leadership impacts on cost efficiency,

revenue and profitability. The IT background of project owner does play a major role in reducing costs in an increasingly competitive business environment.

However, in case of profitability and revenue, our findings were similar to that of Aral and Weill (2007), where they found negative or non-significant relationship with IT investments on profitability. Among possible reasons for the finding, we conjecture that it may be related to project selection. For example, finance executives may be leading projects with direct impacts on financial systems and marketing executives may be leading projects with direct impacts on marketing. But IT executives may be leading more general projects, for example those related to identity management, messaging, and billing. Unfortunately, our dataset did not allow for such identification of projects.

To conclude, our findings suggest that systems integration and other complex IT outsourcing contracts that are motivated by cost-reduction concerns should be managed by CIOs or executives with IT background. This also calls for a need to provide appropriate IT education to executives, to make effective business decisions regarding IT.

Modeling Outsourcing Decisions: An Empirical Analysis of Outsourcing in the US Auto Industry

Introduction. Information technology enabled services (ITES) outsourcing is a wide-spread business phenomenon, with the global market for outsourcing of business and technology services reaching $315 billion in revenues in 2011, as per Gartner Inc. (December 2011). The worldwide IT outsourcing market has grown consistently over the last few years. In a recent survey conducted by InformationWeek (Murphy 2012), only 4% of the 513 business professionals said that they have any plans to decrease their use

of IT outsourcing. While, 17% were weighing their options, and the remaining 79% were maintaining or increasing their level of outsourcing. The increasing trend towards IT outsourcing has been further propelled by the emergence of cloud computing and offshoring activities, with US firms on average spending 14% of their IT budgets on IT outsourcing activities (Han et al. 2013).

In an increasingly globalized and competitive business environment, firms experience increased competition in product, service and labor markets with a continuous requirement to adapt to new markets and technologies (Slaughter et al. 1996). In this context, efficient allocation of resources for IT outsourcing is important because, it can free-up a firm's IT staff for new development. Also, outsourcing can be a potential driver of cost reduction by availing vendors' production cost advantage. In a survey published by AMR Research, it has been noted that a majority of outsourcing contracts are driven by cost reduction targets (Fersht et al. 2009; Han et al. 2013).

In prior research, various theoretical perspective and methodologies have been used to study outsourcing at individual, project, firm and economy levels (Dibbern et al. 2004; Whitaker et al. 2010). Whitaker et al. (2010) have presented a survey on various theoretical perspectives, and different levels of analysis in IS outsourcing research. The theoretical perspectives used in the literature are: Transaction Cost Economics, Agency Theory, Theory of Production, Competitive Strategy, Modularity, Learning and Capabilities Views, and Systems Dynamics. The detailed discussions on various approaches have been provided later.

Although outsourcing has been widely studied in the IS literature, including being modeled as a diffusion process (Loh et al. 1992b), we use a new perspective to

study outsourcing, based on herding behavior. This perspective is based on an interesting observation that CIO decisions are often driven by the objective to maintain industry averages of various financial measures (McDonald 2010).

In this research, we explore the dynamics of ITES outsourcing in the automotive sector. This sector has gone through major structural changes in the last decade due to various bankruptcy filings, government bailouts and several other austerity measures. In the automotive sector in particular, outsourcing has been one of the most widely practiced business strategies to bring down IT costs (Dunn 2005; Techweb 2009).

We model the impact of outsourcing activities of firms using a two-step regression approach. Our approach integrates the impact of peer pressure on firms to undertake outsourcing activities. In our first step, we predict outsourcing decision of firms, which is then used to predict their selling, general and administrative (SG&A) expenses in the second step. The effect of imitative behavior is mediated through the outsourcing decision taken by a firm, on its SG&A expenses. To our knowledge, peer-pressure to model firm behavior is a unique contribution to the outsourcing literature.

This research builds on prior research on conformance behavior, using the perspectives of Information Based Imitation (Lieberman et al. 2006). According to information-based imitation, firms follow other firms that are perceived to have superior information.

Literature review. Outsourcing has been extensively studied in the research literature. A review of relevant literature is presented below. Loh et al. (1992a) have investigated the determinants of IT outsourcing using business and IT competences as represented by various accounting and economic measures. Using factor analysis and

multiple-regressions, they find business and IT cost structures to be positively related to the degree of outsourcing of a firm, while not observing any significant relationship between financial leverage and business performance. Smith et al. (1998) examine the financial characteristics of firms that enter into large scale outsourcing contracts. The impact of outsourcing contracts have been examined for various accounting measures including profitability and SG&A expenses. Chaudhury et al. (1995) investigate the process of IT outsourcing, and the various stages involved in it. Considering cost reduction as a driving force for outsourcing, they propose a bidding mechanism to reduce expected outsourcing costs in the final bidding and vendor selection process.

Slaughter et al. (1996) have used a labor-market perspective to explain IS outsourcing as a response of firms, on the face of increasing costs, and changing technological landscape. Soon et al. (1998) empirically investigate the economic determinants of IS outsourcing, in terms of production cost, transaction cost and financial slack in the context of U. S. banks. Production cost advantages and transaction costs were found as major determinants of IS outsourcing.

In the context of buyer-supplier relationship, Bakos et al. (1993) use economic theory of incomplete contracts to determine the optimal strategy of a buyer to choose the number of suppliers for IT services. Koh et al. (2004) discuss the factors responsible for the success of IT outsourcing contracts. They explore supplier and customer perspectives through the lens of psychological contract of customer and supplier project managers. Mithas et al. (2007) assess the influence of non-contractibility on buyers' use of reverse auctions for supplier relationship in the context of outsourcing. In another instance of vendor-client relationship, Cha et al. (2009) examine how knowledge parameters

characterizing a sourcing relationship between a vendor and a client, interact with production and coordination costs, in affecting the business value of alternative outsourcing strategies. This is then used to determine a firm's optimal rate of outsourcing.

Lee et al. (2004) explore the effects of IT outsourcing strategies on outsourcing success. They identify three dimensions of outsourcing strategies based on residual rights theory: degree of integration, allocation of control, and performance period.

A system dynamics approach is used by Dutta et al. (2005) to explore the mechanics by which different factors interact to produce the observed growth in IT offshoring. They use a computational model for a two-country simulation model of offshoring. Han et al. (2011) use economy level panel data to evaluate the contributions of spending in IT outsourcing using a production function framework, and find IT outsourcing to make a positive and economically meaningful contribution to industry output and labor productivity.

The impact of the choice of sourcing mechanism on the relation between the modularization of business processes and their underlying IT support infrastructure has been investigated by Tanriverdi et al. (2007). Their empirical analysis of large and medium size U.S. firms reveals that domestic outsourcing is preferred for high modularity processes, whereas offshore outsourcing is preferred for processes that are low in modularity.

Ramasubbu et al. (2008) propose a learning-based mediated model of offshore software project productivity and quality. In a related study, Harter et al. (2003)

investigate the impact of software process improvements on infrastructure costs, mediated through software quality.

Chen et al. (2009) use a comprehensive coding scheme to capture contract provisions in terms of monitoring, dispute resolution, property rights protection, and contingency provisions. They investigate the effects of transactional and relational characteristics on the specific contractual provisions through transaction cost, agency, and relational exchange theories.

Dey et al. (2010) present a contract-theoretic model to design software outsourcing contracts, to explore benefits of fixed-price contracts, and time-and-material contracts. They also investigate quality-level agreements and profit-sharing contracts. Fitoussi et al. (2012) use the contract-theory framework to examine how objectives and incentives are related in IT outsourcing contracts. Using a dataset of outsourcing contracts, Susarla et al. (2010) examine whether extensiveness of detailed contracts can alleviate *holdup*, where holdup is described as underinvestment and inefficient bargaining by vendor as a result of relationship-specific investment and contract incompleteness.

In the context of business process outsourcing (BPO), Mani et al. (2010) use the lens of information processing view of firms to theorize heterogeneity across BPO exchanges, as a function of information capabilities that fit the unique information requirements of the exchange. They also provide recommendations on some best practices on BPO design and management. Whitaker et al. (2010) use organizational learning and capabilities to develop a conceptual model of firm-level characteristics that facilitate onshore and offshore BPO.

The impact of IT outsourcing on non-IT operating cost has been examined by Han et al. (2013) using a framework where internal IT investments moderate the relationship between IT outsourcing and non-IT operating costs. Using a panel dataset during the period 1999-2003, they find IT outsourcing to reduce non-IT operating costs in firms such as SG&A. They also find that higher levels of complementary investments in internal IT leads to higher reductions in non-IT operating costs.

Specific to the automotive industry, Mukhopadhyay et al. (1995) find the use of Electronic Data Interchange (EDI) improved savings per vehicle.

From the above review of literature, it could be easily seen that the impact of peer pressure on firms to outsource has not been investigated, given that best practices often evolve by observing and imitating the behaviors of successful peers. This research uses this perspective to make a novel contribution to the literature.

