2019 International Conference on Big Data in Business

October 29th to October 31st, 2019

London, United Kingdom

Table of Contents

Table of Contents

Incomplete Draft. All comments are welcomed. Please do not cite without permission.

Cross-border Acquisitions and Dyadic Distance[1]

EDWARD R. LAWRENCE, MEHUL RAITHATHA, and
IVÁN M. RODRÍGUEZ, JR.

ABSTRACT

We study how dyadic distance influences the initiation, completion, and duration of crossborder deals. Using a sample of 173,616 cross-border deals announced between 1970 and 2016, we find evidence that cross-country differences in culture and geographical distance influences the initiation of cross-border deals; differences in culture and institutions influences the completion of cross-border deals; and the duration of deals is influenced by idiosyncratic factors.

Keywords: Dyadic Distance, Acquisitions, M&A, Cross-country

JEL Code: F20, G30, G34

Cross-border Acquisitions and Dyadic Distance

[1] *Current Version:* November 23, 2018. Edward R. Lawrence and Ivan M. Rodríguez are at the finance Department at the Chapman School of Business, Florida International University, 11200 S.W. 8th St. RB 210 , Miami, Florida, 33199. *Email:* elawrenc@fiu.edu; imrodrig@fiu.edu. Mehul Raithatha is at the Indian Institute of Management, Indore. *Email:* mehulr@iimidr.ac.in.

ABSTRACT

Using a sample of 173,616 cross-border deals announced between 1970 and 2016, and over 8000 country pairs, we find evidence that cross-country cultural, institutional, and geographical distances play a deciding role in the initiation of mergers and acquisitions between countries. The completion of initiated acquisitions is independent of cultural and institutional factors. However, we find that the higher the geographical distances between countries the higher is the probability of completion of deal. In addition, we find the time taken from initiation to completion of deals is independent of cultural, institutional and geographical distances between acquirer and target countries.

In the last few decades cross border mergers and acquisitions have become considerable part of global economic growth. In the five-year period from 2010 to 2015 cross border deals worth $5.8 trillion were announced across the world. In 2015 alone, the value of announced cross border deals exceeded $1.38 trillion.[2] Because of its high economic significance cross border mergers and acquisitions have been an area of active research by academicians. Hymer (1960), in his dissertation which founded the field of international business, noted that there is a "liability of foreignness" when firms expand their operations to other countries. The introduction of the gravity equation to economics by Tinbergen (1962) – which analogizes Newton's law of universal gravitation to cross-border flows – has allowed researchers to explore the empirical structure of the costs and benefits of cross-border mergers and acquisitions, i.e. "liability of foreignness".

[2] See the Deloitte report: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/mergersacqisitions/us-m-a-cross-border-pov-spread.pdf

From these foundations, both the theoretical literature and empirical literature have focused on studying the determinants of foreign direct investment, international trade, and cross border mergers and acquisitions.[3] These determinants may either be firm level factors like firm size, public status, industry affiliation, and mode of payment; or country level factors like differences in institutional and cultural characteristics. In this paper, we contribute to the literature by investigating several new dimensions of cross border mergers and acquisitions that have not been explored so far. We study the impact of differences in cultural, institutional, and geographic factors between an acquirer-target country pair, on initiation, completion and duration of cross border mergers and acquisitions. Our study is the first, to our knowledge, to study these three facets of acquisitions – initiation, completion, and duration – within a large-scale unified framework.[4]

We define deal *Initiation* as the number of cross-border acquisitions from acquirer country to target country relative to the total number of cross-border acquisitions for acquirer country. Deal *Completion* is defined as proportion of completed cross-border acquisitions from acquirer country to target country relative to the number of cross-border acquisitions initiated from acquirer country to target country. finally, deal *Duration* is the average number of days from the announcement of deal to its completion for each acquirer-target country pair.

[3] Theoretical work has a long history and remains a vibrant area of study, e.g., Anderson (1979), Helpman and Krugman (1985), Bergstrand (1985), Davis (1995), Deardoff (1998), and Anderson and van Wincoop (2003). The empirical economics literature has evaluated trade protection (Harrigan, 1993), regional trade agreements (Frankel, Stein, and Wei, 1998), exchange rate variability (Frankel and Wei, 1993; Eichengreen and Irwin, 1998) , border effects (McCallum, 1995; Anderson and van Wincoop, 2003), and cross border mergers & acquitions (Rossi and Volpin, 2004; Erel, Liao, and Weisbach, 2012; Serdar Dinc and Erel, 2013; Aktas, de Bodt, and Roll, 2013 ; Kim and Lu, 2013, Ahern, Daminelli, and Fracassi, 2015).

[4] While we are the first to analyze the initiation, completion, and duration for cross-border mergers and acquisitions, Dikova et al (2010) uses similar definition for completion and duration. However, they study business service industry only. They do not study initiation, and their completion and duration variables are deal level variables instead of country level variables used in our study.

We use a sample of 173,616 cross-border mergers occurring between 1970 and 2016 for our analysis. We find that cultural, institutional, and geographical distances all play a deciding role in the initiation of mergers and acquisitions. The higher these differences the lower the probability of initiation of cross border merges and acquisitions. Next, we find that completion of deal is orthogonal to cultural and institutional factors but largely depends on the geographic distance between acquiring and target countries. The higher the geographical distance the higher is the likelihood of completion. Lastly, our analysis of duration reveals that there is no impact of cross-country cultural or institutional differences on the time between the initiation and completion of the deal.

Differences in culture (e.g., Weber, 1930; Veliz, 1994; Landes, 1998; Guiso, Sapienza, and Zingales, 2003, 2006), institutions (e.g., Alesina and Rodrik, 1994; La Porta, Lopez de

Silanes, Shleifer, and Vishny, 1998; Hirshleifer, 2001), and geography (e.g., Myrdal, 1968 ; Sachs, 2003) have been found to be important fundamental causes of cross-country differences in economic outcomes. Cultural differences between countries – such as differences in religion, language, and general shared experiences – play a key role in shaping economic decision making and outcomes. Different cultures generate diverse sets of beliefs about how to behave, which in turn influence equilibrium outcomes and results in countries coordinating on different equilibria. Differences in institutions shape economic outcomes through various channels; for example, a lack of property rights de-incentivize economic agents while functioning markets help allocate resources efficiently. Differences in geographic endowments naturally determine the preferences and opportunity set of economic agents across different countries. For example, geography may determine the development and availability of different technologies.

These three fundamental factors have also become the recent focus in explaining crossborder acquisition flows between countries. Ahern et al. (2015) isolates the impact of

culture on cross-border acquisitions. They find that countries that are more distant culturally have lower cross-border acquisition activity. Rossi and Volipin (2004) find that the volume of cross-country acquisitions is larger in countries with stronger accounting standards and stronger shareholder protection. In a similar vein, Erel, Liao, and Weisbach (2013) find that geographic distance and differences in various institutional factors – such as accounting disclosures and economic development – are positively related to cross-border acquisition flows between countries. Additionally, the literature has documented that in countries with far-right parties or in times of weak government there is more intervention in large-scale foreign acquisitions (Serdar Dinc and Erel, 2014).

While the prior literature has studied the importance of culture, institutions, and geography, there has been a large emphasis on the flows or propensity of cross-border acquisitions, defined either as dollar volume or as the number of acquisitions between the target-acquirer country pair relative to the sum of domestic mergers in the target country and the number of acquisitions between the target-acquirer country pair. They have also restricted their sample to completed deals where the acquirer acquires majority stakes in the target firm and the value of deal is more than $1 million. Additionally, they exclude all deals that were initiated but could not be completed. By excluding such deals, the sample is biased towards larger firms from more developed countries.[5]

In our study, we alleviate these sample selection issues by considering all the initiated deals irrespective of their completion status, their size, and the proportion of the stake that the acquirer receives. Moreover, instead of dollar volume, we use number of deals, thus alleviating

[5] Size matters because in developing and under developed economies, a company of $1 million value is a bigger firm.

bias towards the bigger companies and companies from the developed world. This allows us to study the holistic view of country-pair differences on cross border mergers.

Based on the findings in existing literature on the cross border mergers and acquisitions we formulate hypothesis concerning the initiation, completion and duration of deals. A priori, we expect that cultural, institutional, and geographic factors should be significant in explaining the initiation of cross-border acquisitions. We argue that managers in acquiring firms already know about the cultural and institutional differences with countries, and that these differences may lead to possible failures of the deals. Shareholder suffer loss when a deal, which should have been initiated, is not considered and when a deal that is initiated gets abandoned. From manager's perspective, cross border mergers and acquisitions are time consuming and resource intensive and the challenges can often be exacerbated by distances involving culture, institution and geography. Hence, they would avoid initiating deals in countries that are culturally, institutionally and geographically apart.

We further argue that if managers do their due diligence at the initiation of deal and will not initiate a deal that is likely to be abandoned because of cultural, institutional and geographical differences between acquirer and target countries then these factors should not influence the completion and the duration of cross border acquisitions. However, despite of the higher dyadic distances firms may initiate those deals that are of significant value creation by making investment in resources that alleviate the negative impacts of the dyadic distances. The higher the dyadic distances, the higher would be the preparation and hence higher will be the likelihood of completion. Our finding that higher geographical distance increases the likelihood of completion indicates that companies, which are well prepared at the time of initiation of mergers and acquisitions, are able to complete the deals. We infer that for

identifying deal synergies and improving shareholder value, managers should emphasize on due diligence in planning and execution of cross border deals.

Prior literature emphasizes the importance of experience in the success of cross border mergers and acquisitions (see for example Bruton, Oviatt, and White, 1994; Loree, Chen and Guisinger, 2000; and Aktas, de Bodt and Roll, 2013). As a robustness, we investigate if experience with cross border mergers and acquisitions complements the impact of culture, institution and geographical distances in the initiation, completion and duration. We find the initiation of deal between acquirer and target to be unaffected by prior experiences. Initiation depends primarily on cultural, geographical, and institutional differences between countries. We also find that prior experience in completing deals increases the probability of completion for subsequent mergers and acquisitions but does not influence the duration of cross border acquisitions.

One may argue that the effect of cultural, institutional and geographical distances on cross border mergers and acquisitions may be time dependent and may have changed over time. To alleviate this concern, we split our sample into pre- and post-2000 periods. We continue to find negative relationship between initiation and cultural, institutional and geographical distances in both periods. Cultural and institutional distances do not affect the completion of cross border acquisitions whereas the geographical distance increases the likelihood of completion in both periods. Lastly, we find no impact of cross-country cultural, institutional or geographical distances on the duration i.e. the time between the initiation and completion of the deal for both periods. Overall we find that our findings are consistent across time.

The remainder of the paper is organized as follows: Section I discusses the empirical strategy; Section II discusses the empirical results; Section III discusses the robustness of our results; and Section IV concludes.

I. Estimation Strategy

In this section, we provide hypotheses which generates predictions for out empirical model. We then turn to the empirical framework and discusses the statistical implementation of our model.

A. Hypothesis Development

Calori, Lubatkin, and Very (1994) state that organizational structures are part of the acquiring firm's administrative heritage that are rooted in their national culture. Large cultural differences between the acquiring and target countries makes the completion of international acquisition deals to be difficult as there is a greater need for cultural sensitivity in resolving incompatibilities (Morosini, Shane, and Singh, 1998; Very, Lubatkin, Calori, and Veiga, 1997; Weber, Shenkar, and Raveh, 1996). Meyer and Altenborg (2008) states that the firm's culture may be inert and difficult to change, increasing the integration cost for the acquiring firm. Dikovo et al. (2010) state that, it is essential to account for cultural differences in the international mergers and acquisitions, as cultural differences can cause an abandonment of the deal. Ahern et al. (2015) establish a negative relationship between the cultural distances and cross border flows where they find that volume of cross border mergers is lower when countries are culturally distant. We argue that managers in acquiring firms already know about the cultural differences with countries, and hence they would avoid

initiating deal in the countries that are culturally apart. Therefore, cultural differences are likely to reduce the probability of initiation of mergers and acquisitions between country pairs.

Geographical distances play a decisive role in the cross border mergers and acquisitions between countries. This argument is supported by literature on the gravity model of trade that documents geographical distance as important determinant of bilateral trade between countries (Hans Linneman, 1966, Frankel et al., 1995, and Frankel, 1997, Frankel and Romer, 1999). Coval and Moskowitz (2001) state that local investors possess an informational advantage when trading local assets. Ragozzino and Reuer (2011) find that acquirers closer to targets are better positioned to assess the resources of target firms and have a lower risk of adverse selection in an acquisition. The authors suggest that remote acquirers are more likely to "lack key relationships and find appraisals of such soft information problematic" (p. 879). Ahern et al (2015) state that geographic distance is related to the costs of cross-border mergers and empirically show that volume of cross border mergers and acquisitions between country pairs is negatively related to the geographical distance between them.