Theoretical framework. Our hypotheses build on prior results of firm profitability, imitation and scale economies. Increased profits enable the firm to invest in the changes necessary to expand outsourcing. Peer imitation encourages firms to increase or decrease outsourcing, and economies of scale enable outsourcing firms to offer services at low costs, which re-inforces the use of outsourcing.

Firm profits. Firms in competitive environments are under pressure to innovate and defend their competitive positions. However, innovations are expensive. The greater the profitability of a firm, the greater its ability to take up innovative projects (Audretsch 1995; Branch 1974), which allow the firm to launch new products, or develop new business processes. Firms with lower profits or losses are likely to try to focus on

lowering costs, shedding unviable business lines and finding ways to streamline their existing operations (Chastain 1984).

However, in highly competitive industries such as the auto industry, where profits fluctuate from year to year, firms are likely to be reluctant to hire new personnel to staff these new projects until the viability of the projects have been demonstrated. Outsourcing is a very effective way to staff such new projects (Gilley et al. 2000). Firms can use the expertise of outsourcing firms to implement projects quickly, deploy them and evaluate their viability. The most successful of such projects, which lead to identifiable long-term improvements in business metrics, are likely to lead to the hiring of permanent workers. We therefore hypothesize that:

H1: Increased profitability of a firm in a time period, is associated with increased outsourcing in the subsequent time period.

Imitation. The notion of Individuals seeking to conform to the behavior of reference groups has been widely used in the economics literature (Benabou 1996; Brock et al. 2001; Schelling 1971). Firms imitate each other in the introduction of new services, product and processes, in organizational practices and in new investments. Those successfully imitating good practices remain competitive and those with inferior performance do not survive the competition. Further, both imitation and innovation guide firms towards the strategies of best payoffs (Lee et al. 2002). Business imitation has been explained using Information-based Theory by Lieberman et al. (2006). A quick overview of this theory is presented below.

Information Based Imitation. Information based imitation explain an environment where managers face uncertainty related to cause-effect relationships and

are often unable to assess the full range of possible outcomes (Lieberman et al. 2006). In such an environment, managers use information implicit in others' action, and may imitate others. Imitating others can be understood as a rational behavior, because decisions of others may reflect information that one doesn't have. This is widely practiced in an environment characterized by uncertainty and ambiguity.

Doing what everybody else is doing is termed as herding behavior. A simple model is presented by Banerjee (1992). The economic theory of herding behavior is cast in terms of information cascades and social learning (Banerjee 1992; Lieberman et al. 2006). In the context of outsourcing, information cascade could occur when a firm observing the decision of profitable firms in the peer set, follows their strategy of outsourcing activity (Bikhchandani et al. 1992). Bikhchandani et al. (1992) argue that informational cascade can lead to conformity of behavior among agents. In some cases, private information of an individual and her prior experience causes her to imitate the action of others (Bikhchandani et al. 1992).

Social and economic decisions are often influenced by what others are doing around us (Banerjee 1992). This interdependence between different decision makers is termed as social influence or social learning (Lopez-Pintado et al. 2008). Social learning which results in herding or conformity behavior is observed in social, psychological and economic phenomena.

In a phenomenon driven by information based imitation, each agent possesses some private information about the state of nature. First an agent acts based purely on her private information, but the agent's actions reveal information to observers. As revealed information accumulates, it becomes more rational for observers to discard their private

128

information and mimic the decisions of others (Lieberman et al. 2006). Imitation can also be triggered by following the practices of larger rivals and firms with superior performances. According to information based imitation, firms with poor performance will be tempted to imitate the outsourcing practices of firms with superior performances.

Imitation is also a response designed to mitigate competitive rivalry (Lieberman et al. 2006). In highly competitive industries, the product and process strategies adopted by profitable rivals have a higher likelihood of being optimal strategies. Differentiation in these markets is likely to lead to sub-optimal outcomes due to uncertainty and unpredictability. This is one reason why imitation to mitigate rivalry is most observed when firms compete with each other in terms of comparable resources and markets (Lieberman et al. 2006).

Thus, it is hypothesized that firms imitate rivals to maintain competitive parity and to gain information from others' decisions. So the extent of outsourcing adopted by other firms within the industry at any given time is taken as optimal level of outsourcing within the industry at that time, that would then be imitated. This is expected to manifest as a tendency for firms to shift their levels of outsourcing towards industry mean levels of outsourcing. We therefore hypothesize that a firm will benchmark its operations against industry standards, and will conform to industry best practices on outsourcing levels in order to benefit from information conveyed by the competitive environment. The more a firm deviates from these levels, the higher the pressure for conformance with industry norms and practices felt by a firm. Hence, we hypothesize that:

H2: The deviation of a firm's outsourcing expenses as a proportion of SG&A expense (OESGA) from the industry average of OESGA at any given time is inversely related to the change in outsourcing activity by the firm.

Firms often outsource to get the benefits from economies of scale accruing to vendors (Han et al. 2011). Due to economies of scale and learning effects (Zimmerman 1982), vendors can provide IT services to a client at a lower cost than a firm's internal IT department (Han et al. 2011). IT service providers typically serve many clients, so they have the opportunity to achieve lower unit costs compared to a single company by leveraging fixed costs and achieving economies of scale (Bryce et al. 1998). Since firms can achieve immediate cost advantages through IT outsourcing by increasing the operational efficiencies of existing processes (Han et al. 2013), it is often considered a promising strategy to improve financial performance (Jiang et al. 2006). Reduction in non-IT operating expenses, as a result of IT outsourcing has been established previously by Han et al. (2013) using a panel data set of approximately 300 U.S. firms from 1999 to 2003. IT services related expenses of a firm are largely included as a part of SG&A expenses. Hence we hypothesize that

H3: Outsourcing investments of a firm is associated with a decrease in its SG&A expenses.

Empirical model. We use a two stage regression model. Using the model in stage 1, we first estimate a firm's outsourcing action for the current period, and then apply the second stage model to predict the SG&A expense for firms, based on predicted outsourcing action. In the first model, we represent action of a firm as a function of its profitability and pressure it experiences, in order to maintain the comparable level of

130

OESGA in the industry. The second model relates, SG&A expenses of a firm to its current outsourcing level controlling for other factors. Hence, the effects of imitation on a firm's SG&A expense are mediated through the actions taken by the firm. While a firm's decision to increase or decrease its outsourcing level is directly influenced by its profit and the mean industry OESGA, these factors only indirectly influence the SG&A expenses in our model.

The model of individual actions to increase or decrease the levels of outsourcing for conforming to the mean industry level, builds on the framework provided by (Brock et al. 2001), in which, an individual's action depends directly on the choices of others. The presence of interaction among firms induces a tendency for conformity in the behavior across agents, in the given reference group. Individual actions are also driven by intrinsic factors that differ across agents due to the heterogeneity of individual characteristics (Bernheim 1994), captured through individual profitability. The interplay of individual heterogeneity and interaction can give rise to complicated behavior of the system (Brock et al. 2001).

The notations used in our model are given in Table 4.8.

Table 4.8. Explanation of Mathematical Notations

Notation	Explanation
F	Index set of agents with cardinality N
S_{it}	State of agent i in period t (such as outsourcing level)
a_{it}	Action of agent i in period t
	(*only those observations in which non-zero change in outsourcing level takes place*)
$SG\&A_{it}$	Selling, general and administrative expenses of agent i in period t
ξ_{it}	Peer pressure (mean *OESGA*) experienced by agent i in period t

Basic unit of analysis is a firm in the automotive industry. The state of firms is described by a $N \times 1$ state vector, s_t for N firms, where s_{it} denotes the state (i.e., outsourcing level) of the ith agent (firm) at time t. At every time period t, an agent i chooses an action: to update its state. We define an action a_{it} of an agent i at time period t as the change in the state (i.e., non-zero change in outsourcing level) of the agent from the previous time period, $t - 1$. In other words, if between two consecutive years, the state of an agent doesn't change, then the agent does not take action during that period of observation.

In (1), we model an agent's actions as a function of its previous profit-margin and peer pressure the agent experiences due to outsourcing activities of other firms. Firms often imitate other firms in the industry, especially when the imitation relates to adopting some of the best practices. As an example, a firm's decision to outsource a business function could be driven by the success stories of outsourcing experienced by similar firms.