Kostova and Zaheer (1999) find that the environmental complexity of the transactions involving mergers and acquisitions increases substantially when the country-specific regulatory institutions of the target firm differ significantly from those of the acquirer firm. This is because the acquirer finds it difficult to understand and adjust to legal institutions that are distinct from their home countries. Dikovo et al. (2010) find that deals between firms from countries with very different legal systems increases the likelihood of deal abandonment. We argue that, if the host country legal systems are difficult to comprehend then acquirer may be reluctant to initiate any merger and acquisition in that target country. Based on our discussion above, we hypothesize that dyadic distances of culture, institution and geography should all negatively influence the initiation of deals between country pairs:

Hypothesis 1. *The higher the cultural, institutional and geographical distances between country pairs the lower the probability of initiation of cross border mergers and acquisitions.*

Next, we provide arguments for our hypothesis for the completion of mergers and acquisitions. Completion of mergers and acquisitions deals with the logistics after the company makes its decision to enter in to a contract with the target. Dikovo et al. (2010) study the completion of mergers and acquisitions and state that the environmental complexity of the cross border merger is higher for the transactions involving significantly higher differences in culture and regulatory institutions between acquirer and target countries. The compliance with host-country rules and laws that a foreign acquirer cannot comprehend may obstruct the deal completion. In addition, completing international acquisition deals can be challenging because of the greater need for cultural sensitivity in resolving incompatibilities. Based on these arguments Dekovo et al. (2010) hypothesize that higher the cultural and institutional distances between acquirer and target countries the lower is the probability of completion of deals between these countries. As stated in our first hypothesis the cultural, institutional and geographical distances play a decisive role in the choice of selecting a target country while initiating cross border merger and acquisition. Hence, unlike Dekovo et al. (2010) , we argue that the cultural, institutional and geographical distances between country pairs should not influence the completion of cross border mergers and acquisitions. firm will have due diligence at the initiation of deal and will not initiate a deal that is likely to be abandoned because of cultural, institutional and geographical differences between acquirer and target countries. We do not state that all initiated deals are going to be completed. Our argument is that the reason of incompletion may be due to factors other than the cross country cultural, institutional and geographical distances. We state our second hypothesis formally as:

Hypothesis 2. *The cultural, institutional and geographical distances between country pairs do not influence the likelihood of completion of cross border mergers and acquisitions.*

In our first hypothesis, we argue that when the cultural, institutional and geographical distances are higher company may decide not to initiate a deal. However, firm may initiate value-creating deals despite of the higher dyadic distances. In these cases, firms may invest in resources to alleviate the negative impacts of these distances. The higher the cultural, institutional and geographical distances, the higher are the costs involved in the preparations. Therefore, if the company has invested significant resources for the targets with higher dyadic distances, then the likelihood of completion is higher. However, lack of preparedness may lead to failures of initiated mergers and acquisitions. In line with the above arguments, we modify our second hypothesis and state our third hypothesis as:

Hypothesis 3. *The higher the cultural, institutional and geographical distances between country pairs the higher is the probability of completion.*

Next, we develop our hypothesis for the duration of merger and acquisition, which is the time between initiation and completion of the deal. Prior work in this field is by Dikovo et al. (2010) who argue that cultural and institutional distances produce hold ups resulting in longer deal completion time. Unlike Dekovo et al. (2010), we argue that the dyadic distances between country pairs influence only the initiation of the deal and do not affect its completion. If the cultural, institutional, and geographical distances do not influence completion of deal, they will not affect the time taken to complete the deal also. We state out third hypothesis formally as:

Hypothesis 4. *The cultural, institutional, and geographical distances between country pairs do not influence the duration of cross border mergers and acquisitions.*

B. Cross-Sectional Analysis

To examine how the differences in culture, institutions, and geography affect the initiation, completion, and duration of cross-border merger and acquisition deals we run various specifications of the following logit models at the cross-sectional level:

$$Initiation_{ij} = G(\alpha + \beta \cdot Cult_{ij} + \gamma \cdot Inst_{ij} + \vartheta \cdot Geo_{ij} + \varsigma \cdot X_{ij'}) + \varepsilon_{ij}, \quad (1)$$

$$Completion_{ij} = G(\alpha + \beta \cdot Cult_{ij} + \gamma \cdot Inst_{ij} + \vartheta \cdot Geo_{ij} + \varsigma \cdot X_{ij'}) + \varepsilon_{ij}, \quad (2)$$

$$Duration_{ij} = G(\alpha + \beta \cdot Cult_{ij} + \gamma \cdot Inst_{ij} + \vartheta \cdot Geo_{ij} + \varsigma \cdot X_{ij'}) + \varepsilon_{ij}, \quad (3)$$

where $Cult_{ij}$ denotes our measures of cultural distance (Culture & Institutions, Culture, Power Distance, Individualism, Masculinity, Uncertainty Avoidance, Long-term Orientation, Indulgence, Language, Religion); $Inst_{ij}$ denotes our measures of institutional distance (Culture & Institutions, Corruption, Law and Order, and Bureaucracy); and Geo_{ij} denotes our measures of geography (Distance).

Since the dependent variables range from zero to one, inclusive, we estimate the parameters of interest (β, γ, and ϑ) using a fractional logistic model (Papke and Wooldridge, 2006 ; Wooldridge, 2011). Following the literature, we cluster the standard errors by the target country in all of our regressions.

II. Results

Table 5 displays results of the cross-sectional regressions from Equation (2) related to the initiation of cross-border mergers and acquisitions. We first run the regressions using the fractional logit model as our dependent variable is the probability of initiation of cross border mergers and acquitions and varies between 0 and 1. Additionally we use Poisson pseudo-maximum likelihood methods (PPML) of Silva and Tenreyro (2006) which has been extensively

used in the studies that estimate gravity equation ((Tenreyro (2007), Fally (2013) , Irarrazabal, Moxnes, and Opromolla (2013), Karolyi and Taboda (2015)). As the data of independent variables that we study is retrieved from different sources, there is difference in the number of observations across these variables. In our regressions, we start with the variables that has maximum number of observations (8813 in column 1) and then include other variables based on a decreasing trend in data availability (3844 in column 6). This allow us to maximize the use of independent variable observations thereby improving the power of test. We observe several interesting patterns in the regression results. From the fractional logit results in Column (1), we observe that that the coefficients for language and geographic distance are negative and significant whereas the coefficient for prior experience is significantly positive. Large differences in language is related to lower initiation but differences in religion is not significant. The larger the geographical distance, the less initiation there is between countries; more experience leads to more initiations. In column 4, we report PPML estimates for the equation similar to column 1 and find our results to be qualitatively similar.

Next, we add the differences in economy size (GDP) between the acquirer and target countries to our regressions. This reduces our number of observations to 7862 in column (2). We find the coefficient on size to be negative and highly significant indicating that larger differences in size is related to lower initiations. We have similar results with PPML in column (5).

finally, we add the cultural and institutional differences between acquirer and target countries to the regression equation and report the results in column (3). We observe negative and significant coefficient for power distance, masculinity, uncertainty avoidance, long term orientation, and indulgence. This supports our argument that there is lower probability of initiation of cross border mergers and acquitions if there are larger cultural differences between

nations. We also find that the coefficient for differences in institutional measures of corruption, law and order and bureaucracy are all insignificant. Again, in Column (6) we observe that our results are qualitatively similar using the PPML method.

Overall our results indicate that cultural differences between target and acquirer countries are important determinants in the decision to initiate a cross border merger or acquisitions. Surprisingly, we do not find any evidence of the institutional differences between acquirer and target countries to impact the initiation of cross border mergers. We also find that differences in size of economy, in the language spoken, religion followed and geographical distances among acquirer and target also play an important role in initiation. finally, the significance of experience indicates that past-experience of mergers and acquisitions in the same country increases the probability further acquisitions in that target country.

Table 6 displays results of the cross-sectional regressions from Equation (2) related to the completion of cross-border mergers and acquisitions. We see that differences in individualism, language, religion, and the size of the country affects the completion of the deal. In general, we do not find any effect of cultural factors on the probability of completion of cross border mergers and acquitions. We find that religion difference does matter in probability of completion of cross border mergers and acquitions. There is lower probability of completion when the religion of acquirer and target countries are different. We find that larger the difference in size of the acquirer and target countries the lower is the probability of completion. Among institutional factors, we find that difference in bureaucracy between target and acquirer country affect the probability of completion. For the full sample of observations (5302) we do not find language differences to impact the probability of completion but for the reduced sample with all the variables (2007 observations), we find the coefficient of language to be significantly positive. This empirical finding is counter intuitive as it indicates that if the acquirer and target

countries have same language then the probability of completion of cross border mergers and acquisition is lower. Overall, our findings suggest that cultural factors do not affect the probability of completion, however completion of the deals is mainly driven by bureaucracy. Similar level of bureaucracy across acquirer and target increases the probability of completion of the deal.

In Table 7, we report the results for OLS and PPML regressions with the duration as dependent variable. We do not find any effect of cultural, institutional, or geographical factors on the duration of deals between acquirer and target nation.

Overall our results indicate that the lower the cultural and bureaucratic distance among the nations, the higher the initiation of cross border merger and acquisitions between them. We also find that the completion of deal does not depend on cultural factors however the higher difference in bureaucracy between the nations the lower is the probability of completion. finally we find that the duration of the deals from announcement to effective date does not depend on any cultural or institutional factors.

III. Robustness Analysis

A. Effect of Prior Completion and Duration of Completed Deals

It can be argued that the initiation of deal may depend on whether the firms from the acquiring nation were able to complete mergers and acquisitions with the firms from the target nations in past and how much time did it took for them to complete the deals. If there is a history of successful completion where the deals are completed in a reasonable time, then firms may be motivated to initiate further acquisitions. Similarly, firms may be deterred from

initiating mergers and acquisitions if there is history of failed completions or it takes firms unreasonably longer time to complete acquisitions.

It may also be argued that firms from acquiring nations may learn from prior acquisitions in a target nation and hence completing a deal in any year may depend on its history of successful completions in prior years. Also, all acquiring firms who initiate a deal may want to complete the acquisition in shortest possible time. The history of prior completions may provide them useful information on how to complete the deals in reasonable time.

In the context of our dependent variables based on the above arguments, we expect initiation in any year to be dependent on the completions and the duration of completing the deals in prior years. Similarly, the completion and duration of a deal between acquirer and target nations in any year should be influenced by the completion of deals between the same country pairs in the prior years. As these arguments are based on the level of experience between the acquirer and target nation pairs, we use the data on cross border mergers and acquisitions for the latest year in our data when the experience in completing the deals is at a maximum, 2015. We measure the prior year completion experience as the number of completed deals out of the initiated deals between the acquiring and target nation pairs in all the years prior to 2015. Similarly, we measure duration experience between the acquiring and target nation pairs as the average duration between announcement and the deal effective dates prior to 2015. We include the completion experience and duration experience in our prior models for initiation, completion, and duration and report our results in table 8.

In column (1), we find the initiation of deal between acquirer and target to be unaffected by prior completion and duration experiences. Initiation depends mainly on cultural, geographical, and institutional differences between countries. Next, for completion as dependent variable, in column (2) we find the coefficient of completion experience to be

significantly positive indicating that prior experience in completing deals increases the probability of completion for subsequent mergers and acquisitions. In column (3) where we include prior completion experience in explaining the duration, we find the coefficient for completion experience to be statistically insignificant indicating that prior completion experience the deals does not play any role in deciding the time taken to complete a deal.

B. Different Measures of Distance

To test whether our measures of culture, institutions, and geography are adequete proxies, in Table 9 we re-run our regressions using 9 different measures of distance from Berry, Guillen, and Zhou (2010). $\Delta Culture_{j-i}$ measures the log difference in attitudes toward authority, trust, individuality, and the importance of work and family using data from the World Value Survey. $\Delta Geography_{j-i}$ is the log great circle distance between the geographic center of each country. $\Delta Administrative_{j-i}$ measures the log differences in colonial ties, language, religion, and legal systems. $\Delta Demographic_{j-i}$ measures the differences in demographic characteristics. $\Delta Political_{j-i}$ measures the differences in political stability, democracy, and trade-bloc membership. $\Delta Economic_{j-i}$ measures the difference in economic development and macroeconomic characteristics. $\Delta financial_{j-i}$ measures the difference in financial sector development. $\Delta Integration_{j-i}$ measures the differences in tourism and internet usage. $\Delta Innovation_{j-i}$ measures the differences in patents and scientific production.

Our results using these independently calculated distance measures reconfirm our results. Culture, Geography, Economic, financial, Integration, and Innovative distance are influence inititations. Culture and Economic distance drop out while Political distance influence completions. Only Innovation differences affect duration, while none of the cultural, geographic, or institutional factors matter.

C. Is the Effect Time Dependent?

We investigate cross border mergers and acquitions over a longer time period of 4 decades. It may be argued that the effect of dyadic distances on cross border mergers and acquitions may have changed over time and may be different from the later years than the initial years. To alleviate this concern, we split our sample into 2 time periods: (i) 1985 to 2000 and (ii) 2001 to 2016.[6] We report our results for initiation, completion, and duration respectively in Tables X to XV.

In Table X and XI, we observe that our results are similar for both the time periods and consistent with the overall results reported in Table V. We continue to find negative relationship between initiation and all measures of cultures. A notable difference is in the significance for Individualism and long term orientation. Individualism loses its significance, whereas long term orientation gains the significance in the later time period. Another difference is in the law and order, which has marginal significance in the 1985-2000 period but becomes insignificant in the later time period.

Our sub sample results for completion of deals reported in Table XII and XIII are qualitatively similar in both the time periods and to the overall results in TableVI. There are minor difference in statistical significance only for the coefficients of long term orientation and Indulgence?... They are marginally significant during the 1985 to 2000 period however, in line with the overall results, they are insignificant for the later time period of 2001 to 2016.

[6] Our initial sample on cross border mergers and acquitions starts from 1976. However, cross border mergers were rare events until 1984, where the number of acquitions were less than XXX. From 1985, we observe a significant increase in cross border mergers. Hence, we choose time period from 1985 to 2016 for this robustness analysis.

In Tables XIV and XV, we do not find any significant influence of cultural and institutional factors on the duration of the deal for our sub sample analysis. This is similar to the findings reported in Table VII for overall sample.

Overall our sub sample analysis shows that impact of cultural and institutional factors on initiation, completion, and duration of cross border mergers and acquisitions is consistent over time.

IV. Conclusion

All in all, we contribute to this growing area by studying the effects of differences in cultural, institutional, and geographic factors between an acquirer-target country pair, or the dyadic distance, on the initiation, completion, and duration of cross-border acquisition deals. In particular, using a sample of 173,616 cross-border mergers occurring between 1970 and 2016, we estimate the factors that affect the likelihood that a firm will initiate a cross-border deal, complete the cross-border deal, and the time between the announcement and effective date of the deal (see Section 3). Our study is the first, to our knowledge, to study these three facets of acquisitions – initiation, completion, and duration – within a large-scale unified framework. Our results show that culture, institutions, and geography are important determinants in the initiation and completion of cross-border acquisitions. More specifically, culture differences, institutional, and geographical distance are of paramount importance in whether firms decide to initiate cross-border mergers and acquisitions. Cultural and geographic distance do not play a large role in determining whether firms follow through and complete the deal; what matters in the completion of initiated deals are institutional differences.

Appendix A. Data Definitions

This appendix provides further description of our data sources and variable definitions used in the paper.

Dependent Variables

- ***Initiation***$_{ij}$: The number of cross-border acquisitions from acquirer country *j* to target country *i* relative to the total number of cross-border acquisitions for acquirer country *j*. *Source: Thomson Reuters' SDC Platinum database.*

- ***Completion***$_{ij}$: The proportion of completed cross-border acquisitions from acquirer country *j* to target country *i* relative to the number of cross-border acquisitions from from acquirer country *j* to target country *i*. *Source: Thomson Reuters' SDC Platinum database.*

- ***Duration***$_{ij}$: The average difference between the announcement date and effective date for each acquirer-target pair, removing acquisitions with the same announcement and effective date. *Source: Thomson Reuters' SDC Platinum database.*

Independent Variables

- Δ***Power Distance***$_{j-i}$: Squared difference in Power Distance Index (PDI) scores between acquirer and target scaled by its variance. The PDI expresses the degree to which the less powerful members of a society accept and expect that power is distributed unequally. The fundamental issue here is how a society handles inequalities among people. People in societies exhibiting a large degree of Power Distance accept a hierarchical order in which everybody has a place and which needs no further justification. In societies with low Power Distance, people strive to equalise the distribution of power and demand justification for inequalities of power. *Source: Hofstede (1980).*

- Δ***Individualism***$_{j-i}$: Squared difference in Individualism Index (IDV) scores between acquirer and target scaled by its variance. The high side of this dimension, called individualism, can be defined as a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families. Its opposite, collectivism, represents a preference for a tightly-knit framework in society in which individuals can expect their relatives or members of a particular in-group to look after them in exchange for unquestioning loyalty. *Source: Hofstede (1980).*

- Δ***Masculinity***$_{j-i}$: Squared difference in Masculinity Index (MAS) scores between acquirer and target scaled by its variance. The Masculinity side of this dimension represents a preference in society for achievement, heroism, assertiveness and material rewards for success. Society at large is more competitive. Its opposite, femininity, stands for a preference for cooperation, modesty, caring for the weak and quality of life. Society at large is more consensus-oriented. In the business context Masculinity versus Femininity is sometimes also related to as "tough versus tender" cultures.*Source: Hofstede (1980).*

- Δ***Uncertainty Avoidance***$_{j-i}$: Squared difference in Uncertainty Avoidance Index (UAI) scores between acquirer and target scaled by its variance. The Uncertainty Avoidance dimension expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity. Countries exhibiting strong UAI maintain rigid codes of belief and behaviour and are intolerant of unorthodox behaviour and ideas. Weak UAI societies maintain a more relaxed attitude in which practice counts more than principles.*Source: Hofstede (1980).*

- Δ***Long-Term Orientation***$_{j-i}$: Squared difference in Long-Term Orientation Index (LTO) scores between acquirer and target scaled by its variance. Societies who score low on this dimension, for example, prefer to maintain time-honoured traditions and norms while viewing societal change with suspicion. Those with a culture which scores high, on the other hand, take a more pragmatic approach. *Source: Hofstede (1980).*

- Δ***Indulgence***$_{j-i}$: Squared difference in Indulgence Index (IND) scores between acquirer and target scaled by its variance. Indulgence stands for a society that allows relatively free gratification of basic and natural human drives related to enjoying life and having fun. Restraint stands for a society that suppresses gratification of needs and regulates it by means of strict social norms. *Source: Hofstede (1980).*

- Δ***Culture***$_{j-i}$: Cultural Index score based on Kogut and Singh (1988) index, which uses the differences in the scores on Hofstede's (1980) dimensions of national culture between the acquirer country and target country as individual components. These differences are corrected for differences in the variance of each dimension and then arithmetically averaged. *Source: User calculated using data from using data from Hofstede (1980).*

- Δ***Corruption***$_{j-i}$: Squared difference in Corruption Index scores between acquirer and target scaled by its variance. A measure of corruption within the political system that is a threat to foreign investment by distorting the economic and financial environment, reducing the efficiency of government and business by enabling people to assume positions of power through patronage rather than ability, and introducing inherent instability into the political process. *Source: International Country Risk Guide (ICRG).*

- Δ***Law and Order***$_{j-i}$: Squared difference in Law and Order Index scores between acquirer and target scaled by its variance. There are two measures comprising the Law and Order index. Each sub-component equals half of the total. The "law" subcomponent assesses the strength and impartiality of the legal system, and the "order" sub-component assesses popular observance of the law.*Source: nternational Country Risk Guide (ICRG).*

- Δ***Bureaucracy***$_{j-i}$: Squared difference in Bureaucracy Index scores between acquirer and target scaled by its variance. Institutional strength and quality of the bureaucracy is a shock absorber that tends to minimize revisions of policy when governments change. In low-risk countries, the bureaucracy is somewhat autonomous from political pressure.*Source: International Country Risk Guide (ICRG).*

- Δ***Institutions***$_{j-i}$: Institutions Index score based on Kogut and Singh (1988) index, which uses the differences in the scores on Corruption, Law and ORder, and Bureaucracy between the acquirer country and target country as individual components. These differences are corrected for differences in the variance of each dimension and then arithmetically averaged. *Source: User calculated using data from using data from the International Country Risk Guide (ICRG).*

- Δ***Size***$_{j-i}$: The percentage difference in the gross domestic product (GDP) between acquirer and target. *Source: Mayer and Zignano (2008).*

- ***Language***$_{i}$: Dummy variable coded as 1 if the nation language of acquirer nation is different from target nation, 0 otherwise. *Source: 2000 CIA Factbook.*

- ***Religion***$_{i}$: Dummy variable coded as 1 if the religion of acquirer nation is different from target nation, 0 otherwise. *Source: 2000 CIA Factbook.*

- ***Distance***$_{ij}$: Natural logarithm of the kilometer distance difference between acquirer and target nation. *Source: Mayer and Zignano (2008).*

- ***Experience***$_{ij}$: No. of deals between acquirer and target nation. *Source: Thomson Reuters' SDC Platinum database.*

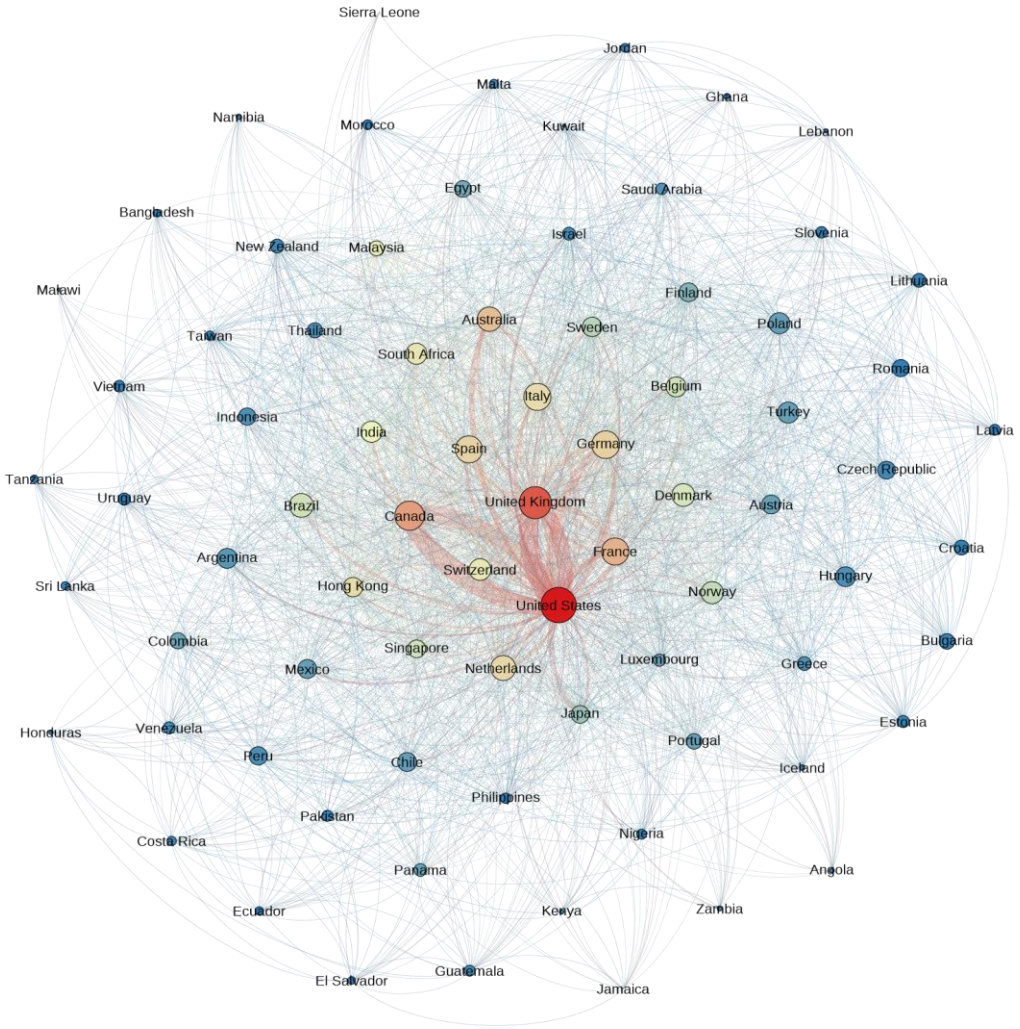


Figure 1. Network of Completed Deals Between Dyads. The lines, or edges, between the country pairs are based on the number of completed acquisitions. The larger the width of the edge, the larger the total number of acquisitions between the countries and vice-versa. The size of the node of each country is based on a measure of importance known as *Authority*. The intensity of the color indicates the importance of the country to the overall network measured by the *Betweenness Centrality*. *Reads: The network of countries which complete cross-border acquisitions is dominated by a few 'countries.*

Table I

Summary Statistics of Cultural Variables.

This tables presents the country-level measures of culture from Hofstede (1980) that are used to construct the distance measures. These variables are the Power Distance Index (PDI), Individualism Index (IDV), Masculinity

Index (MAS), Uncertainty Avoidance Index (UAI), Long-Term Orientation Index (LTO), Indulgence Index (IND), and the Overall Culture Index (OCI).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Country	PDI	IDV	MAS	UAI	LTO	IND	OCI
Afghanistan	41	47	57	75	18	62	1.48
Angola	83	18	20	60	10	83	2.56
Argentina	41	47	57	75	18	62	1.48
Australia	27	99	62	41	22	71	2.40
Austria	0	58	82	60	43	63	2.60
Bangladesh	74	16	56	50	41	20	1.72
Belgium	58	81	54	83	87	57	2.22
Brazil	62	38	49	65	50	59	1.14
Bulgaria	63	28	39	74	72	16	1.83
Canada	30	87	52	38	27	68	1.95
Chile	56	20	26	75	30	68	1.69
China	74	16	68	21	100	24	3.14
Colombia	60	8	66	69	10	83	2.24
Costa Rica	26	11	18	75	—	—	—
Croatia	67	32	39	69	60	33	1.37
Czech Republic	49	61	58	63	73	29	1.54
Denmark	8	80	12	14	34	70	3.37
Dominican Rep	58	28	67	36	10	54	1.10
Ecuador	72	2	64	57	—	—	—
Egypt	63	22	44	69	3	4	2.41
El Salvador	59	15	39	83	18	89	2.23
Estonia	31	64	28	50	87	16	2.35
fiji	72	9	46	38	—	—	—
finland	24	67	23	49	38	57	1.73
France	61	76	42	75	66	48	1.59
Germany	26	72	68	55	57	40	1.71
Ghana	74	11	39	55	0	72	2.18
Greece	53	34	58	100	46	50	1.82
Guatemala	90	0	36	89	—	—	—
Honduras	74	16	39	40	—	—	—
Hong Kong	61	22	58	20	74	17	2.30
Hungary	38	87	92	71	60	31	2.66
Iceland	20	64	6	40	27	67	2.48
India	71	49	57	31	52	26	1.60
Indonesia	72	9	46	38	64	38	1.75
Ireland-Rep	18	75	70	26	22	65	2.38
Israel	2	56	47	70	38	—	—

(Continued on next page)

Table I – continued from previous page

Country	(1) PDI	(2) IDV	(3) MAS	(4) UAI	(5) LTO	(6) IND	(7) OCI
Italy	42	82	72	64	63	30	1.92
Jamaica	37	39	70	5	—	—	—
Japan	46	47	100	81	82	42	2.91
Jordan	63	28	44	55	13	43	1.42
Kenya	63	22	61	40	—	—	—
Kuwait	85	22	39	69	—	—	—
Latvia	35	75	4	53	72	13	2.80
Lebanon	69	40	67	40	11	25	1.86
Lithuania	33	64	16	55	87	16	2.56
Luxembourg	31	64	50	60	67	56	1.46
Malawi	63	28	39	40	—	—	—
Malaysia	100	24	50	27	41	57	2.27
Malta	48	62	47	85	48	66	1.49
Mexico	75	28	71	71	22	97	2.46
Morocco	63	22	53	58	11	25	1.67
Namibia	58	28	39	36	29	—	—
Netherlands	29	87	10	43	53	68	2.48
New Zealand	12	86	59	39	28	75	2.44
Nigeria	74	28	61	45	10	84	2.08
Norway	22	74	3	40	34	55	2.53
Pakistan	47	9	50	60	23	0	2.26
Panama	90	6	43	75	—	—	—
Peru	57	12	41	76	23	46	1.55
Philippines	89	31	66	35	20	42	2.03
Poland	61	64	66	82	31	29	1.71
Portugal	56	25	29	92	27	33	1.93
Romania	85	28	41	79	53	20	1.95
Saudi Arabia	90	22	61	69	36	52	1.83
Serbia	81	22	42	81	53	28	1.77
Sierra Leone	63	16	39	40	—	—	—
Singapore	68	16	48	0	58	46	2.51
Slovenia	65	25	16	77	50	48	1.76
South Africa	41	69	64	39	33	63	1.54
South Korea	53	14	38	74	87	29	2.11
Spain	49	53	41	75	49	44	1.22
Sri Lanka	74	34	6	36	38	—	—
Surinam	80	48	36	81	—	—	—
Sweden	22	76	0	20	40	78	3.31
Switzerland	25	73	72	48	78	66	2.20
Taiwan	51	13	44	59	88	49	1.86
Tanzania	63	22	39	40	33	38	1.37
Thailand	57	16	32	54	40	45	1.35
Trinidad	39	12	59	45	10	80	1.44
Turkey	59	36	44	74	47	49	1.19

United Kingdom	26	98	68	26	52	69	2.62

(Continued on next page)

Table I – continued from previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Country	PDI	IDV	MAS	UAI	LTO	IND	OCI
United States	40	91	62	46	29	68	1.90
Uruguay	54	35	37	88	24	53	1.59
Venezuela	75	7	76	65	13	100	2.98
Vietnam	63	16	39	21	59	35	1.86
Zambia	53	34	39	40	29	42	1.23
Average	53.82	40.21	46.86	55.20	42.01	49.25	2.01

Table II

Summary Statistics of Institutional Variables.

This tables presents the country-level measures of institutions from International Country Risk Guide. These variables are the Corruption Index, the Law and Order Index, the Bureaucracy Index, and the overall Institutions Index.