The peer pressure term in (1) can be understood as the deviation from the mean OESGA of peers.

$$a_{it} = \beta_0 + \beta_1 * \frac{profit_{it-1}}{revenue_{it-1}} + \beta_2 * \left(\frac{S_{it-1}}{SG\&A_{it-1}} - \xi_{it-1} \right) + \epsilon \qquad (1)$$

We conceptualize this effect of peer pressure for an agent i, as per the following expression:

$$\xi_{it} = \frac{1}{N-1} \left(\sum_{\substack{i \in F \\ i \neq j}} \frac{S_{it}}{SG\&A_{it}} \right) \qquad (2)$$

Action corresponds to non-zero observations and is given by,

132

$$a_{it} = \begin{cases} S_{it} - S_{it-1} & S_{it-1} > 0 \\ S_{it}(entry) & S_{it-1} = 0 \end{cases} \quad (3)$$

SG&A expenses of firms are estimated using the following equation:

$$\ln(SG\&A_{it+1}) = \gamma_0 + \gamma_1 * \ln(SG\&A_{it}) + \gamma_2 * \ln(S_{it} + a_{it}) + \gamma_3 * \ln(GDP_t) + \epsilon \quad (4)$$

A log-linear specification has been used to incorporate the effects of economies of scale experienced by a client through outsourcing contracts (Harter et al. 2003; Hitt et al. 2002). In (4), we control for previous SG&A level and GDP. The parameter estimates obtained from (1) have been used to estimate a firm's action in (4).

Data and measures. Data on outsourcing contracts of firms in the automotive sector was obtained from IDC BuyerPulse deals database, which is the largest database for outsourcing contracts and has data on outsourcing contracts signed by US companies during the period of: 1995 − 2010 (reference). In this study, we use data for automotive firms from the database, which enter into multiple outsourcing contracts during the aforementioned period. Different kinds of contracts signed by firms include but not limited to: Application Management, Business Process Outsourcing, Custom Application Development, IS Outsourcing, Network and Desktop Outsourcing and System Integration.

To supplement the data from IDC database, news wire services at LexisNexis was also searched during the time period of 1995-2011, using terms: "information technology", "information systems", "outsource", "outsourcing" and "company Name". In total, there were 107 contracts.

The data from IDC and news wire search was supplemented with accounting measures of firms obtained from COMPUSTAT. The accounting measures for some of the firms were not available for the complete period of 1995-2010. Hence, a subset of

outsourcing contracts was created to match the corresponding accounting measures. The data for implicit price deflators and the economy wide gross domestic product (GDP) was obtained from Bureau of Economic Analysis (BEA).

We were interested in those years for analysis, when a firm's outsourcing level changed. This condition further reduced the sample size to 60 observations. From these 60 observations, the data points for the last contract year were used for holdout sample, to determine the predictive ability of empirical model. The remaining 50 observations were used for training set to develop the regression model.

Measures. We describe below the methodology to calculate the outsourcing state of a firm in a given year.

Based on the information on contract value and contract length, the dollar value of a specific contract was divided by its contract length (in years), and then summed over all the active contracts a firm had in a given year. Hence, the total contract value for i^{th} firm in year t was given by:

$$contract\ value\ (S_{it}) = \sum_{j \in all\ active\ contracts} \left(\frac{Value\ of\ outsourcing\ contract}{Length\ of\ contract\ in\ years} \right)_j$$

The summary of constructs used in the analysis is provided in Table 4.9. Note, all quantities were converted to 2005 constant million dollars.

Table 4.9. Variable Description and Data Sources

Variable	Source	Units	Deflator
Revenue	COMPUSTAT	Million dollars	2005 implicit price deflators from BEA
Profit	COMPUSTAT	Million dollars	2005 implicit price deflators from BEA
Contract value	IDC and News wire	Million dollars	2005 implicit price deflators from BEA
Contract duration	IDC	Months converted in years	
Selling general & administrative expenses	COMPUSTAT	Million dollars	2005 implicit price deflators from BEA
GDP	BEA	Million dollars	2005 implicit price deflators from BEA

Analysis and results. The descriptive statistics of variables used in modeling is given in Table 4.10.

For (1), we analyzed the scatter plot of the dependent variable against both independent variables. Few observations were identified as outliers. First OLS regression was performed to investigate the different modeling assumptions. Very low R-square value was obtained using OLS regression. Further, the assumption of normality of residuals was strongly rejected ($p < .01$) using Shapiro-Wilk's test. When the normality assumptions are violated, the regression equation may generate unreliable p-value or t-statistics (Schwab et al. 2011). Based on the Durbin-Watson and Durbin-H tests, we found no first order autocorrelation among residuals. The White's test showed no evidence of heteroskedasticity of residuals.

Table 4.10. Descriptive Statistics of Variables

Variable	Mean (s. d.)	Profit/revenue	Peer factor	Action	Current State	SG&A	GDP (in billions)
Profit/revenue	0.004 (0.160)						
Peer factor	0.004 (0.09)	-0.007 (0.95)					
Action	58.283 (562.664)	0.007 (0.9613)	-0.09 (0.55)				
Current State	903.31 (1439)	0.100 (0.489)	0.72 (<.01)	0.185 (0.196)			
SG&A	8272 (8410)	0.053 (0.713)	0.5 (<.01)	0.129 (0.370)	0.872 (<.01)		
GDP (in billions)	12130 (939.463)	-0.038 (0.789)	-0.352 (0.01)	-0.439 (<.01)	-0.3 (0.03)	-0.197 (0.17)	

Pearson correlation coefficients are reported, with the p-values are given in the brackets.

When normality assumption of residuals is violated, robust regression procedures are used to mitigate the effects of unusual observations, as the estimates from robust regression are more reliable (Stuart 2011). These robust regression methods behave like traditional methods when data satisfy the assumptions of those models, but behave differently when data violate the assumptions such as: normality, and homoscedasticity. One of such widely used robust regression techniques is robust-MM regression (Schwab et al. 2011; Stuart 2011). MM-estimators combine the high asymptotic relative efficiency of M-estimators with the high breakdown of a class of S-estimators. Hence MM estimates have properties of both robust regression M-estimates and S-estimates. 'MM' refers to multiple M-estimates carried out in the computation of the estimator. We used robust MM regression to test hypotheses, and the estimates can be seen in Table 4.11 and Table 4.12.

Table 4.11. Parameter Estimates for Action (Equation 1)

Parameter	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept (β_0)	14.06	5.47	6.6	0.01
Profit margin (β_1)	**381.54*****	99.73	14.64	<.01
PeerFactor (β_2)	**141.44****	61.96	5.21	0.02
R-square = .023				

For (1), we found that higher level of profit margin (profit/revenue) was related to higher level of outsourcing $\left(\beta_1 = 381.54, p < .01\right)$. Further, higher the effect of peer pressure for outsourcing, the more likely firms were to outsource $\left(\beta_2 = 141.44, \ p = .02\right)$.

Table 4.12. Parameter Estimates for Equation 2

Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept (γ_0)	5.41	2.6	4.25	0.04
ln(SG&A) (γ_1)	**1.02*****	0.02	4542.88	<.01
ln(updated state) (γ_2)	**-0.02***	0.01	3	0.08
ln(GDP) (γ_3)	**-0.34****	0.16	4.43	0.04
R-square = .73				

In (4), we controlled for SG&A expenses from the previous year to account for firm-specific factors that affect SG&A in the current period (Harter et al. 2003; Reger et al. 1994). We also controlled for gross domestic product of US economy, to account for economy-wide factors.

For (4) used in predicting the selling general and administrative expenses (SG&A), we found that investments in outsourcing led to decrease in SG&A expenses $\left(\gamma_2 = -0.02, p < .1\right)$. Also the lagged value of SG&A expenses strongly

influenced the current SG&A level $\left(\gamma_0 = 1.02, p < .01 \right)$. GDP negatively influenced the SG&A expenses ($\gamma_3 = -0.34, p < .05$) in the automotive industry.

Accuracy prediction on the holdout dataset. Holdout dataset consisted of all contracts signed by firms in the sample, in the year corresponding to the last contract announced by a firm. In total, there were 10 contracts. The mean accuracy of estimation was calculated using mean magnitude of relative error (Agrawal et al. 2007), and was defined in the following way:

$$MMRE\% = \frac{\sum_{i=1}^{N} \left| \frac{Actual\ SGA - Estimated\ SGA}{Actual\ SGA} \right| \times 100}{N}$$

To assess the predictive ability of model introduced in the present research, first MMRE metric for the two-step regression was obtained. We then calculated MMRE metric, when future SGA for each firm was estimated using the following heuristic[10]: "mean of all observations corresponding to SG&A expenses for a particular firm in the training-set". The MMRE for our two step regression and the mean SG&A estimates were: 12.47 and 53.25 respectively, clearly demonstrating the improvement in predicting SGNA over mean SG&A heuristic, the proposed method in the present research provides.

Contributions and implications. This study makes important contribution to the IS literature in following ways. First, to the best of our knowledge, it is one of the first studies that empirically investigates the outsourcing behavior of firms through the lens of social influence and peer pressure. It establishes that outsourcing among automotive firms lead to reductions in SG&A expenses. It also suggests that outsourcing plays a major role in fulfilling CEOs' objectives to cut IT spending (Murphy 2012).

---

[10] Loosely it can be understood as moving average estimate for each firm.

Second, this research complements the previous literature on the impact of IT outsourcing (Han et al. 2011; Han et al. 2013), albeit, through a different theoretical perspective. We believe that the use of Information-Based Imitation to explain herding behavior of firms in the IT outsourcing context is a novel theoretical contribution IS literature, and it can be adapted to explain various business imitation processes.

In an increasingly competitive environment, where automotive firms such as GM and Ford are investing heavily on consolidating data centers and applications, centralizing IT planning and execution, maintaining privacy of customer data, and bringing various IT services in-house to cut cost and improve firm profit, our research has important implications as we show, that outsourcing has indeed positively contributed to bringing down the SG&A expenses of firms.

In the case of automotive firms, where up to 90% of different IT services are provided by outsourcing vendors (Murphy 2012), making restructuring decision of IT services calls for careful analysis by managers. By not availing the market oriented services, firms often lose the expertise of IT service providers, and face the risk of eroding their competitive advantage to manage operating costs.

The herding behavior approach taken in this research can be extended to study a variety of network and social influence phenomena. For instance, Oh et al. (2007) have used the herding behavior perspective to understand the membership dynamics in the open source software community. In an environment, where world is becoming closely interlinked due to the wider penetration of social media and web-services, modeling the herding behavior of agents to study IT driven phenomena will have important implications for policies and business issues.

When firms imitate each other for any phenomenon such as outsourcing, in an uncertain environment, they reduce the risk of falling behind rivals. Thus, imitation can spur productive innovation, or amplify the error of early movers (Lieberman et al. 2006). Herding behavior can also lead to bubbles, and waste of resources in mimicking others' strategies (Lieberman et al. 2006). Recent financial crisis and internet boom of the late 90's are some examples of this. Hence, deeper analysis of outsourcing phenomena via the herding behavior perspective, on other measures of firm, could be possible extensions of current research.

Investigating cases where agents with sufficiently extreme preferences do not conform to industry norms, yet are still able to successfully manage their operations, can be another possible extension of the current research. Finally, weighing the actions of peers based on some similarity metrics (Segev et al. 1999), to model the herding behavior among agents can provide us with fresh insights on the conforming behavior of firms.

Finally, we note that the phenomena of herd behavior and peer influence are also modeled through agent-based simulation (Lewis, Gonzalez, & Kaufman, 2012; Oh & Jeon, 2007; Zhao et al., 2011). The extension of the current research, through the simulation based approach of imitation model, in the context of outsourcing, can provide novel insights into outsourcing behavior of firms.

References

Agrawal, M., and Chari, K. 2007. "Software effort, quality, and cycle time: A study of CMM level 5 projects," *Software Engineering, IEEE Transactions on* (33:3), pp 145-156.

Aral, S., & Weill, P. (2007). IT Assets, Organizational Capabilities, and Firm Performance: How Resource Allocations and Organizational Differences Explain Performance Variation. *Organization Science, 18*(5), 763-780.

Armstrong, C. P., & Sambamurthy, V. (1999). Information technology assimilation in firms: The influence of senior leadership and IT infrastructures. *Information systems research, 10*(4), 304-327.

Audretsch, D. B. 1995. "Firm profitability, growth, and innovation," *Review of Industrial Organization* (10:5), pp 579-588.

Banker, R. D., Hu, N., Pavlou, P. A., & Luftman, J. (2011). CIO reporting structure, strategic positioning, and firm performance. *MIS Quarterly, 35*(2), 487-504

Bakos, J. Y., and Brynjolfsson, E. 1993. "Information Technology, Incentives, and the Optimal Number of Suppliers," *Journal of Management Information Systems* (10:2), pp 37-53.

Banerjee, A. V. 1992. "A simple model of herd behavior," *The Quarterly Journal of Economics* (107:3), pp 797-817.

Bassellier, G., Benbasat, I., & Reich, B. H. (2003). The influence of business managers' IT competence on championing IT. *Information systems research, 14*(4), 317-336.

Baum, J. A., and Haveman, H. A. 1997. "Love thy neighbor? Differentiation and agglomeration in the Manhattan hotel industry, 1898-1990," *Administrative Science Quarterly*), pp 304-338.

Benabou, R. 1996. "Equity and efficiency in human capital investment: the local connection," *The Review of Economic Studies* (63:2), pp 237-264.

Bernheim, B. D. 1994. "A theory of conformity," *Journal of political Economy*), pp 841-877.

Bikhchandani, S., Hirshleifer, D., and Welch, I. 1992. "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*), pp 992-1026.

Branch, B. 1974. "Research and development activity and profitability: a distributed lag analysis," *The Journal of Political Economy* (82:5), pp 999-1011.

Brock, W. A., and Durlauf, S. N. 2001. "Discrete choice with social interactions," *The Review of Economic Studies* (68:2), pp 235-260.

Bryce, D. J., and Useem, M. 1998. "The impact of corporate outsourcing on company value," *European Management Journal* (16:6), pp 635-643.

Cha, H. S., Pingry, D. E., and Thatcher, M. E. 2009. "A learning model of Information Technology outsourcing: Normative implications," *Journal of Management Information Systems* (26:2), pp 147-176.

Chang, R. M., Oh, W., Pinsonneault, A., and Kwon, D. 2010. "A network perspective of digital competition in online advertising industries: A simulation-based approach," *Information Systems Research* (21:3), pp 571-593.

Chastain, C. E. 1984. "Streamlining: Necessary strategy for licking a profit crunch," *Business Horizons* (27:2), pp 69-76.

Chatterjee, D., Richardson, V. J., & Zmud, R. W. (2001). Examining the shareholder wealth effects of announcements of newly created CIO positions. *MIS Quarterly*, 43-70.

Chaudhury, A., Nam, K., and Rao, H. R. 1995. "Management of Information Systems Outsourcing: A Bidding Perspective," *Journal of Management Information Systems* (12:2), pp 131-159.

Chen, Y., and Bharadwaj, A. 2009. "An empirical analysis of contract structures in IT outsourcing," *Information Systems Research* (20:4), pp 484-506.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression* (Vol. 5): Chapman and Hall New York.

Deephouse, D. L. 1999. "To be different, or to be the same? It's a question (and theory) of strategic balance," *Strategic Management Journal* (20:2), pp 147-166.

Dey, D., Fan, M., and Zhang, C. 2010. "Design and analysis of contracts for software outsourcing," *Information Systems Research* (21:1), pp 93-114.

Dibbern, J., Goles, T., Hirschheim, R., and Jayatilaka, B. 2004. "Information systems outsourcing: a survey and analysis of the literature," *ACM SIGMIS Database* (35:4), pp 6-102.

Dunn, D. 2005. "General Motors Signs Biggest Java Deal Ever," in *InformationWeek*.

Dutta, A., and Roy, R. 2005. "Offshore outsourcing: A dynamic causal model of counteracting forces," *Journal of Management Information Systems* (22:2), pp 15-35.

Earl, M. J., & Feeny, D. F. (1994). Is your CIO adding value? *Sloan Management Review, 35*, 11-11.

Feeny, D. F., Edwards, B. R., & Simpson, K. M. (1992). Understanding the CEO/CIO relationship. *MIS Quarterly*, 435-448.

Fersht, P., and Stiffler, D. 2009. "State of the Outsourcing Industry in Mid-2009: Activity To Resume With a More Cautious and Global Focus," *AMR Research*).

Fitoussi, D., and Gurbaxani, V. 2012. "IT outsourcing contracts and performance measurement," *Information Systems Research* (23:1), pp 129-143.

Gefen, D., Ragowsky, A., Licker, P., & Stern, M. (2011). The Changing Role of the CIO in the World of Outsourcing: Lessons Learned from a CIO Roundtable. *Communications of the Association for Information Systems, 28*(1), 15.

Gilley, K. M., and Rasheed, A. 2000. "Making more by doing less: an analysis of outsourcing and its effects on firm performance," *Journal of management* (26:4), pp 763-790.

Gore, A. K., Matsunaga, S., & Eric Yeung, P. (2011). The role of technical expertise in firm governance structure: Evidence from chief financial officer contractual incentives. *Strategic Management Journal, 32*(7), 771-786.

Han, K., Kauffman, R. J., and Nault, B. R. 2011. "Research Note—Returns to Information Technology Outsourcing," *Information Systems Research* (22:4), pp 824-840.

Han, K., and Mithas, S. 2013. "Information technology outsourcing and non-IT operating costs: An empirical investigation," *Management Information Systems Quarterly* (37:1), pp 315-331.

Harter, D. E., and Slaughter, S. A. 2003. "Quality improvement and infrastructure activity costs in software development: A longitudinal analysis," *Management science* (49:6), pp 784-800.

Hitt, L. M., Wu, D., and Zhou, X. 2002. "Investment in enterprise resource planning: Business impact and productivity measures," *Journal of Management Information Systems* (19:1), pp 71-98.

Inc., G. December 2011. "Forecast Analysis: IT Outsourcing, Worldwide, 2010-2015, 4Q11 Update,").

Jiang, B., Frazier, G. V., and Prater, E. L. 2006. "Outsourcing effects on firms' operational performance: an empirical study," *International Journal of Operations & Production Management* (26:12), pp 1280-1300.