	(1)	(2)	(3)	(4)
Country	Corruption	Law and Order	Bureaucracy	Institutions
Albania	2	2.5	2	2.06
Algeria	2	3	2	1.90
Angola	1.5	2.5	1.5	2.22
Argentina	2	2.5	3	1.61
Armenia	1.5	3	1	1.59
Australia	4.5	5.5	4	3.32
Austria	4.5	6	4	3.43
Azerbaijan	1.5	3.5	1	1.84
Bahamas	4	4.5	3	1.47
Bahrain	3	4.5	2	1.14
Bangladesh	3	2	2	1.72
Belarus	1.5	3.5	1	2.38
Belgium	5	5	4	3.06
Bolivia	1.5	2.5	2	2.55
Botswana	3.5	3.5	2	1.34
Brazil	2.5	2	2	1.73
Brunei	2.5	5	3.5	1.16
Bulgaria	2	2.5	2	1.73
Burkina Faso	2	3	1	—
Cameroon	2	2	1.5	—
Canada	5	5.5	4	4.09
Chile	4.5	4.5	3	1.75
Colombia	2.5	2	2	1.73
Costa Rica	2.5	3	2	1.36
Croatia	2	4.5	3	1.31
Cuba	2.5	3	2	1.97
Cyprus	4	5	4	2.01
Czech Republic	2.5	5	3	1.28
Denmark	5.5	6	4	4.25
Ecuador	2.5	2.5	2	1.49
Egypt	2	3	2	1.45
El Salvador	2	2	2	2.02
Estonia	3.5	4	2.5	1.11
Ethiopia	1.5	4.5	1.5	1.73

finland	5.5	6	4	4.06
France	4.5	5	3	2.24
Gabon	2	3	1.5	3.12
Gambia	2	3.5	2	1.64
Germany	5	5	4	3.09

(Continued on next page)

Table II – continued from previous page

	(1)	(2)	(3)	(4)
Country	Corruption	Law and Order	Bureaucracy	Institutions
Ghana	2.5	2.5	2.5	1.49
Greece	2	4.5	3	1.36
Guatemala	1.5	1.5	2	2.50
Guinea	1.5	2.5	2	4.31
Guyana	1.5	1.5	3	5.11
Honduras	1.5	1.5	2	2.50
Hong Kong	4.5	5	3	2.11
Hungary	3	4	3	1.09
Iceland	5	6	4	3.52
India	2.5	4	3	1.12
Indonesia	3	3	2	1.25
Iran	1.5	4	2	2.03
Iraq	1	1.5	1.5	4.93
Israel	3.5	5	4	2.11
Italy	2.5	4	2.5	1.10
Jamaica	2	2	3	1.91
Japan	4.5	5	4	2.72
Jordan	2.5	4	2	1.15
Kazakhstan	1.5	3.5	2	1.82
Kenya	1.5	2	2	2.05
Kuwait	2.5	5	2	1.36
Latvia	2.5	5	2.5	1.29
Lebanon	1.5	4	2	1.42
Liberia	2.5	2.5	0	5.85
Libya	1	4	1	2.51
Lithuania	2.5	4	2.5	1.09
Luxembourg	5	6	4	3.69
Malawi	2	2.5	2.5	1.76
Malaysia	2.5	4	3	1.11
Mali	1.5	3	0	3.79
Malta	3.5	5	3	1.44
Mexico	2	1.5	3	2.28
Moldova	2	4.5	1	0.13
Mongolia	2	4	2	1.59
Morocco	2	4.5	2	1.40
Namibia	3	5	2	1.33
Netherlands	5	6	4	4.24

New Zealand	5.5	5.5	4	3.88
Nicaragua	1.5	3.5	1	2.78
Niger	1.5	2	1.5	1.01
Nigeria	1.5	2	1	2.58
Norway	5.5	6	4	4.26
Oman	2.5	5	2	1.25
Pakistan	2	3.5	2	1.41
Panama	2	3	2	1.46

(Continued on next page)

Table II – continued from previous page

	(1)	(2)	(3)	(4)
Country	Corruption	Law and Order	Bureaucracy	Institutions
Paraguay	1.5	2	1	4.99
Peru	2	3	2	1.49
Philippines	2	2.5	3	1.59
Poland	3	4.5	3	1.25
Portugal	3.5	5	3	1.52
Qatar	3	5	2	1.27
Romania	2	4	1	1.88
Saudi Arabia	2.5	5	2	1.32
Senegal	2	3	1	5.10
Sierra Leone	1.5	3.5	0	3.32
Singapore	4.5	5	4	2.60
Slovenia	3.5	4.5	3	1.34
South Africa	2.5	2.5	2	1.39
Spain	4	5	3	1.83
Sri Lanka	2.5	2.5	2	1.49
Sudan	0.5	2.5	1	4.55
Sweden	5.5	6	4	4.27
Switzerland	5	5	4	3.18
Syria	1.5	4.5	1.5	2.04
Taiwan	3	5	3	1.32
Tanzania	2	5	1	2.08
Thailand	2	2.5	2	1.69
Togo	1.5	3	0	1.21
Tunisia	2.5	5	2	1.10
Turkey	2.5	3.5	2	1.19
Uganda	1.5	3.5	2	2.04
Ukraine	1.5	4	1	2.59
United Kingdom	4.5	5	4	3.13
United States	4	5	4	2.68
Uruguay	4	2.5	2	1.70
Venezuela	1	1	1	3.87
Vietnam	2.5	4	2	1.19
Zambia	2.5	4	1	1.73
	1	3	1.5	3.20

Zimbabwe	2.68	3.77	2.32	2.21
Average				

III

Summary Statistics of Main Regression Variables.

Panel A: Country-Level Measures						
	N	M ean	Median	Std. Dev.	Min	Max
Culture$_i$	56	1.77	1.71	1.00	0.00	3.85
Power$_i$	79	2.19	0.91	3.21	0.00	19.86
Individualism$_i$	79	1.89	0.77	2.39	0.00	10.00
Masculinity$_i$	79	1.62	0.73	2.12	0.00	10.36
Uncert. Aviod.$_i$	79	2.05	0.87	2.66	0.00	15.35
Long-Term Orient.$_i$	61	1.39	0.62	1.74	0.00	7.44
Indulgence$_i$	56	2.01	0.90	3.17	0.00	19.53
Institutions$_{j-i}$	112	2.09	1.51	1.98	0.05	9.20
Corruption$_i$	112	2.28	0.68	2.81	0.00	10.95
Law$_i$	112	2.07	1.30	2.47	0.00	11.72
Bureaucracy$_i$	112	1.93	0.82	2.21	0.00	13.15
Panel B: Country-Pair Measures						
	N	Mean	Me dian	Std. Dev.	Min	Max
Initiation$_{ij}$	3,844	0.30	0.01	1.58	0.00	49.13
Completion$_{ij}$	2,007	0.79	0.81	0.18	0.14	1.00
Duration$_{ij}$	2,007	43.60	0.00	155.34	0.00	3,653.00
Δ*Culture*$_{j-i}$	3,844	1.97	1.84	1.13	0.07	6.89
Δ*Power*$_{j-i}$	3,844	1.85	0.91	2.46	0.00	20.68
Δ*Individualism*$_{j-i}$	3,844	2.14	1.06	2.45	0.00	12.02
Δ*Masculinity*$_{j-i}$	3,844	2.29	1.01	3.20	0.00	25.29
Δ*Uncert. Aviod.*$_{j-i}$	3,844	1.93	0.87	2.56	0.00	21.75
Δ*Long-Term Orient.*$_{j-i}$	3,844	1.74	0.90	2.18	0.00	12.92
Δ*Indulgence*$_{j-i}$	3,844	1.89	0.95	2.43	0.00	18.38
Δ*Institutions*$_{j-i}$	3,844	1.84	1.22	1.78	0.00	9.20
Corruption$_i$		3.28	2.50	1.27	1.50	5.50
Law$_i$	3,844	4.17	4.50	1.27	1.50	6.00
Bureaucracy$_i$	3,844	2.81	3.00	0.92	1.00	4.00
Corruption$_j$	3,844	3.30	2.75	1.27	1.50	5.50
Law$_j$	3,844	4.18	4.50	1.28	1.50	6.00
Bureaucracy$_j$	3,844	2.84	3.00	0.90	1.00	4.00
Size$_{ij}$	3,844	−8.51	0.04	49.38	−1,694.95	1.00
Δ*Language*$_{j-i}$	3,844	0.95	1.00	0.22	0.00	1.00
Δ*Religion*$_{j-i}$	3,844	0.72	1.00	0.45	0.00	1.00
Δ*Distance*$_{j-i}$	3,844	8.57	8.89	0.95	4.39	9.89
Experience$_{ij}$	3,844	0.19	0.00	0.27	0.00	1.00

This table presents the summary statistics of our dependent and independent variables of interest. We include both the country-level summary in Panel A and the country-pair data that is actually used in our regressions in Panel B.

Table IV

Correlation Table.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
(1)	1																					
(2)	0.99*	1																				
(3)	0.02	0.02	1																			
(4)	−0.12*	−0.19*	0.01	1	1																	
(5)	−0.06*	−0.10*	0.05*	0.52*	0.35*																	
(6)	−0.07*	−0.12*	0.03	0.53*	0.11*																	
(7)	−0.02	−0.05*	−0.00	0.51*	0.01	1	1															
(8)	−0.04*	−0.08*	−0.01	0.32*	−0.07*	0.00	−0.01															
(9)	−0.07*	−0.08*	−0.02	0.35*	−0.02	0.01	0.01	1														
(10)	−0.07*	−0.09*	−0.04	0.40*	0.42*	0.05*	−0.00	0.10*	1													
(11)	−0.10*	−0.16*	0.02	0.38*	0.11*	0.05*	−0.00	0.08*	0.20*	1	1											
(12)	0.12*	0.13*	−0.03	0.13*	0.12*	0.37*	0.09*	0.01	0.03*	0.09*	0.11*	1			1							
(13)	0.09*	0.10*	−0.03	0.10*	0.13*	0.10*	0.19*	0.10*	−0.06*	−0.14*	0.02	0.74*	1		0.74*							
(14)	0.14*	0.16*	−0.01	0.17*	0.14*	0.05*	0.19*	0.09*	−0.08*	−0.17*	0.05*	0.82*	0.69*	1	0.82*							
(15)	0.16*	0.12*	0.01	0.15*	0.13*	0.09*	0.21*	0.11*	−0.04*	−0.12*	0.13*	−0.02	−0.01	−0.01	0.04*							
(16)	0.13*	0.10*	0.01	0.10*	0.17*	0.10*	0.19*	0.11*	−0.07*	−0.14*	0.03	−0.01	−0.02	−0.02		1						
(17)	0.19*	0.16*	0.03	0.18*	0.05*	0.05*	0.19*	0.09*	−0.08*	−0.17*	0.03	−0.01	−0.01	−0.14*	−0.01	0.70*	1					
(18)	0.02	0.02	0.01	−0.00	−0.01	0.10*	0.20*	0.12*	−0.05*	−0.11*	0.08*	−0.01	−0.01	−0.02		0.01	0.06*	1				
(19)	−0.19*	−0.24*	−0.05	0.09*	0.06*	0.03*	−0.05*	0.03	0.00	0.01	0.01	−0.09*	−0.06*	−0.08*	0.07*	0.03	−0.07*	0.08*	1			
(20)	−0.09*	−0.11*	−0.02	0.07*	0.09*	0.07*	0.12*	0.10*	0.09*	−0.01	−0.01	0.01	0.03	−0.12*	0.09*	0.05*	−0.12*	0.02	0.07*	1		
(21)	−0.15*	−0.15*	0.03	0.17*	0.06*	0.07*	−0.05*	0.02	0.08*	0.04*	−0.00	−0.16*	−0.21*			0.22*		0.01	0.08*	0.03*	1	
(22)	0.08*	−0.03	0.02	0.00	0.09*	0.21*	−0.11*	0.11*	0.07*	0.13*	0.18*	−0.09*	−0.23*			0.21*	0.28*				0.21*	1
					0.01	0.09*	0.04*	0.01	−0.08*	−0.09*	0.03	0.03	0.00	0.04*	0.24*							

(1)	*Initiation*$_{ij}$	(12)	*Corruption*$_i$
(2)	*Completion*$_{ij}$	(13)	*Law and Order*$_i$
(3)	*Duration*$_{ij}$	(14)	*Bureaucracy*$_i$
(4)	Δ*Culture*$_{j-i}$	(15)	*Corruption*$_j$
(5)	Δ*Power*$_{j-i}$	(16)	*Law and Order*$_j$
(6)	Δ*Individualism*$_{j-i}$ Δ*Masculinity*$_{j-i}$	(17)	*Bureaucracy*$_j$
(7)		(18)	
(8)	Δ*Uncert. Aviod.*$_{j-i}$	(19)	Δ*Size*$_{j-i}$
(9)	Δ*Long-Term Orientation*$_{j-i}$	(20)	Δ*Language*$_{j-i}$
(10)	$\Delta\Delta$*Indulgence*$_{jj--ii}$	(21)	Δ*Religion*$_{j-i}$
(11)	*Institutions*	(22)	*Distance*$_{ij}$
			Experience$_{ij}$

This table presents the correlation matrix between our dependent and independent variables of interest. Significance at 10% is denoted by (*).

V

Cross-section of Deal Initiations.

	Fractional Logit			PPML		
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	0.287***	0.290***	0.968***	0.279***	0.288***	0.967***
	(0.099)	(0.105)	(0.112)	(0.076)	(0.082)	(0.092)
ΔLanguage$_{j-i}$	−0.822***	−0.891***	−0.909***	−0.762***	−0.833***	−0.846***
	(0.125)	(0.125)	(0.112)	(0.087)	(0.086)	(0.093)
ΔReligion$_{j-i}$	−0.236*	−0.270*	−0.075	−0.234**	−0.271**	−0.085
	(0.132)	(0.140)	(0.129)	(0.107)	(0.117)	(0.095)
ΔDistance$_{j-i}$	−0.853***	−0.874***	−0.999***	−0.784***	−0.803***	−0.933***
	(0.048)	(0.048)	(0.045)	(0.028)	(0.030)	(0.029)
ΔGDP$_{j-i}$		0.004	−0.015***		0.004	−0.015***
		(0.004)	(0.005)		(0.003)	(0.004)
ΔPower$_{j-i}$			−0.015			−0.015
			(0.020)			(0.016)
ΔIndv.$_{j-i}$			−0.033*			−0.029**
			(0.020)			(0.014)
ΔMasc.$_{j-i}$			−0.067***			−0.065***
			(0.012)			(0.009)
ΔUncert. Aviod.$_{j-i}$			−0.129***			−0.126***
			(0.020)			(0.013)
ΔLong Term Orient.$_{j-i}$			−0.053*			−0.053**
			(0.029)			(0.023)
ΔIndulgence$_{j-i}$			−0.057**			−0.055***
			(0.026)			(0.018)
ΔCorruption$_{j-i}$			−0.056**			−0.057***
			(0.022)			(0.018)
ΔLaw$_{j-i}$			0.021			0.016
			(0.025)			(0.018)
ΔBureaucracy$_{j-i}$			0.004			0.012
			(0.038)			(0.032)
Constant	1.691***	2.219***	2.220***	0.791	1.263**	−0.048
	(0.506)	(0.531)	(0.657)	(0.499)	(0.575)	(0.799)
Obs.	8,813	8,387	3,969	8,736	8,311	3,969
(pseudo) R^2	0.347	0.336	0.233	0.749	0.736	0.743

The dependent variables is the number of cross-border acquisitions from acquirer country *j* to target country *i* relative to the total number of cross-border acquisitions for acquirer country *j*. We use a logit fractional response and PPML model to estimate the cross sectional regressions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

Table VI

Cross-section of Deal Completions.