Kambil, A., & Lucas, H. C. (2002). The board of directors and the management of information technology. *Communications of the Association for Information Systems, 8*(1), 26.

Keen, P. G. W. (1991). *Shaping the future: business design through information technology*: Harvard Business School Pr.

Koh, C., Ang, S., and Straub, D. W. 2004. "IT outsourcing success: a psychological contract perspective," *Information Systems Research* (15:4), pp 356-373.

Lakonishok, J., Shleifer, A., and Vishny, R. W. 1992. "The impact of institutional trading on stock prices," *Journal of financial economics* (32:1), pp 23-43.

Lee, J.-N., Miranda, S. M., and Kim, Y.-M. 2004. "IT outsourcing strategies: Universalistic, contingency, and configurational explanations of success," *Information Systems Research* (15:2), pp 110-131.

Lee, J., Lee, K., and Rho, S. 2002. "An evolutionary perspective on strategic group emergence: a genetic algorithm-based model," *Strategic Management Journal* (23:8), pp 727-746.

Lewis, K., Gonzalez, M., & Kaufman, J. 2012. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1): 68-72.

Lieberman, M. B., and Asaba, S. 2006. "Why do firms imitate each other?," *Academy of Management Review* (31:2), pp 366-385.

Loh, L., and Venkatraman, N. 1992a. "Determinants of information technology outsourcing: a cross-sectional analysis," *Journal of Management Information Systems*), pp 7-24.

Loh, L., and Venkatraman, N. 1992b. "Diffusion of information technology outsourcing: influence sources and the Kodak effect," *Information Systems Research* (3:4), pp 334-358.

Lopez-Pintado, D., and Watts, D. J. 2008. "Social influence, binary decisions and collective dynamics," *Rationality and Society* (20:4), pp 399-443.

Luftman, J., & Kempaiah, R. (2008). Key issues for IT executives 2007. *MIS Quarterly Executive, 7*(2), 99-112.

Mani, D., Barua, A., and Whinston, A. 2010. "An empirical analysis of the impact of information capabilities design on business process outsourcing performance," *MIS quarterly* (34:1), pp 39-62.

Mateyaschuk, J. (1999). CIOs Head for the Top. *Informationweek*, 128.

Overby, S. (2003). The incredible shrinking CIO. *CIO Magazine, 17*(2), 15.

McDonald, M. P. 2010. "IT spend as a percent of revenue – a dubious metric at best," in *Gartner Inc.*

Mithas, S., and Jones, J. L. 2007. "Do Auction Parameters Affect Buyer Surplus in E-Auctions for Procurement?," *Production and Operations Management* (16:4), pp 455-470.

Mukhopadhyay, T., Kekre, S., and Kalathur, S. 1995. "Business Value of Information Technology: A Study of Electronic Data Interchange," *MIS quarterly* (19:2), pp 137-156.

Murphy, C. 2012. "General Motors will Slash Outsourcing in IT Ovehaul."

Novak, S., and Stern, S. 2008. "How does outsourcing affect performance dynamics? Evidence from the automobile industry," *Management science* (54:12), pp 1963-1979.

Oh, W., and Jeon, S. 2007. "Membership herding and network stability in the open source community: The Ising perspective," *Management science* (53:7), pp 1086-1101.

Preston, R. 2012. "Randy Mott's Journey To General Motors."

Ramasubbu, N., Mithas, S., Krishnan, M., and Kemerer, C. F. 2008. "Work dispersion, process-based learning, and offshore software development performance," *MIS quarterly* (32:2), pp 437-458.

Reger, R. K., Gustafson, L. T., Demarie, S. M., and Mullane, J. V. 1994. "Reframing the organization: Why implementing total quality is easier said than done," *Academy of Management Review* (19:3), pp 565-584.

Rockart, J. F., Earl, M. J., & Ross, J. W. (1996). Eight imperatives for the new IT organization. *Sloan Management Review, 38*(1), 43-55.

Schelling, T. C. 1971. "Dynamic models of segregation†," *Journal of mathematical sociology* (1:2), pp 143-186.

Schwab, A., Abrahamson, E., Starbuck, W. H., and Fidler, F. 2011. "PERSPECTIVE—Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests," *Organization Science* (22:4), pp 1105-1120.

Segev, E., Raveh, A., and Farjoun, M. 1999. "Conceptual maps of the leading MBA programs in the United States: core courses, concentration areas, and the ranking of the school," *Strategic Management Journal* (20:6), pp 549-565.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3/4), 591-611.

Slaughter, S., and Ang, S. 1996. "Employment outsourcing in information systems," *Communications of the ACM* (39:7), pp 47-54.

Smith, M. A., Mitra, S., and Narasimhan, S. 1998. "Information Systems Outsourcing: A Study of Pre-Event Firm Characteristics," *Journal of Management Information Systems* (15:2), pp 61-93.

Soon, A., and Straub, D. W. 1998. "Production and Transaction Economies and IS Outsourcing: A Study of the U. S. Banking Industry," *MIS quarterly* (22:4), pp 535-552.

Starbuck, W. H. (2006). *The production of knowledge: The challenge of social science research*: Oxford University Press, USA.

Stuart, C. 2011. "Robust Regression,").

Susarla, A., Subramanyam, R., and Karhade, P. 2010. "Contractual provisions to mitigate holdup: Evidence from information technology outsourcing," *Information Systems Research* (21:1), pp 37-55.

Tanriverdi, H., Konana, P., and Ge, L. 2007. "The choice of sourcing mechanisms for business processes," *Information Systems Research* (18:3), pp 280-299.

Techweb 2009. "Renault Outsources To Capgemini."

Westerman, G., & Hunter, R. (2007). *IT risk: turning business threats into competitive advantage*: Harvard Business School Press Boston.

Whitaker, J., Mithas, S., and Krishnan, M. 2010. "Organizational learning and capabilities for onshore and offshore business process outsourcing," *Journal of Management Information Systems* (27:3), pp 11-42.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817-838.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 642-656.

Zhao, L., Yang, G., Wang, W., Chen, Y., Huang, J., Ohashi, H., & Stanley, H. E. 2011. Herd behavior in a complex adaptive system. *Proceedings of the National Academy of Sciences*, 108(37): 15058-15063.

Zhu, K., Kraemer, K. L., & Dedrick, J. (2004). Information technology payoff in e-business environments: An international perspective on value creation of e-business in the financial services industry. *Journal of management information systems, 21*(1), 17-54.

Zimmerman, M. B. 1982. "Learning Effects and the Commercialization of New Energy Technologies: The Case of Nuclear Power," *The Bell Journal of Economics* (13:2), pp 297-310.

Chapter 5 : Appendices

Appendix 1: Proofs

Table 5.1A. Summary of terms used in analytical modeling

Term	Definition
n_0	Initial natural count of the article-a
m_0	Initial natural count of the article-b
n	Total number of iterations
p	Probability that a reader reads recommended article upon arrival
$1-p$	Probability that a reader reads un-recommended article upon arrival
A_n^h	The count of article-a in hard cutoff NRS, after n iterations
A_{nl}^p	The count of article-a in probabilistic NRS when $p = 0$ at every time step, after n iterations
A_{nu}^p	The count of article-a in probabilistic NRS when $p = 1$ at every time step, after n iterations
$p_t(read)$	Probability of article-a being read in probabilistic NRS at time t
p_{at}, p_{bt}	Probabilities of article-a and article-b being recommended in probabilistic NRS
$p_{tl}(read)= 1 - p_{at}$	Probability of article-a being read in probabilistic NRS when $p = 0$ at every time step
$p_{tu}(read) = p_{at}$	Probability of article-a being read in probabilistic NRS when $p = 1$ at every time step
τ_n	Total count of articles 'a' and 'b' after n iterations
Z_n	Random variable defined as $Z_n = A_{nl}^p - \frac{\tau_n}{2}$
r_i	Probability of recommended article being read in 2nd user model

Appendix 1 (Continued)

Proposition 1. Let A_n^h represents the count of article-a after n^{th} iteration in hard cutoff NRS. Then

$$E(A_n^h) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (n_0 + k)$$

$$= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (n_0) + \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (k)$$

Hence, $E(A_n^h) = n_0 + n * p$

Proposition 2. First the expression for $E(A_{nl}^p)$ has been derived.