	Fractional Logit			PPML		
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	−0.030	−0.070	−0.095	−0.006	−0.010	−0.019
	(0.073)	(0.075)	(0.113)	(0.010)	(0.010)	(0.021)
ΔLanguage$_{j-i}$	0.116	0.072	0.083	0.019	0.011	0.015
	(0.085)	(0.087)	(0.122)	(0.013)	(0.013)	(0.021)
ΔReligion$_{j-i}$	−0.002	0.028	0.179	0.005	0.009	0.041*
	(0.077)	(0.080)	(0.148)	(0.012)	(0.012)	(0.022)
ΔDistance$_{j-i}$	0.275***	0.285***	0.228***	0.051***	0.053***	0.049***
	(0.035)	(0.037)	(0.046)	(0.004)	(0.004)	(0.007)
ΔGDP$_{j-i}$		0.008*	0.006		0.001***	0.001
		(0.004)	(0.006)		(0.001)	(0.001)
ΔPower$_{j-i}$			0.009			0.003
			(0.015)			(0.003)
ΔIndv.$_{j-i}$			−0.042***			−0.010***
			(0.014)			(0.003)
ΔMasc.$_{j-i}$			−0.014			−0.002
			(0.009)			(0.001)
ΔUncert. Aviod.$_{j-i}$			−0.002			0.000
			(0.012)			(0.002)
ΔLong Term Orient.$_{j-i}$			0.008			0.001
			(0.020)			(0.003)
ΔIndulgence$_{j-i}$			0.023			0.004
			(0.020)			(0.003)
ΔCorruption$_{j-i}$			0.002			0.000
			(0.023)			(0.004)
ΔLaw$_{j-i}$			0.026			0.004
			(0.021)			(0.004)
ΔBureaucracy$_{j-i}$			−0.077**			−0.016**
			(0.034)			(0.006)
Constant	17.660***	16.390***	−0.515	−0.226	−0.153	−0.322***
	(1.456)	(1.474)	(0.559)	(0.204)	(0.203)	(0.079)
Obs.	5,284	4,962	2,042	5,284	4,962	2,042
(pseudo) R^2	0.090	0.086	0.044	0.270	0.265	0.210

The dependent variables is the proportion of completed cross-border acquisitions from acquirer country *j* to target country *i* relative to the number of cross-border acquisitions from from acquirer country *j* to target country *i*. We use a logit fractional response model to estimate the cross sectional regres-

sions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

VII

Cross-section of Deal Duration.

	Fractional Logit			PPML		
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	−3.718	−2.876	6.965	−0.064	−0.043	0.171
	(4.780)	(5.043)	(8.683)	(0.096)	(0.100)	(0.188)
ΔLanguage$_{j-i}$	8.111	11.500	−1.523	0.126	0.191*	−0.034
	(7.045)	(7.543)	(8.647)	(0.106)	(0.109)	(0.162)
ΔReligion$_{j-i}$	−1.780	−2.558	−10.280	−0.043	−0.067	−0.276*
	(5.457)	(5.503)	(6.894)	(0.102)	(0.101)	(0.161)
ΔDistance$_{j-i}$	0.083	0.506	−4.284**	0.006	0.015	−0.081*
	(1.978)	(2.036)	(2.129)	(0.037)	(0.039)	(0.047)
ΔGDP$_{j-i}$		−0.109	−0.505		−0.003	−0.011
		(0.263)	(0.704)		(0.006)	(0.014)
ΔPower$_{j-i}$			2.101			0.056**
			(1.620)			(0.026)
ΔIndv.$_{j-i}$			0.006			−0.008
			(1.260)			(0.024)
ΔMasc.$_{j-i}$			0.631			0.025
			(0.684)			(0.021)
ΔUncert. Aviod.$_{j-i}$			0.690			0.019
			(0.832)			(0.016)
ΔLong Term Orient.$_{j-i}$			−0.460			−0.007
			(0.890)			(0.027)
ΔIndulgence$_{j-i}$			−0.761			−0.023
			(0.720)			(0.025)
ΔCorruption$_{j-i}$			−1.622			−0.046
			(1.394)			(0.037)
ΔLaw$_{j-i}$			1.676			0.039
			(2.315)			(0.042)
ΔBureaucracy$_{j-i}$			2.708			0.072
			(2.354)			(0.051)
Constant	60.080	45.600	13.590	4.112***	3.815***	3.691***
	(54.500)	(57.100)	(35.260)	(0.886)	(0.963)	(1.231)
Obs.	5,285	4,963	2,043	5,197	4,879	2,034
(pseudo) R^2	0.169	0.174	0.091	0.266	0.286	0.139

The dependent variables is the average difference between the announcement date and effective date for each acquirer-target pair. We use an OLS model to estimate the cross sectional regressions. These measures

are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

Table VIII

Robustness to Prior Experience.

	Initiation		Completion	Duration		
	(1)	(2)	(3)	(4)	(5)	(6)
	FL	PPML	FL	PPML	OLS	PPML
Experience$_{ij}$	−0.097	−0.095	0.248	0.021	13.840	0.432*
	(0.104)	(0.081)	(0.215)	(0.024)	(9.463)	(0.229)
ΔLanguage$_{j-i}$	−0.026	−0.024	0.013	0.005	−0.437	−0.008
	(0.020)	(0.018)	(0.045)	(0.007)	(1.316)	(0.044)
ΔReligion$_{j-i}$	−0.089***	−0.086***	−0.036	−0.008	0.601	0.023
	(0.022)	(0.017)	(0.035)	(0.006)	(1.285)	(0.034)
ΔDistance$_{j-i}$	−0.062***	−0.059***	0.056***	0.008***	−0.635	−0.022
	(0.010)	(0.009)	(0.019)	(0.003)	(0.537)	(0.020)
ΔGDP$_{j-i}$	−0.070***	−0.068***	−0.052**	−0.007	−0.313	−0.018
	(0.023)	(0.015)	(0.024)	(0.004)	(0.979)	(0.029)
ΔPower$_{j-i}$	−0.057**	−0.054**	−0.017	−0.004	0.436	0.016
	(0.027)	(0.025)	(0.045)	(0.006)	(2.050)	(0.048)
ΔIndv.$_{j-i}$	0.018	0.017	−0.052	−0.004	−1.838	−0.058
	(0.034)	(0.029)	(0.045)	(0.007)	(1.739)	(0.051)
ΔMasc.$_{j-i}$	−0.047*	−0.050**	0.118*	0.015*	0.075	0.013
	(0.029)	(0.022)	(0.063)	(0.008)	(2.257)	(0.069)
ΔUncert. Aviod.$_{j-i}$	0.021	0.021	0.062	0.003	1.013	−0.002
	(0.048)	(0.036)	(0.085)	(0.009)	(2.791)	(0.077)
ΔLong Term Orient.$_{j-i}$	−0.088*	−0.080*	−0.105	−0.016	3.857	0.199*
	(0.049)	(0.045)	(0.097)	(0.013)	(3.178)	(0.116)
ΔIndulgence$_{j-i}$	−0.694***	−0.634***	0.272	0.051	−2.809	−0.082
	(0.074)	(0.081)	(0.225)	(0.034)	(6.061)	(0.205)
ΔCorruption$_{j-i}$	0.194	0.182	0.331	0.054	0.721	0.122
	(0.147)	(0.122)	(0.301)	(0.048)	(9.585)	(0.313)
ΔLaw$_{j-i}$	−0.585***	−0.538***	0.262***	0.049***	−0.426	−0.037
	(0.048)	(0.030)	(0.076)	(0.012)	(2.153)	(0.077)
ΔBureaucracy$_{j-i}$	−0.019***	−0.018***	0.015*	0.002	−0.185	−0.007
	(0.004)	(0.006)	(0.008)	(0.003)	(0.265)	(0.013)
ΔCompletion Exp.$_{j-i}$	0.237	0.211	2.130***	0.220***	18.180	0.936
	(0.367)	(0.289)	(0.819)	(0.084)	(23.660)	(0.696)
ΔDuration Exp.$_{ij}$	0.000	0.000	—	—	—	—
	(0.001)	(0.001)	—	—	—	—

Constant	0.438	−3.860***	12.700***	−0.525***	−61.650*	3.269***
	(0.454)	(0.581)	(1.348)	(0.203)	(31.600)	(1.263)
Obs.	901	901	782	782	782	760
(pseudo) R^2	0.110	0.796	0.110	0.281	0.219	0.280

The dependent variables is the average difference between the announcement date and effective date for each acquirer-target pair. We use an OLS model to estimate the cross sectional regressions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

Table

IX

Robustness using Different Measures of Distance.

	(1)	(2)	(3)
	$Initiation_{ij}$	$Completion_{ij}$	$Duration_{ij}$
$\Delta Culture_{j-i}$	−0.131*	−0.117	−3.449
	(0.075)	(0.074)	(5.269)
$\Delta Geography_{j-i}$	−0.354***	−0.332***	0.999
	(0.040)	(0.062)	(3.746)
$\Delta Administrative_{j-i}$	−0.207***	−0.233***	−3.069
	(0.025)	(0.047)	(2.323)
$\Delta Demographic_{j-i}$	−0.398***	−0.355***	4.214
	(0.044)	(0.077)	(5.514)
$\Delta Political_{j-i}$	−0.040	−0.238***	−6.959
	(0.037)	(0.091)	(4.915)
$\Delta Economic_{j-i}$	−0.194***	−0.061	3.165
	(0.054)	(0.065)	(2.779)
$\Delta financial_{j-i}$	0.090***	0.187***	−4.515
	(0.035)	(0.054)	(3.835)
$\Delta Integration_{j-i}$	0.064*	−0.128**	8.009
	(0.038)	(0.057)	(5.120)
$\Delta Innovation_{j-i}$	0.071***	0.253***	−5.271***
	(0.015)	(0.026)	(1.839)
Constant	0.058	4.859***	160.900***
	(0.328)	(0.555)	(32.740)
N	5,938	5,938	2,032
(Pseudo) R^2	0.051	0.118	0.012

The dependent variables are: (1) $Initiation_{ij}$: the number of cross-border acquisitions from acquirer country j to target country i relative to the total number of cross-border acquisitions for acquirer country j; (2) $Completion_{ij}$: the proportion of completed cross-border acquisitions from acquirer country j to target country i relative to the number of cross-border acquisitions from from acquirer country j to target country i; and (3) $Duration_{ij}$: the average difference between the announcement date and effective date for each acquirer-target pair, removing acquisitions with the same announcement and effective date. We use a logit fractional response model for models (1) and (2) and an OLS model for model (3). The cross-national distance measures are from Berry, Guillen, and Zhou (2010). $^{\Delta}Culture^{j-i}$ importancemeasures theof worklog differenceand family.in attitudes toward authority, trust, individuality, and the

Table

ligion,country.Δ*Geography*andΔ*Administrative*legal*j*−*i* issystems. the log*j*−greatΔ*i**Demographic*measurescircle distancethe*j*−log*i*differencesmeasuresdifferencesbetweeninthethepoliticalindifferencesgeographiccolonialstability,ties,incenterdemographiclanguage,democracy,of eachre-

characteristics. Δ*Political**j*−*i* measures the
and trade-bloc membership. Δ*Economic**j*−*i* measures the difference in economic de-

velopmentin
financialandsectormacroeconomicdevelopment.characteristics.Δ*Integration**j*−Δ*i*
*financial*measuresdifferences*j*−the*i*
measuresindifferencespatentstheandindifferencescientifictourism and internet usage. Δ*Innovation**j*−*i* measures the
production. These measures are constructed using the Mahalanobis method. Timevarying distance measures use the time-varying covariance matrix in computing the distance and are averaged across each acquirer-target pair. Significance at 10%, 5 %, and 1% is denoted by (*, **, ***) with t-statistics based on standard errors clustered on 84 target countries in parentheses.

X

Cross-section of Deal Initiations, 1985-2000.

Fractional Logit

	(1)	(2)	(3)	(4)	(5)	(6)
Experience*ij*	−0.016	−0.040	0.126	−0.013	−0.033	0.129
	(0.090)	(0.094)	(0.151)	(0.072)	(0.076)	(0.102)
Δ Language*j*−*i*	−0.612***	−0.696***	−0.662***	−0.543***	−0.632***	−0.614***
	(0.132)	(0.143)	(0.141)	(0.095)	(0.101)	(0.111)
Δ Religion*j*−*i*	−0.086	−0.169	0.230	−0.099	−0.181	0.186
	(0.168)	(0.186)	(0.205)	(0.120)	(0.134)	(0.141)
Δ Distance*j*−*i*	−0.728***	−0.733***	−0.736***	−0.653***	−0.652***	−0.666***
	(0.045)	(0.044)	(0.044)	(0.033)	(0.034)	(0.033)
Δ GDP*j*−*i*		0.005	−0.016**		0.005	−0.013**
		(0.004)	(0.006)		(0.004)	(0.006)
Δ Power*j*−*i*			−0.027			−0.026
			(0.022)			(0.018)
Δ Indv.*j*−*i*			−0.075***			−0.072***
			(0.025)			(0.017)
Δ Masc.*j*−*i*			−0.050***			−0.048***
			(0.013)			(0.009)
Δ Uncert. Aviod.*j*−*i*			−0.090***			−0.086***
			(0.024)			(0.016)
Δ Long Term Orient.*j*−*i*			0.011			0.013
			(0.040)			(0.029)
Δ Indulgence*j*−*i*			−0.077**			−0.068***
			(0.030)			(0.022)

Table

ΔCorruption$_{j-i}$			−0.067**			−0.065***
			(0.027)			(0.022)
ΔLaw$_{j-i}$			0.088**			0.082***
			(0.036)			(0.024)
ΔBureaucracy$_{j-i}$			−0.089*			−0.083**
			(0.048)			(0.036)
Constant	2.178***	2.870***	20.570***	0.954**	1.590***	1.465***
	(0.629)	(0.720)	(1.215)	(0.443)	(0.516)	(0.466)
Obs.	2,893	2,762	1,508	2,893	2,762	1,508
(pseudo) R^2	0.369	0.361	0.205	0.860	0.854	0.743

The dependent variables is the number of cross-border acquisitions from acquirer country *j* to target country *i* relative to the total number of cross-border acquisitions for acquirer country *j*. We use a logit fractional response and PPML model to estimate the cross sectional regressions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

XI

Cross-section of Deal Initiations, 2001-2015.