As $p_{tl}(read) = 1 - p_{at}$,

At any time n we have the recurrence relation

$$P(A_{n+1l}^p = A_{nl}^p + 0 | A_{nl}^p) = \frac{A_{nl}^p}{\tau_n} \text{ and } P(A_{n+1l}^p = A_{nl}^p + 1 | A_{nl}^p) = 1 - \frac{A_{nl}^p}{\tau_n}. \text{ Hence,}$$

$$E(A_{n+1l}^p | A_{nl}^p) = (A_{nl}^p + 0) \frac{A_{nl}^p}{\tau_n} + (A_{nl}^p + 1) \left(1 - \frac{A_{nl}^p}{\tau_n}\right) = 1 + A_{nl}^p \left(1 - \frac{1}{\tau_n}\right)$$

Taking expectation on both sides and using the property of conditional expectation

$$E(A_{n+1l}^p) = 1 + (1 - \frac{1}{\tau_n}) E(A_{nl}^p)$$

Using the transformation, $A_{nl}^p = Z_n + \frac{\tau_n}{2}$ results in following relation

$$E(Z_{n+1}) = \left(1 - \frac{1}{\tau_n}\right) E(Z_n) \text{ or } E(Z_n) = \left(1 - \frac{1}{\tau_{n-1}}\right) E(Z_{n-1}) \tag{5}$$

Iteratively using the recurrence relation (5) results in

$$E(Z_n) = \prod_{j=0}^{n-1} \left(\frac{\tau_j - 1}{\tau_j}\right) (A_{0l}^p - \frac{\tau_0}{2}) \tag{6}$$

Using the relation $\tau_{j-1} = \tau_j - 1$ for $n \geq 1$ the expression in (6) results in

Appendix 1 (Continued)

$$E(Z_n) = \left(\frac{\tau_0 - 1}{\tau_0}\right) \prod_{j=1}^{n-1} \left(\frac{\tau_{j-1}}{\tau_j}\right) \left(A_{0l}^p - \frac{\tau_0}{2}\right) = \left(\frac{\tau_0 - 1}{\tau_{n-1}}\right) \left(A_{0l}^p - \frac{\tau_0}{2}\right)$$

Substituting values $A_{0l}^p = n_0$, $\tau_0 = n_0 + m_0$, and $\tau_{n-1} = n_0 + m_0 + n - 1$

$$E(Z_n) = \left(\frac{n_0 + m_0 - 1}{n_0 + m_0 + n - 1}\right) \left(n_0 - \frac{n_0 + m_0}{2}\right) = \left(\frac{n_0 + m_0 - 1}{n_0 + m_0 + n - 1}\right) \left(\frac{n_0 - m_0}{2}\right)$$

Hence, $E\left(A_{nl}^p - \frac{\tau_n}{2}\right) = \left(\frac{n_0 + m_0 - 1}{n_0 + m_0 + n - 1}\right) \left(\frac{n_0 - m_0}{2}\right)$

Finally, we have $E(A_{nl}^p) = \left(\frac{n_0 + m_0 - 1}{n_0 + m_0 + n - 1}\right) \left(\frac{n_0 - m_0}{2}\right) + \frac{n_0 + m_0 + n}{2}$ (7)

Now, we derive the expression for $E(A_{nu}^p)$.

Probability of article-a being read at time t and hence probability of increase in the count of the article-a at any given time t

is $p_{tu}(read) = p_{at}$ (8)

Suppose $1 \le i_1 \le i_2 \le \cdots \ldots \ldots \le i_k \le n$ be the time indices when an article-a was read by a reader. Then the probability of this particular string will be given by

$$\left(\frac{m_0}{n_0 + m_0}\right) * \left(\frac{m_0 + 1}{n_0 + m_0 + 1}\right) * \ldots \ldots \left(\frac{n_0}{n_0 + m_0 + i_1 - 1}\right) * \left(\frac{m_0 + i_1 - 1}{n_0 + m_0 + i_1}\right)$$

$$* \ldots \left(\frac{n_0 + 1}{n_0 + m_0 + i_2 - 1}\right) * \left(\frac{m_0 + i_2 - 2}{n_0 + m_0 + i_2}\right) * \ldots * \left(\frac{n_0 + k - 1}{n_0 + m_0 + i_k - 1}\right)$$

$$* \ldots * \left(\frac{m_0 + n - k - 1}{n_0 + m_0 + n - 1}\right)$$

These indices can be chosen in $\binom{n}{k}$ ways so, probability of having an article-$'a'$ being read k times and hence $A_{nu}^p = n_0 + k$ is given by

$$p(A_{nu}^p = n_0 + k)$$

$$= \binom{n}{k} \frac{n_0(n_0 + 1) \dots (n_0 + k - 1) * m_0(m_0 + 1) \dots (m_0 + n - k - 1)}{(n_0 + m_0)(n_0 + m_0 + 1) \dots (n_0 + m_0 + n - 1)} \quad (9)$$

Hence, $E(A_{nu}^p) = \sum_{k=0}^n p(A_{nu}^p = n_0 + k) * (n_0 + k)$ \quad (10)

$$= n_0 \sum_{k=0}^n p(A_{nu}^p = n_0 + k) + \sum_{k=0}^n k * p(A_{nu}^p = n_0 + k)$$

The expressions $\sum_{k=0}^n p(A_{nu}^p = n_0 + k)$ and $\sum_{k=0}^n k * p(A_{nu}^p = n_0 + k)$ has been

calculated separately. From the result in equation (9) we have,

$$\sum_{k=0}^n p(A_{nu}^p = n_0 + k) = \sum_{k=0}^n \binom{n}{k} \frac{n_0(n_0 + 1) \dots (n_0 + k - 1) * m_0(m_0 + 1) \dots (m_0 + n - k - 1)}{(n_0 + m_0)(n_0 + m_0 + 1) \dots (n_0 + m_0 + n - 1)}$$

$$= Total\ selection\ probability\ of\ either\ of\ the\ articles\ in\ 'n' iterations = 1$$

Now,

$$\sum_{k=0}^n k * p(A_{nu}^p = n_0 + k) = \sum_{k=1}^n k * \binom{n}{k} \frac{n_0(n_0 + 1) \dots (n_0 + k - 1) * m_0 \dots (m_0 + n - k - 1)}{(n_0 + m_0)(n_0 + m_0 + 1) \dots (n_0 + m_0 + n - 1)}$$

Using the property $k * \binom{n}{k} = n\binom{n-1}{k-1}$, the expression earlier takes the following form

$$n \sum_{k=1}^n \binom{n-1}{k-1} \frac{n_0(n_0 + 1) \dots (n_0 + k - 1) * m_0(m_0 + 1) \dots (m_0 + n - k - 1)}{(n_0 + m_0)(n_0 + m_0 + 1) \dots (n_0 + m_0 + n - 1)}$$

$$= n \frac{n_0}{n_0 + m_0} \sum_{k=1}^n \binom{n-1}{k-1} \frac{(n_0 + 1) \dots (n_0 + k - 1) * m_0(m_0 + 1) \dots (m_0 + n - k - 1)}{(n_0 + m_0 + 1) \dots (n_0 + m_0 + n - 1)}$$

$$= n \frac{n_0}{n_0 + m_0} \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{(n_0 + 1) \dots (n_0 + k) * m_0(m_0 + 1) \dots (m_0 + n - k - 2)}{(n_0 + m_0 + 1) \dots (n_0 + m_0 + n - 1)} \quad (11)$$

$$= n \frac{n_0}{n_0 + m_0} * 1 = n \frac{n_0}{n_0 + m_0}$$

Appendix 1(continued)

The last expression in equation (11) is the total selection probability of either of the articles in' $n-1'$ iterations but with the initial counts of n_0+1 and m_0 for the article a and b respectively.

So expression in (10) becomes,

$$E(A_{nu}^p) = n_0 + n\frac{n_0}{n_0+m_0} = \frac{n_0}{n_0+m_0}(n_0+m_0+n) \qquad (12)$$

Proposition 4. $E(B_n^{,h}) = \sum_{k=0}^{n-\epsilon}\binom{n-\epsilon}{k}p^k(1-p)^{n-k}(m_0+\epsilon+k)$

$$= \sum_{k=0}^{n-\epsilon}\binom{n-\epsilon}{k}p^k(1-p)^{n-k}(m_0+\epsilon) + \sum_{k=0}^{n-\epsilon}\binom{n-\epsilon}{k}p^k(1-p)^{n-k}(k)$$

$$= m_0 + \epsilon + (n-\epsilon)*p$$

Appendix 2: Sensitivity Analysis

In case of news articles, where majority of queries are driven by front page display or recommended articles, we expect popularity to exhibit some kind of power law distribution. The rationale for power-law distribution of popularity, especially in web-based systems, has been suggested by Easley and Kleinberg (2010). This assumption of popularity is also consistent with the effect of social influence discussed by Salganik, et al. (2006). In their experiment for artificial music market, they found that in the presence of social influence, such as media sites, we observe greater inequality – popular entities are more popular and unpopular entities are less popular. From a given power-law distribution its corresponding Zipf distribution can also be obtained (Adamic 2000).

To validate the popularity distribution of articles, we obtained data on popularity of articles from DailyMe Inc., a company that provides news personalization technology to a large number of media sites. There are five datasets from five different local news websites serving markets in Connecticut, Pennsylvania, New York, Colorado and Massachusetts, collected during the period of February 2012 to April 2012. The data provided listed specific articles along with cookie IDs and time stamps read across the five different local news websites.

Figure 5.2A shows the normalized frequency distribution on a log-log scale using the *logarithmic binning* with multiplier of 2 – similar to the procedure described by Newman (2005). The X-axis corresponds to natural log value of bins and Y-axis corresponds to the natural log value of normalized frequencies. Data from these five real local news websites show the pattern of power-law in popularity. Based on the findings,

Appendix 2 (Continued)

the power-law exponent used to discuss results in the sensitivity analysis is given by $c = 1.7$. We used this exponent value in the modified simulation model, where the initial distribution of article counts was generated using power-law distribution with exponent 1.7. We note that these are relatively smaller local news Web sites and we do not therefore make broader generalizations about the power law based on these alone. However, it provides valuable insights for this sensitivity analysis.

The probability density function of power law is given by $f(k) = \frac{a}{k^c}$, for some exponent c and the constant of proportionality a. $f(k)$ represents the fraction of articles which have popularity k. Cumulative distribution of power law follows Zipf's law (Newman 2005). The Zipf's probability mass function of an article ranked k, when the total number of articles in the system is N, is given by:

$$p(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$$

In the above expression the value of s characterizes the behavior of the system. Further, between a given Zipf's distribution and its corresponding power law distribution the following relation holds between the exponents $c = 1 + \frac{1}{s}$ (Adamic 2000). We will use this relation between exponents in the simulation model.

Empirical analysis of popularity distribution of articles. Figure 5.1A depicts the histogram plot for the popularity of articles on each of the five sites. In all cases popularity distribution is L-shaped. The X axis is article counts, binned in intervals of width 100. The Y axis is the number of news articles in the period that have the corresponding count.
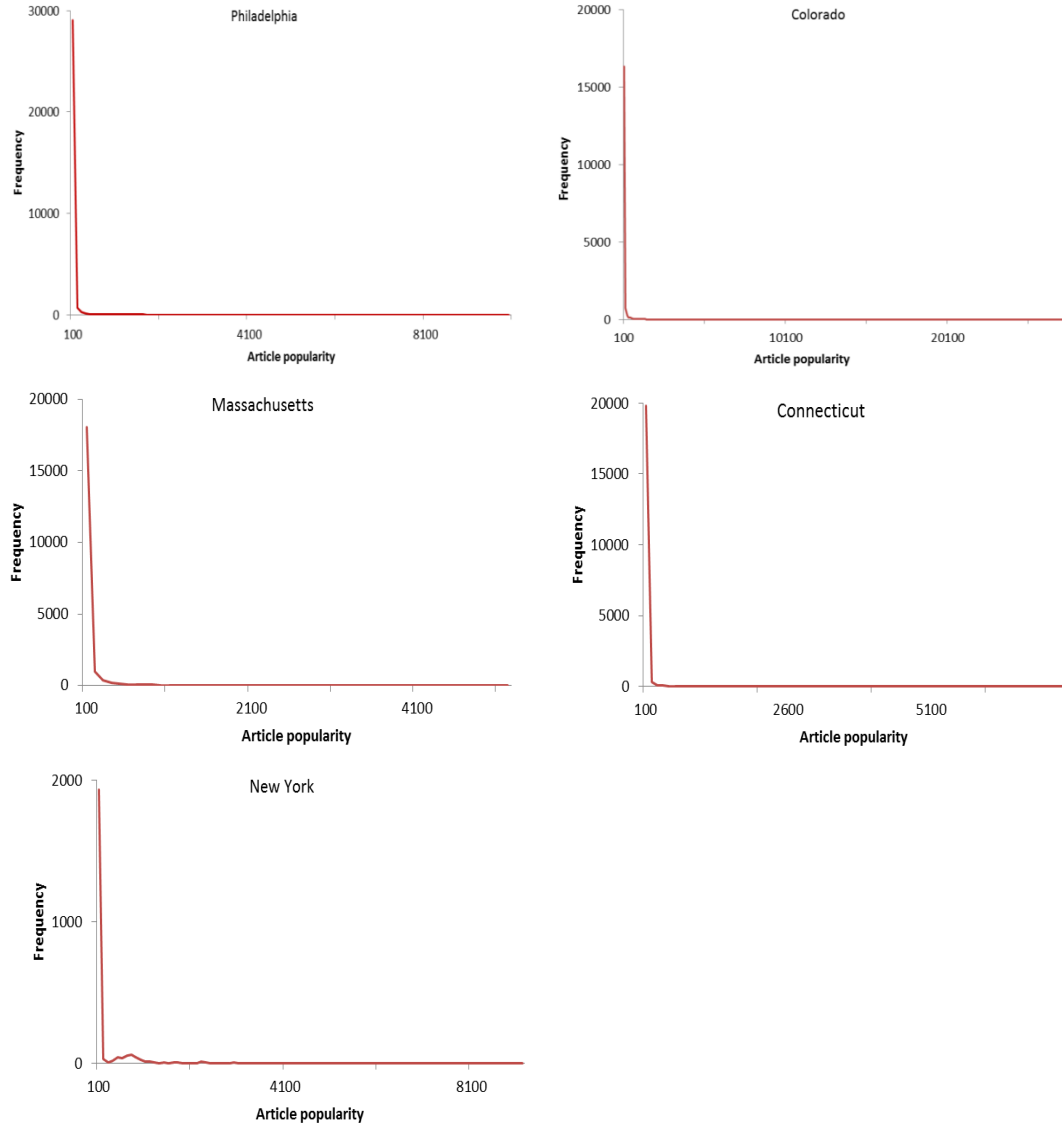
Appendix 2 (Continued)

Figure 5.1A. Distribution of the number of articles receiving a given number of counts. To plot the histogram, X-axis has been binned in the intervals of length 100. Y-axis corresponds to the number articles falling in that range.

Further we plotted the normalized frequency distribution on log-log scale using the *logarithmic binning* with multiplier of 2 – similar to the procedure described by Newman (2005). Findings in this case are produced in the Figure 5.2A, with the slope of curves. X-axis corresponds

155

to natural log value of bins and Y-axis corresponds to the natural log value of normalized

frequencies.
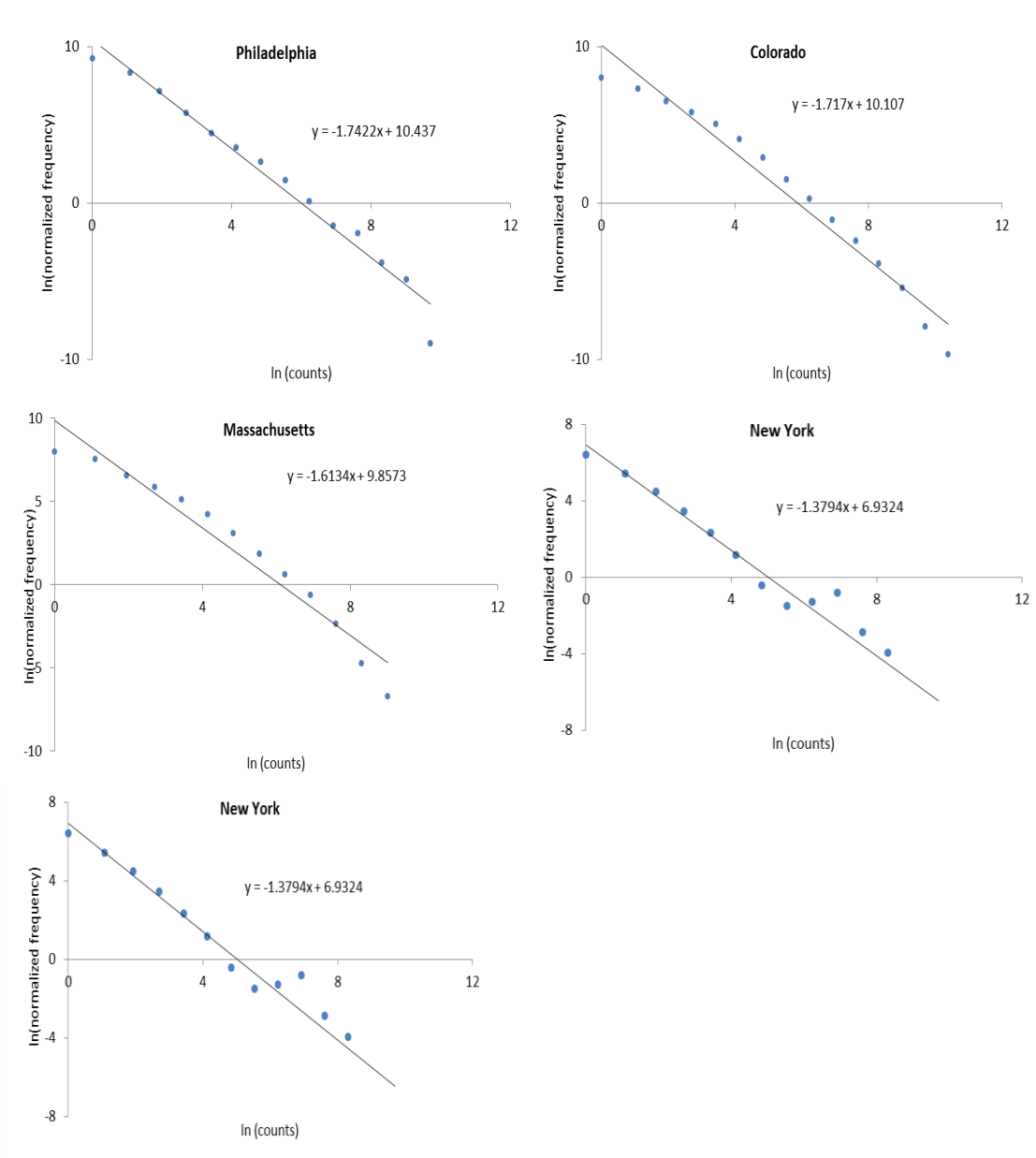

Figure 5.2A. log-log plot for popularity of articles at five different sites

Data from these five real local news websites show the pattern of power-law in popularity. Based on the findings, the exponent used to discuss results in the sensitivity analysis is given by $c = 1.7$ Hence for the simulation model that follows, the value of s is given by $s = 1.4$.