	Fractional Logit					
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	0.242**	0.240**	0.857***	0.226***	0.231***	0.858***
	(0.104)	(0.113)	(0.099)	(0.074)	(0.080)	(0.091)
ΔLanguage$_{j-i}$	−0.821***	−0.845***	−0.956***	−0.762***	−0.785***	−0.891***
	(0.135)	(0.134)	(0.113)	(0.098)	(0.098)	(0.090)
ΔReligion$_{j-i}$	−0.256*	−0.292**	−0.134	−0.264**	−0.299***	−0.142
	(0.137)	(0.141)	(0.118)	(0.112)	(0.115)	(0.089)
ΔDistance$_{j-i}$	−0.866***	−0.890***	−1.001***	−0.795***	−0.818***	−0.936***
	(0.049)	(0.050)	(0.049)	(0.031)	(0.032)	(0.029)
ΔGDP$_{j-i}$		0.004	−0.016***		0.003	−0.016***
		(0.004)	(0.005)		(0.003)	(0.004)
ΔPower$_{j-i}$			−0.008			−0.009
			(0.019)			(0.016)
ΔIndv.$_{j-i}$			−0.033*			−0.029**
			(0.019)			(0.014)
ΔMasc.$_{j-i}$			−0.066***			−0.064***
			(0.012)			(0.009)

Table

ΔUncert. Aviod.$_{j-i}$			−0.121***			−0.118***
			(0.019)			(0.013)
ΔLong Term Orient.$_{j-i}$			−0.061**			−0.061***
			(0.028)			(0.022)
ΔIndulgence$_{j-i}$			−0.055**			−0.053***
			(0.025)			(0.018)
ΔCorruption$_{j-i}$			−0.052**			−0.052***
			(0.022)			(0.018)
ΔLaw$_{j-i}$			0.013			0.009
			(0.021)			(0.018)
ΔBureaucracy$_{j-i}$			−0.011			−0.004
			(0.036)			(0.031)
Constant	3.040***	3.405***	2.485***	1.584***	1.845***	1.660***
	(0.440)	(0.456)	(0.662)	(0.540)	(0.581)	(0.491)
Obs.	8,279	7,892	3,832	8,202	7,816	3,832
(pseudo) R^2	0.367	0.352	0.229	0.791	0.772	0.752

The dependent variables is the number of cross-border acquisitions from acquirer country *j* to target country *i* relative to the total number of cross-border acquisitions for acquirer country *j*. We use a logit fractional response and PPML model to estimate the cross sectional regressions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

XII

Cross-section of Deal Completions, 1985-2000.

	Fractional Logit					
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	0.033	0.036	0.031	0.005	0.004	0.003
	(0.095)	(0.102)	(0.155)	(0.014)	(0.014)	(0.023)
ΔLanguage$_{j-i}$	0.089	0.056	0.174	0.013	0.007	0.035
	(0.112)	(0.110)	(0.107)	(0.018)	(0.018)	(0.028)
ΔReligion$_{j-i}$	−0.114	−0.070	0.305	−0.006	−0.002	0.062*
	(0.145)	(0.150)	(0.189)	(0.019)	(0.020)	(0.032)
ΔDistance$_{j-i}$	0.315***	0.326***	0.264***	0.055***	0.056***	0.052***
	(0.038)	(0.040)	(0.047)	(0.005)	(0.006)	(0.008)
ΔGDP$_{j-i}$		0.009**	0.003		0.001	0.001
		(0.004)	(0.005)		(0.001)	(0.001)

Table

ΔPower$_{j-i}$			0.016			0.005
			(0.019)			(0.004)
ΔIndv.$_{j-i}$			−0.070***			−0.013***
			(0.019)			(0.004)
ΔMasc.$_{j-i}$			−0.009			−0.002
			(0.015)			(0.002)
ΔUncert. Aviod.$_{j-i}$			0.023			0.005
			(0.019)			(0.003)
ΔLong Term Orient.$_{j-i}$			−0.057**			−0.010**
			(0.025)			(0.005)
ΔIndulgence$_{j-i}$			0.048**			0.007*
			(0.024)			(0.004)
ΔCorruption$_{j-i}$			0.007			0.003
			(0.030)			(0.005)
ΔLaw$_{j-i}$			0.051**			0.007
			(0.025)			(0.005)
ΔBureaucracy$_{j-i}$			−0.063			−0.012
			(0.056)			(0.008)
Constant	34.270	34.870***	14.050***	−0.021	0.037	−0.380***
	(.)	(2.542)	(1.481)	(0.067)	(0.084)	(0.131)
Obs.	2,649	2,530	1,397	2,649	2,530	1,397
(pseudo) R^2	0.104	0.104	0.061	0.278	0.278	0.224

The dependent variables is the proportion of completed cross-border acquisitions from acquirer country *j* to target country *i* relative to the number of cross-border acquisitions from from acquirer country *j* to target country *i*. We use a logit fractional response model to estimate the cross sectional regres-
sions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

Table XIII

Cross-section of Deal Completions, 2001-2015.

	Fractional Logit			PPML		
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	−0.017	−0.066	−0.218*	−0.007	−0.012	−0.040**
	(0.083)	(0.084)	(0.127)	(0.010)	(0.011)	(0.020)
ΔLanguage$_{j-i}$	0.085	0.045	0.038	0.009	0.003	0.005
	(0.081)	(0.086)	(0.139)	(0.014)	(0.014)	(0.023)
ΔReligion$_{j-i}$	0.059	0.093	0.173	0.014	0.019	0.035
	(0.089)	(0.090)	(0.154)	(0.012)	(0.012)	(0.023)
ΔDistance$_{j-i}$	0.291***	0.295***	0.261***	0.052***	0.052***	0.054***
	(0.038)	(0.040)	(0.047)	(0.004)	(0.004)	(0.007)
ΔGDP$_{j-i}$		0.005	0.004		0.001*	0.000
		(0.004)	(0.006)		(0.000)	(0.001)
ΔPower$_{j-i}$			0.021			0.005*
			(0.015)			(0.003)
ΔIndv.$_{j-i}$			−0.054***			−0.012***
			(0.016)			(0.003)
ΔMasc.$_{j-i}$			−0.014			−0.002
			(0.012)			(0.002)
ΔUncert. Aviod.$_{j-i}$			−0.008			−0.001
			(0.012)			(0.002)
ΔLong Term Orient.$_{j-i}$			0.018			0.003
			(0.023)			(0.003)
ΔIndulgence$_{j-i}$			−0.022			−0.004
			(0.021)			(0.004)
ΔCorruption$_{j-i}$			0.010			0.002
			(0.025)			(0.004)
ΔLaw$_{j-i}$			0.018			0.002
			(0.024)			(0.004)
ΔBureaucracy$_{j-i}$			−0.048			−0.010
			(0.035)			(0.006)
Constant	34.040***	31.920***	−1.354**	−0.041	−0.005	−0.399***
	(2.254)	(1.512)	(0.569)	(0.101)	(0.079)	(0.135)
Obs.	4,767	4,479	1,899	4,767	4,479	1,899
(pseudo) R^2	0.101	0.097	0.058	0.284	0.281	0.242

The dependent variables is the proportion of completed cross-border acquisitions from acquirer country *j* to target country *i* relative to the number of cross-border acquisitions from from acquirer

country *j* to target country *i*. We use a logit fractional response model to estimate the cross sectional regres-
sions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

Table XIV

Cross-section of Deal Duration, 1985-2000.

	Fractional Logit			PPML		
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	0.058	0.754	6.648	0.039	0.064	0.253
	(7.975)	(8.344)	(12.950)	(0.154)	(0.159)	(0.227)
ΔLanguage$_{j-i}$	−9.173	−3.594	3.595	−0.118	−0.021	0.035
	(10.760)	(11.070)	(11.350)	(0.158)	(0.166)	(0.218)
ΔReligion$_{j-i}$	5.777	6.250	−13.780	0.024	−0.000	− 0.360
	(13.440)	(14.090)	(11.850)	(0.169)	(0.170)	(0.268)
ΔDistance$_{j-i}$	−5.021	−4.747	0.255	−0.111*	−0.108*	0.032
	(3.959)	(4.077)	(3.516)	(0.060)	(0.061)	(0.078)
ΔGDP$_{j-i}$		0.519	−0.018		0.003	− 0.003
		(0.713)	(0.623)		(0.011)	(0.016)
ΔPower$_{j-i}$			−2.241			− 0.038
			(2.534)			(0.031)
ΔIndv.$_{j-i}$			1.138			0.007
			(1.738)			(0.030)
ΔMasc.$_{j-i}$			0.428			0.009
			(0.777)			(0.020)
ΔUncert. Aviod.$_{j-i}$			−0.406			− 0.009
			(1.287)			(0.029)
ΔLong Term Orient.$_{j-i}$			−1.610			− 0.030
			(1.939)			(0.040)
ΔIndulgence$_{j-i}$			0.550			0.006
			(1.888)			(0.038)
ΔCorruption$_{j-i}$			−1.692			− 0.041
			(2.951)			(0.057)
ΔLaw$_{j-i}$			4.922			0.110*
			(3.625)			(0.056)
ΔBureaucracy$_{j-i}$			4.988			0.099
			(4.150)			(0.071)
Constant	199.100***	233.600***	−87.670*	4.517***	4.692***	2.559
	(47.190)	(85.800)	(52.280)	(0.949)	(1.404)	(1.584)

Obs.	2,644	2,525	1,394	2,514	2,410	1,378
(pseudo) R^2	0.267	0.268	0.112	0.411	0.415	0.180

The dependent variables is the average difference between the announcement date and effective date for each acquirer-target pair. We use an OLS model to estimate the cross sectional regressions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

Table XV

Cross-section of Deal Duration, 2001-2015.

	Fractional Logit			PPML		
	(1)	(2)	(3)	(4)	(5)	(6)
Experience$_{ij}$	−1.560	−1.282	9.197	−0.014	−0.009	0.238
	(4.947)	(5.167)	(7.505)	(.)	(.)	(0.187)
ΔLanguage$_{j-i}$	8.805*	10.680**	5.720	0.192	0.229	0.107
	(4.495)	(4.655)	(10.100)	(.)	(.)	(0.172)
ΔReligion$_{j-i}$	−0.506	−0.523	−9.369	−0.031	−0.034	−0.217
	(5.749)	(5.671)	(8.641)	(.)	(.)	(0.179)
ΔDistance$_{j-i}$	−0.311	0.182	−4.217	0.001	0.013	−0.086*
	(1.718)	(1.784)	(2.564)	(.)	(.)	(0.051)
ΔGDP$_{j-i}$		0.011	−0.553		−0.001	−0.013
		(0.298)	(0.796)		(.)	(0.014)
ΔPower$_{j-i}$			3.557*			0.085**
			(1.908)			(0.035)
ΔIndv.$_{j-i}$			−0.102			0.001
			(1.152)			(0.023)
ΔMasc.$_{j-i}$			0.618			0.024*
			(0.394)			(0.013)
ΔUncert. Aviod.$_{j-i}$			−0.001			−0.001
			(1.020)			(0.019)
ΔLong Term Orient.$_{j-i}$			−0.716			−0.020
			(1.411)			(0.033)
ΔIndulgence$_{j-i}$			−2.943			−0.067*
			(2.133)			(0.037)
ΔCorruption$_{j-i}$			−1.673			−0.055*
			(1.388)			(0.031)
ΔLaw$_{j-i}$			0.178			0.005
			(1.990)			(0.033)
ΔBureaucracy$_{j-i}$			1.124			0.047
			(2.370)			(0.054)

Constant	22.530	16.290	31.420	−15.520	−15.710	2.852**
	(15.210)	(20.980)	(40.150)	(.)	(.)	(1.320)
Obs.	4,767	4,479	1,899	4,681	4,398	1,890
(pseudo) R^2	0.151	0.158	0.107	0.202	0.215	0.174

The dependent variables is the average difference between the announcement date and effective date for each acquirer-target pair. We use an OLS model to estimate the cross sectional regressions. These measures are constructed using the Mahalanobis method restricting the covariance terms to zero following Kogut and Singh (1988). Significance at 10%, 5%, and 1% is denoted by (*, **, ***) with standard errors clustered on the target countries in parentheses.

Integration of Big Data into the Business Curriculum: Evidence from Top US Undergraduate Business Programs

Rafiq H. Hijazi
College of Natural and Health Sciences
Zayed University
P. O. Box 144534, Abu Dhabi, UAE
Email: rafiq.hijazi@zu.ac.ae

ABSTRACT

The rapid emergence of "Big Data" has substantially changed the business process and showed strong potential to significantly improve the business decision-making practices. This necessitates the need to introduce big data in undergraduate business curriculum to ensure the preparedness of business graduates to the information-based labor market. Hence, the purpose of this study was to examine the current state of big data coverage in the top 50 US undergraduate business programs. Relevant data have been collected based on content analysis of the program curricula, syllabi of required statistics courses and adopted statistics textbooks. Results revealed that all business programs require at least one course of statistics or business analytics. However, there is a lack of big data presence in the majority of these programs. Moreover, only very few of the commonly used business statistics and business analytics textbooks have incorporated a brief introduction to big data concepts and practices. Finally, recommendations on efficient integration of big data into the undergraduate business curriculum are provided.

Keywords: big data, business curriculum, statistics and business analytics

Evolution of a Sustainable Innovation Ecosystem

Josephine Chong
Northumbria University
Josephine.chong@northumbria.ac.uk

ABSTRACT

Business disruptions are changing business strategies. Recently, organizations are moving towards in building a sustainable future to respond to business disruptions. Sustainability has been realized as an important strategic driver in many organizations for achieving competitive advantages. This study introduces the concept of ecosystem lifecycle to shed light on how innovation ecosystems overcome business disruptions which are dynamic and constantly changing.

KEYWORDS

Innovation ecosystems; Sustainability; Lifecycle; Data analytics

1. INTRODUCTION

Sustainability has been realized as an important strategic driver in many organizations for achieving competitive advantages. With the pressure from the market for more environmentally friendly products, organizations which adopt green processes and practices in manufacturing their products tend to reap more competitive advantages (El-Kassar et al. 2018; Lin et al 2016). Recently, organizations are moving towards in building a sustainable future to respond to business disruptions. The notion of future sustainability refers to using resources to meet the needs of the present without compromising the ability of future generations to meet their own needs (WCED 1987). Researchers have asserted that organizations can overcome future sustainability challenges through technological innovations (Kramer and Porter, 2011). As such, many organizations are leveraging technological innovations to slow down the depletion of natural resources and to reduce the negative impact on environment, thus enhancing sustainability.

In today's business world, no single organization is able to facilitate technological innovation on its own. Complementary collaboration with other relevant organizations is essential in creating valuable sustainable products for customers continually. Implicitly, an innovation ecosystem is established, which involves organizations working together to plan and execute operations toward environmental sustainability goals. Innovation ecosystem research has

proliferated over the last decade as it focuses on the joint participation of organizations towards innovation. All stakeholders co-operate within a respective innovation ecosystem with the goal of achieving collaborative advantage.

Huxham and Macdonald (1992, p.50) defined collaborative advantage as "the creation of synergy between organizations towards the achievement of common goals". As such, the value of the ecosystem is determined by the degree of interdependency among participated organizations and their ability to innovate the product successfully (Adner and Kapoor 2010).