Simulation setup. In this case initially articles were assigned random counts between 0 and 1000 generated using Zipf distribution with the exponent of 1.4. Further, as before the difference in the counts of Nth and $(N + 1)^{th}$ article was kept to be 1. Other simulation parameters remain same as in the case on uniform distribution with $p = 0.9$. This particular choice of p has been chosen to illustrate the case with influential NRS.
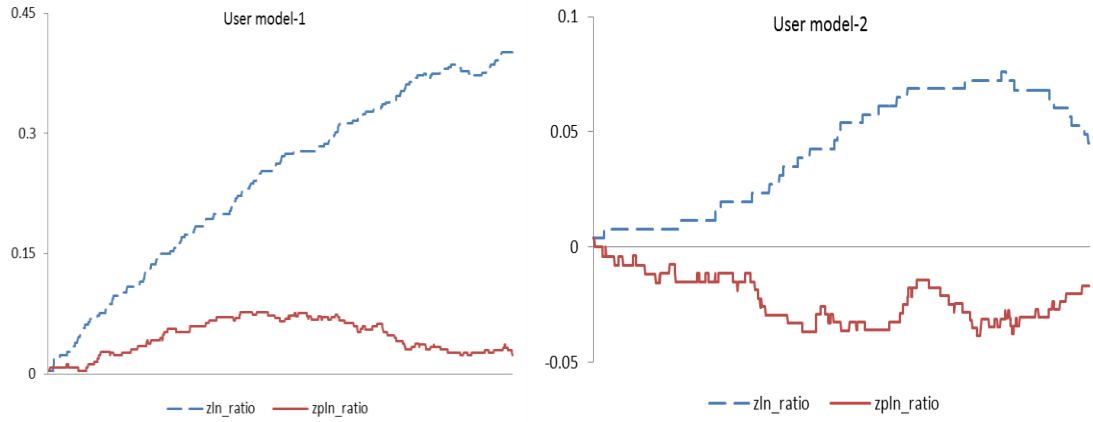

Figure 5.3A. Sample simulation path for boundary amplification

The count of the selected articled is increased by 1. For the recommended articles (i.e. DL), we follow the selection process based on two user models. In the first user model a recommended article is selected randomly. Whereas in the second user model a

recommended article is selected based on linear decrease in selection probability of the recommended articles. In both cases, the count of selected article is increased by 1.

The discussion in this section is based on findings that resulted after running the simulation model multiple times. One sample path in each case is produced in the figures 5.3A, 5.4A & 5.5A.

Figure-5.3A shows that the issue of count amplification between Nth and (N+1)th article still exists in the modified simulation setup. The issue of count amplification (based on M1) for the hardcutoff scenario was observed for both reader models. Also in the probablisitic selection mechansim, the path of M1 stays close to its initial value (i.e. ~0) for both reader model.

Comparing to the results in the case of the uniform distribution considered before, one difference is in terms of the highest value M1 takes at the end of simulation. But this is mainly due to difference in the initial distribution of articles. In case of power-law, the inital counts of 10th and 11th article were almost 4 times lower than in the case of uniform distribution. Besides this the findings remain consistent - for probabilistic mechanism, in presence of the second reader model, M1 has random fluctuations close to its initial value – similar to the observation in case of uniform distribution.

Manipulation. We consider two cases of manipulation that are of major interest: (i) early little (10, 100) and (ii) early heavy – (50, 100). Overall findings for manipulation remain similar to our prior findings, although the benefits of manipulation appear slightly lower for the manipulator. When popularity of articles follow power-law, coupled with increasingly focused attention for the top ranked articles, even in the recommended

articles (user model-2), we observe that to make manipulation activity successful, requires substantially more clicks to maintain higher popularity of the target article in some cases (right panel, Figure 5.4A & 5.5A). For example in Figure 5.5A (right panel) – even the case of heavy early manipulation, results do not appear to be as encouraging for a manipulator as in the case of uniform distribution. Earlier, for the case of heavy manipulation in presence of uniform distribution (Figure 3.8), once a manipulator stopped the manipulation activity, he was able to leverage self-reinforcing nature of hardcutoff. While that phenomenon still exists here (the downtrend in M1 continues), the trend is less pronounced than was the case under the uniform distribution. However the consistent theme remains – probabilistic selection continues to offer benefits in terms of offering significant resistance to manipulation.
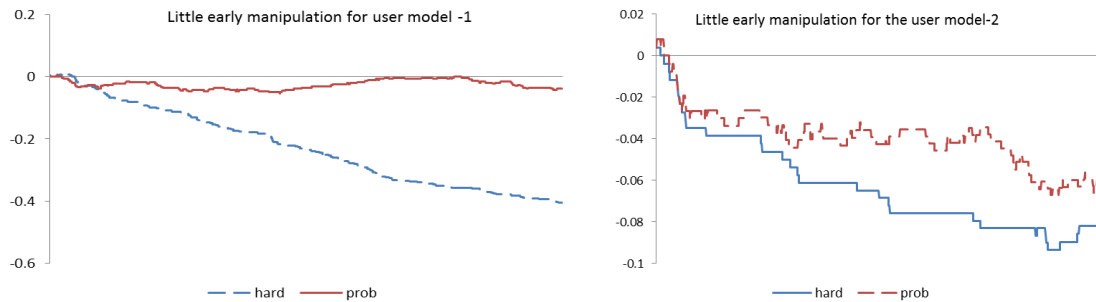

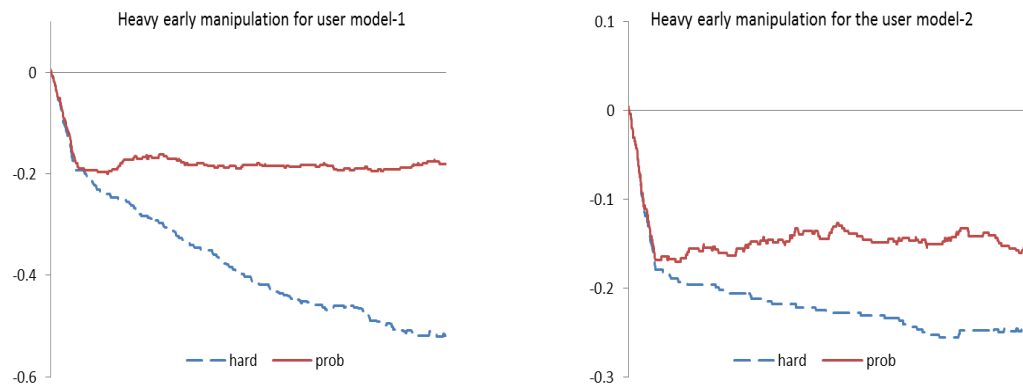Figure 5.4A. Little early manipulation for Zipf distribution

Figure 5.5A. Heavy early manipulation for Zipf distribution

Appendix 3: Frequency Table

Table 5.2A. Frequency Table

	Other industries	Retail and trade	Finance	Services
Non IT background-*profit*	8	22	13	10
IT background-*profit*	10	6	6	8
Non IT background-*op exp*.	11	27	21	12
IT background-op *exp*.	12	11	7	11
Non IT background-*SG&A*	10	7	4	9
IT background-*SG&A*	11	12	14	7

References

Adamic, L.A., "Zipf, power-laws, and pareto-a ranking tutorial," *Xerox Palo Alto Research Center, Palo Alto, CA,* http://ginger. *hpl. hp. com/shl/papers/ranking/ranking. html*, (2000),

Bilton, Nick, "Disruptions: Top 10 Lists Lead to Less Choice on the Web," *The New York Times*, April 1, 2012,

Easley, D. and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge Univ Pr, 2010.

Matthew J. Salganik, Peter S. Doddsand Duncan J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *Science*, 311, 5762, (2006), 854-856.

Newman, M.E.J., "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, 46, 5, (2005), 323-351.

Salganik, Matthew J., Peter Sheridan Dodds and Duncan J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, 311, 5762, (2006), 854-856.