There is not much research on innovation ecosystem evolution (Dedehayir et al 2018). Successful innovation ecosystem evolution is the basis for organizations to maintain sustainable advantages (Gao et al., 2019). This study introduces the concept of ecosystem lifecycle to shed light on how innovation ecosystems overcome business disruptions which are dynamic and constantly changing. Furthermore, this study seeks to understand the role of agility capability which is a critical mechanism in sensing and responding to the changes during the evolution of innovation ecosystem.

In this study, we analyze how VF Corporation, parent company of Wrangler, has succeeded in creating its sustainable innovation ecosystem. How did VF Corporation develop and harness agility capability to facilitate its sustainability innovations for Wrangler? We draw upon the concept of an ecosystem lifecycle, developed by Moore (1996), to provide a better understanding of how Wrangler evolved its ecosystem.

2. THEORETICAL BACKGROUND

2.1 Agility Capability

Agility capability is an important mechanism in overcoming uncertainties and turbulences in the business environment. More significantly, agility capability can exploit volatile external and internal changes to enhance and refine value creation, capture and competitive performance (Sambamurthy et al. 2003). There are many conceptualizations of agility capability: sensing and responding capacity to detect threats and opportunities (Liang et al., 2017; Sambamurthy et al., 2003; Tallon and Pinsonneault, 2011); collaborating capacity with customers and stakeholders to expand competitiveness (Ngai et al., 2011; Richardson et al., 2014); ability of organizational processes to exploit opportunities for market capitalization and respond to threats (Chen et al., 2014; Huang et al., 2014); and information capacity for knowledge exploration and exploitation (Nazir and Pinsonneault, 2012; Roberts and Grover, 2012). Although prior research has conceptualized different types of agility, the primary dimension of agility capability is measured as an ability to sense and respond swiftly to business disruptions to capture market opportunities. As such, we conceptualize "sense" and "response" as the two major components that comprise agility capability.

Innovation ecosystems need to possess sensing competency to address rapid environmental changes. More importantly, the innovation ecosystem needs to establish sensing tasks to assimilate and dissimilate critical information, and filter out irrelevant information about environmental uncertainties (Park et al., 2017). Technological

resources and information processing competency are important sensing mechanisms, which are enablers of agility capability (Huang et al., 2014; Roberts and Grover, 2012; Tallon et al., 2018). Nowadays, Internet of Things (IoTs) are crucial vehicles for driving interactions among the stakeholders in the ecosystem. IoTs are defined as "devices that have network connectivity and the ability to send or receive data and information to other connected objects" (Akhtar et al., p. 308). For instance, Airbnb and Uber rely on mobile devices, cloud computing and social media to transfer a huge amount of data to the customers who processed the data into information.

Besides developing a sensing competency to detect opportunities and threats, an innovation ecosystem has to leverage its resources to re-configure processes and create new processes, as well as to redesign its structure in order to cope with environmental dynamism. Nowadays, the ability to harness big data is vital which can provide more insights for how to discover changing market patterns ahead of competitors. More specifically, big data enables knowledge creation and knowledge update which can influence new product innovation, new service offerings and process improvement. Innovation ecosystems need to develop data analytics capability to leverage their analytic tools to extract meaningful information from the collected raw data. In the empirical study of Srinivasan and Swink (2018), supply chains are leveraging analytical techniques (e.g., simulation, optimization, regression) and visualization techniques (e.g., dashboards) to improve their operational decision-making in terms of demand visibility and supply visibility. As defined by Grant (1996, p. 377), capabilities are "the ability to perform repeatedly a productive task which relates either directly or indirectly to a firm's capacity for creating value through effecting the transformation of inputs into output." More specifically, data analytics capability is regarded as a special, higher-level type of resource that is evolved from distinct organizational structures, functional processes, routines, and skills (Teece at al., 1997). Therefore, we suggest that data analytics capability enables innovation systems to transform collected structured and unstructured data into insightful information in responding swiftly to environmental disruptions.

2.2 Innovation Ecosystem Lifecycle

According to Moore (1993), there are four phases in an innovation ecosystem lifecycle: birth, expansion, leadership and self-renewal. The birth phase emphasizes on devising a value proposition of a new product and service, also defining the best strategy in delivering it. The expansion phase emphasizes on scaling up the capacity of the ecosystem to expand into new territories. The leadership phase emphasizes in developing leadership to create a more value-added ecosystem by stabilizing the ecosystem's sub-systems and processes, thus achieving sustained competitive advantage. Importantly, the leaders need to establish a clear vision of future development, which in turn, steers all the stakeholders and customers to be more committed to the ecosystem. The self-renewal stage phase emphasizes on the ecosystem's competency in responding to new ecosystems; starting new ecosystems; and creating new ideas to balance stability and change.

An innovation ecosystem evolves when organizations (on the supply side of the market) collaborate interactively in leveraging the ecosystem's core resources and capabilities to innovate products. Example of stakeholders in the innovation ecosystems including producers; suppliers; governmental organizations; organizations; and financial and

research institutions. Sustainable innovation entails an interplay among supply chain stakeholders, sustainable innovation process, and the environment. Closely knitted interactions are crucial mechanisms for an effective facilitation of sustainable innovative activities. Furthermore, sustainable innovation ecosystems are characterized as dynamic as they must respond quickly to the changing environmental environment in a real time manner. Sustainable innovation process begins at the birth stage and ends at self-renewal stage of the business ecosystem lifecycle.

In the birth stage, the innovation ecosystem focuses on two business objectives (Moore, 1993). The first objective is to identify a seed innovation that can bring about revolutionary products. The second objective is to discover new opportunities that deliver value added products to customers. The innovation ecosystem must scan effectively for feasible explorative activities in order to achieve both objectives. Large-scale information processing is an essential mechanism to facilitate the innovation ecosystem to make timely decisions for defining creative innovative opportunities. For instance, Apple and Samsung possess superior information processing capabilities that place them in a better strategic position to adapt in fast changing environment. The smartphone ecosystem is are constantly collecting data about customer needs and competitor competencies. Subsequently, the collected data is coordinated and processed into insightful information which can enhance innovation capabilities. In 2018, Apple sent out survey to iMac Pro buyers to ask about their preferences regarding the new features. Apple used the collected feedback to help direct Mac Pro features so that the improvised Mac Pro can be released in 2019. Therefore, IoTs are important technological enablers for capturing customer data from all device-to-device connections.

In the expansion stage, the innovation ecosystem focuses on three business objectives (Moore, 1993). The primary objective is to capture a bigger market share value by competing against other ecosystems. The second objective is to stimulate more customer demand for its product offerings. The third objective is to meet demand with adequate supply. The innovation ecosystem needs to develop a sensing competency that enable the stakeholders to see opportunities sooner, think critically, and put learning into action. For an innovation ecosystem to expand into new territories, it must possess technological innovation knowledge and creative thinking competency. Inter-organizational learning within an innovation ecosystem can increase the speed to market for new products development. Through the learning process, stakeholders' assets, specialized competencies, knowledge are converged to discover a new technique or a new process to innovate a product. More importantly, the innovation ecosystem needs to establish vigilant learning which is "characterized by alertness, curiosity, and a willingness to act on partial information" (Day and Schoemaker, 2016, p. 62).

In the leadership stage, the primary focus of the innovation ecosystem is to enhance profitability through its constant innovation (Moore, 1993). In a sense, profitability is achieved by creating new products, services and processes. The innovation ecosystem has to develop a responding capacity to cope with customers' changing behavior.

The innovation ecosystems has to rely on big data analytical tools to understand customers' needs and preferences. Enhanced customer value is created when there is an improvement in the delivery of customer services. Such analytic applications can be leveraged in different domains (e.g., marketing, logistics, manufacturing and operations) within

the innovation ecosystem. Examples of big data analytical tools are digital dashboards, data warehouses, and web analytics. The analytical insights collected through the IoTs are processed into knowledge that is crucial for transforming innovation processes. For instance, Ford Motor Company deployed consumer analytics to gather data from four million consumers through sensors and remote app management software to revolutionize its own noise cancelling technology (Erevelles et al., 2016). By using car's voice recognition system to analyze the collected data, Ford was inspired to use noise cancelling headphones to minimize background noise. This product innovation provides a quieter driving environment that can enhance driver concentration.

In the self-renewal stage, the ecosystem further enhances agility capability in order to stay ahead of other innovation ecosystems by embracing and leading the change. At this stage, the innovation ecosystem must be extremely responsive to abrupt new environmental changes that include changes in government regulations, disaster events and macroeconomic conditions. In particular, the innovation ecosystem needs to re-configure its innovation processes to enhance positive and reduce negative environmental impacts. It is vital for the innovation ecosystem to develop data analytics capability to manage sustainable operations practices. To realize these practices , the ecosystem needs to have a sustainable innovative design that includes life cycle assessments, the built environment, and services configuration (Tseng et al., 2018).

Figure 1 illustrates the lifecycle of an innovation ecosystem. At the birth stage, the innovation ecosystem establishes information-processing capacity to capture innovation opportunities. At the expansion stage, stakeholders within the ecosystem adopt vigilant learning to overcome innovation challenges and identifying external threats. At the leadership stage, data analytics tools are deployed to respond to the fast-moving developments of innovative products and applications. At the self-renewal stage, data analytics capacity is developed to enhance the innovation ecosystem's responsiveness in overcoming emerging sustainability challenges.

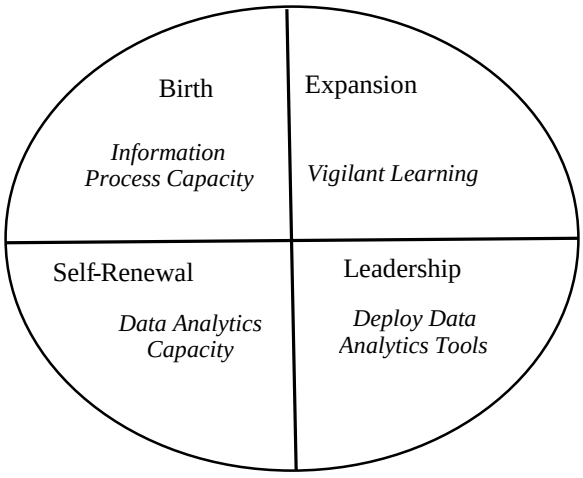


Fig. 1: Sustainable Innovation Ecosystem Lifecycle

3. THE CASE OF WRANGLER SUSTAINABLE INNOVATION ECOSYSTEM

Our primary source of data was the archival data available from the websites of VF Corporation. We obtained and analyzed numerous website documents and press releases to come to an initial understanding of its sustainability strategy. We then specifically investigated how VF Corporation developed agility capability in planning and designing its sustainable innovative process. Drawing on the archival data, we were able to gain a better understanding of the planning and decision-making processes of VF Corporation's sustainable operational practices.

3.1 Wrangler and Sustainable Cotton Farming

In 1899, VF Corporation is an international apparel and footwear company founded in October 1899 by John Barbey and his group of investors. Over the decades, the company has acquired companies that encompass Lee, Wrangler, Rustler, Jantzen, JanSport, Red Kap, Bulwark, The North Face, Eastpak, Nautica, Vans, Kipling, Eagle Creek, Timberland and Icebreaker. The acquired brands are a mixture of workwear, sportwear, jeanswear and outdoor accessories which are further categorized into Outdoor, Active, Work and Jeans. VF Corporation is regarded as a sustainable organization with a blend of for-profit coupled with environmental objectives. In 2017, the organization announced its global sustainability and responsibility strategy – Made for Change. The primary objective of the strategy is to reduce environmental and social footprint by exploring how products are manufactured and sold, and how the materials are sourced.

Among the brands owned by VF Corporation, Wrangler is the most active subsidiary organization in undertaking many environmental and social initiatives to cope with sustainability. It is moving away from the conventional linear system of production (take, make, waste) which is more costly and produces more waste. More strategically, the sustainable organization is adopting a circular supply chain that seeks to nurture the future generation of farming by exploring and exploiting innovative means to improve sustainability practices. The circular production process enables materials to be reused again and again, and old products can be recycled or are turned into new ones, generating very little waste. Wrangler strives to deliver sustainable innovation products to their consumers by engaging in active collaboration with its ecosystem of stakeholders. By partnering with farmers and organizations, Wrangler is able to embrace technology to reduce, reuse and recycle materials and water consumption, as well as to enhance soil health and dying techniques.

3.2 Wrangler Sustainable Innovation Ecosystem: Birth Stage

With the constant and fast changing apparel retail market, it is vital for Wrangler to turn its attention to producing sustainable products. VF Corporation took two years during the birth stage of the innovation ecosystem to assimilate information from stakeholder engagement, trends analyses and consumer research to manage sustainability. Importantly, information processing capacity was established to seek an understanding what were the parameters for its sustainability strategy. Sensing competency is an essential mechanism in scanning and filtering critical information

for identifying strategic solutions and sustainable means that can reduce VF Corporation's overall environmental impact by half.

The first sensing task gathered data from the customers to gain more insights to understand their expectations from purchasing Wrangler produced apparels. After conducting a consumer research, VF Corporation was informed of the current trends about customers' behavior: they are conscious spenders who only purchase products that have a positive impact on society. By sensing customers' changing social and environmental values, VF Corporation transited from linear to circular supply chain that recycles and reuses valuable limited resources in the manufacturing process to create more sustainable products. Furthermore, consumers can experience better shopping experiences where they can either engage in re-commerce activities (buying and selling used apparels) or rent apparels they need, rather than buying new apparels. This changing shopping behavior can reduce environmental footprint both for VF Corporation and the customers.

The second sensing task was to conduct comprehensive environmental risk assessments to study how VF Corporation's operational processes had an impact on the environment. Furthermore, VF Corporation performed a sustainability audit to ensure VF Corporation had adopted best environmental practices by conducting a materiality assessment in 2017. The materiality assessment report highlighted several main priorities: climate actions; ethical labor practices and workplace health and safety; energy reduction and energy efficiency; waste generated in manufacturing; water use in textile and manufacturing; and ensuring consumer safety of products and materials. Consequently, VF Corporation shaped and developed its sustainability strategy by addressing these sustainability and social issues.

3.3 Wrangler Sustainable Innovation Ecosystem: Expansion Stage

Wrangler innovation ecosystem developed vigilant learning to gain a better understanding of how to improve soil health and sustainability. Wrangler collaborates with external stakeholders (cotton producers; Soil Health Institute; and the USDA's National Resources Conservation Service) to learn about farming best practices aimed at improving soil health (Seeding soils potential). Wrangler was expanding its innovation ecosystem by implementing techniques including conservation tillage, cover crops and conservation crop rotations to supplement organic matter and microbial organisms in the ground, and reduce associated carbon emissions. Consequently, such newly learnt soil health practices yielded more profits for cotton producers. Furthermore, they enhanced environmentally quality by preserving natural habitats and biodiversity. The vigilant learning that took place at the expansion stage enabled Wrangler in exploring new ways to improve soil health, create innovative circular products and keep up with customers' preferences and needs. Consequently, Wrangler gained two percent more of the market share as its innovation ecosystem expanded.

3.4 Wrangler Sustainable Innovation Ecosystem: Leadership Stage

In order to lead the transformation towards sustainability through innovation, VF Corporation invested heavily in data analytics tools to overcome challenges such as climate change and limited raw materials. Wrangler depended heavily on data analytical tools to collect the sustainability data in managing cotton production. The innovation ecosystem is leveraging on smart farming technologies to capture farming data to enhance planting precision, minimize seed and fertilizer input and monitor soil health on a field-by-field basis. Through Global Positioning System (GPS) interfaced to tractors, real-time kinetics field-specific data can be collected for easier farm monitoring and management. Crop management decisions can be analyzed with digital field mapping connected to GPS location.

Furthermore, Wrangler's cotton pickers can deploy a unique identifier using barcode and/or Radio Frequency Identification technology (RFID) to record crop yield and moisture, mark bales of cotton and co-ordinate their locations, subsequently move from one acre to the next.

Wrangler takes on an active role in shaping future directions for the blue jean industry by adopting a data-intensive approach to sustainability and farm management. The innovation ecosystem has formed a sustainable cotton coalition with governmental organizations (i.e., Texas Alliance for Water Conversation (TAWC) and The Nature Conservancy), Field to Market and North Carolina State University to provide collaborative training and improvement programs for cotton producers across the country. The coalition aims to enhance environmental and social practices in the global cotton industry by embedding more sustainable farming practices.

3.5 Wrangler Sustainable Innovation Ecosystem: Self-Renewal Stage

Entering the self-renewal stage, Wrangler innovation ecosystem aimed to transform the jeans wear industry towards the next generation pf sustainable fashion. The innovation ecosystem is leveraging on consumer analytics to better inform of customers' changing preferences in product design and inventory transparency throughout the supply chain. Wrangler collects a vast amount of customer data from its brick and mortar and online channels. By leveraging on machine learning, a predictive analytics capability, to scan 210,000 metrics daily, this enables Wrangler to gain a better understating of customers' behavior. Examples of the metrics are data from the e-commerce point of sale, social media, page load time, customer views, media spends and many more.

4. CONCLUSION

This study focuses on innovation ecosystem evolution. Using a case study, this article provides an understanding how an innovation ecosystem has evolved to overcome business disruptions. There are two main contributions. First, it is one of the few studies that study the evolution of an innovation ecosystem. Second, the study provides an understanding on how an innovation ecosystem needs to develop agility capability to sense opportunities and respond to threats. The main limitation is that this exploratory study has certain representatives and interpretability. Future research can examine innovation ecosystem evolution with in depth case studies from exemplary organizations, comparing different industries in different markets. In addition, future research can consider deploying quantitative analysis to further confirm the research model.

REFERENCES

Adner, Ron and Kapoor, Rahul (2010) "Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations". Strategic Management Journal, 31(3), pp. 306–333.

Akhtar, Pervaiz, Khan, Zaheer, Tarba, Shlomo and Jayawickrama, Uchitha (2018) "The internet of things, dynamic data and information processing capabilities, and operational agility". Technological Forecasting and Social Change, 136, pp. 307–316.

Chen, Yang, Wang, Yi, Nevo, Saggi, Jin, Jiafei, et al. (2014) "IT capability and organizational performance: The roles of business process agility and environmental factors". European Journal of Information Systems, 23(3), pp. 326–342.

Day, G. S., & Schoemaker, P. J. (2016). Adapting to fast-changing markets and technologies. California Management Review, 58(4), 59-77.
Dedehayir, Ozgur, Mäkinen, Saku J. and Roland Ortt, J. (2018) "Roles during innovation ecosystem genesis: A literature review". Technological Forecasting and Social Change, 136, pp. 18–29.

El-Kassar, A. N., & Singh, S. K. (2018). Green innovation and organizational performance: the influence of big data and the moderating role of management commitment and HR practices. Technological Forecasting and Social Change.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897-904.

Gao, Y., Liu, X., & Ma, X. (2019). How do firms meet the challenge of technological change by redesigning innovation ecosystem? A case study of IBM. *International Journal of Technology Management*, *80*(3-4), 241-265.

Grant, Robert M. (1996) "Prospering in dynamically-competitive environments: Organizational capability as knowledge integration". Organization Science, 7(4), pp. 375–387.

Huang, Pei-Ying, Pan, Shan L and Ouyang, Tao Hua (2014) "Developing information processing capability for operational agility: implications from a Chinese manufacturer". European Journal of Information Systems, 23(4), pp. 462–480.

Huxham, Chris and Macdonald, David (1992) "Introducing collaborative advantage: Achieving inter- organizational effectiveness through meta- strategy". Management Decision, 30(3). [online] Available from: https://www.emeraldinsight.com/doi/abs/10.1108/00251749210013104 (Accessed 22 June 2019)

Kramer, M. R., & Porter, M. (2011). Creating shared value. Harvard business review, 89(1/2), 62-77.

Liang, Huigang, Wang, Nianxin, Xue, Yajiong and Ge, Shilun (2017) "Unraveling the alignment paradox: how does business - IT alignment shape organizational agility?" Information Systems Research, 28(4), pp. 863–879.

Lin, Y. H., & Tseng, M. L. (2016). Assessing the competitive priorities within sustainable supply chain management under uncertainty. Journal of cleaner production, 112, 2133-2144.

Moore, James F. (1993) "Predators and prey: A new ecology of competition". Harvard business review, 71(3), pp. 75–86. Nazir, Salman and Pinsonneault, Alain (2012) "IT and firm agility: An electronic integration perspective". Journal of the Association for Information Systems, 13(3). [online] Available from: https://aisel.aisnet.org/jais/vol13/iss3/2

Ngai, Eric W. T., Chau, Dorothy C. K. and Chan, T. L. A. (2011) "Information technology, operational, and management competencies for supply chain agility: Findings from case studies". The Journal of Strategic Information Systems, 20(3), pp. 232–249.

Richardson, Sandra, Kettinger, William, Banks, Michael and Quintana, Yuri (2014) "IT and agility in the social enterprise: A case study of St Jude children"s esearch hospital"s "Cure4Kids" IT-platform for international outreach". Journal of the Association for Information Systems, 15(1). [online] Available from: https://aisel.aisnet.org/jais/vol15/iss1/2

Roberts, Nicholas and Grover, Varun (2012) "Leveraging information technology infrastructure to facilitate a firm"s customer agility and competitive activity: An empirical investigation". Journal of Management Information Systems, 28(4), pp. 231–270.

Sambamurthy, Vallabh, Bharadwaj, Anandhi and Grover, Varun (2003) "Shaping agility through digital options: Reconceptualizing the role of information technology in contemporary firms". MIS Quarterly, pp. 237–263.

Srinivasan, Ravi and Swink, Morgan (2018) "An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective". *Production and Operations Management*, 27(10), pp. 1849–1867.

Tallon, Paul P. and Pinsonneault, Alain (2011) "Competing perspectives on the link between strategic information technology alignment and organizational agility: insights from a mediation model". MIS Quarterly, 35(2), pp. 463–486.

Teece, David J., Pisano, Gary and Shuen, Amy (1997) "Dynamic capabilities and strategic management". Strategic Management Journal, 18(7), pp. 509–533.

Tseng, M. L., Lim, M. K., Wong, W. P., Chen, Y. C., & Zhan, Y. (2018). A framework for evaluating the performance of sustainable service supply chain management under uncertainty. International Journal of Production Economics, 195, 359-372.

VF Corporation, (2017) "Made for change: VF"s new sustainability & responsibility strategy". [online] Available from: https://www.vfc.com/news/company-news/detail/54204/made-for-change-vfs-new-sustainabilityresponsibility (Accessed 24 June 2019)

WCED (World Commission on Environment and Development) (1987), Our Common Future, Oxford University Press, Oxford, UK.

Approaches to Using Big Data for Business Analysis

Peter Dahlin
Mälardalen University
Email: peter.dahlin@mdh.se

ABSTRACT

A large variety of datasets are joined under the term 'Big Data'. They differ regarding volume, velocity and variety, but also regarding the context they are in, and the phenomenon they relate to.
Generating or collecting data poses some challenges; making use of it implies others. This paper outlines a framework of approaches to using big datasets in business analysis, with examples from projects studying companies and managers. Exemplified datasets contain, for example, standardized data on company finances, formal relationships such as board appointments, and unstructured text describing events.
The suggested framework consists of two dimensions. The first concerns how central the dataset is to the study and analysis; ranging from being the only data used, to being a minor complement to other data. The other dimension regards the refinement of the data. At one extreme, the data are directly analyzed, while at the other, data needs to be calculated or created to represent the sought variables. The paper uses four example studies to illustrate the four approaches.
In the approaches requiring much refinement, calculations and data generation, the analysis layer is very central. Whereas this aspect offers great opportunities, it also complicates the understanding of what the underlying dataset is. In contrast to quantitative survey research, the dataset is not ready for statistical analysis. Instead, it should be considered an "empirical model"; a representation of the studied phenomenon. Similar to asking questions to respondents, questions can be asked to such datasets. Calculated or generated variables then form the data in the analysis layer, which can be used to answer the research question.
In summary, Big Data approaches can take many different forms, somewhat differing from "standard" research methods. The framework could aid in planning as well as explaining how big datasets can be used in business analysis.

Big Data Analytic for Sustaining the Growth of Dubai Hospitality Knowledge Economy

Farhi Marir

College of Technological Innovation, Zayed University, Dubai, UAE Farhi.marir@zu.ac.ae

ABSTRACT

The Hospitality (aka Travel and Tourism) industry has a high impact on Dubai's knowledge economy with direct contribution of 20% to its GDP and 6% to its total employment. Data analytical tools are the response to the 21st century's challenges on the hospitality industry. Mining such a large volume of data (aka as Big Data) and online booking systems, social networks, Web and mobile services to establish the profile of people visiting Dubai and understand their needs and requirements is almost non-existent. This research paper capitalizes on the big data magnet to develop Big Data Analytic Hospitality (BDAH) tools that capture and mine big data from authentic and reliable sources to support Dubai hospitality managers to instantly respond to customers' needs and meet their expectations, adapt their businesses and develop products and services capable of bringing back old and attracting new customers.

KEYWORDS

Big data analytics; Hospitality knowledge economy; Social networks; Web and mobile services.

1. INTRODUCTION

The Hospitality (aka Travel and Tourism) industry has a high impact on Dubai's knowledge economy with direct contribution of 20% to its GDP and 6% to its total employment. The endless stream of brand new ``educated'' consumers booking online is receding - it is no longer enough to have a Web site with a booking engine, the online experience must be such that it attracts customers and keeps them coming back. Over 5 billion people are calling, tweeting, texting and browsing from mobile phones 90% of global data was created in the last two years (http://wikibon.org/blog/big-data-statistics/). Data analytical tools are the response to the 21st century's challenges on the hospitality industry. Mining such a large volume of data (aka as Big Data) and online booking systems like Booking.com and Velocity.com, social networks, Web and mobile services to establish the profile of people visiting Dubai and understand their needs and requirements is almost non-existent.

This research paper capitalizes on the big data magnet to develop analytical tools that capture and mine big data from authentic and reliable sources on tourist's feedback to support Dubai hospitality managers to instantly respond to customers' needs and meet their expectations, adapt their businesses and develop products and services capable of bringing back old and attracting new tourists.

The paper is organized as follows. Section 2 is devoted to report literature review, section 3 will detail the proposed Big Data Analytic Hospitality (BDAH) analytical tool and the last section presents the conclusion on the developed BDAH tool.

2. LITERATURE REVIEW

Over the years, different development waves have shaped the Web. Started as a simple browsing tool to screen Web sites, the Web now is a dynamic and robust platform upon which companies conduct business and initiate cross-border collaborative activities (Lijun Duan and Hao Tian, 2017). The latest development wave in the Web triggered by among other things the pressure on organizations to remain agile and Web2.0 widespread adoption sheds the light on two major research streams (Punita Duhan & Anurag Singh, 2014)

- Research on loosely-coupled Web applications. The execution of these applications spans several distributed and heterogeneous data sources and hence, has to cross organization boundaries transparently.
- Research on social computing illustrated with the massive deployment of social applications like Facebook, Twitter, and LinkedIn. These applications capitalize on the ability and willingness of users to interact, share, collaborate, and recommend.

Such research streams have a common element that is the Web. Many organizations are getting the message about Web2.0: "*enterprise spending on Web2.0 technologies will grow strongly over the next five years, reaching $4.6billion globally by 2013, with social networking, Mashups, and RSS capturing the greatest share"*. The social organization (aka Organization2.0) is expected to maintain contacts with customers, suppliers, and competitors (Linda Jane Coleman ET al.,2019). To this end, the existing practices for managing business processes need to be re-visited in a way that permits to capture social relations that arise inside and outside the organization, establish guidelines and techniques to assist IT practitioners integrate social relations into their design, development, and maintenance efforts, identify and tackle challenges that prevent capturing social relations, and last but not least define techniques to assess the impact of social relations on the organization performance.

This proposal considers the readiness of hospitality companies in Dubai (that could be generalized to the whole UAE) to respond to the challenges associated with capturing and analyzing Big Data for the sake of offering high-quality online services. It aims at making these companies rethink the way they conduct business. Big Data analytics offer new opportunities that will enable them to enhance their performance when it comes to gauging the opinions of customers, dealing with good suppliers, and being aware of competitors. Being able to implement big data analytics, Dubai hospitality and travel companies would be able to gain important insight through external market intelligence generated by the end users, prevent potential risks by detecting negative ratings so they can act swiftly, compete and develop successful products and services based on real data, develop superior customer experience by understanding what customers values most (who & when) and finally and manage opinions & gain feedback through sentiment analysis to generate immediate feedback and develop new insight.

3. THE DEVELOPED BDAH ANALYTICAL Tool

For effective and efficient monitoring and harvesting of web and social based data and for the maximum flexibility we propose to build Media & Content Aggregation Layer using our Web Intelligence Data Aggregation & Enrichment framework (WIDA framework) as shown in Fig.1 below. This architecture has proved to be especially efficient for harvesting the social sites, discussion groups, article comments and various types of individually generated human information on external sites. The key benefits of such web scripting framework are: ability to easily and quickly add new web sites to be harvested, Ability to work with different level of granularity of harvested content and associated HTML metadata or content attributes, ability to detect the changes of web source structures and formats for management & reconfiguration of harvesting scripts, easy and efficient deployment, all scripts produce common IDX format for direct indexing, ability to further enrich the standard IDX text format before indexing harvested data to

Autonomy IDOL and modular design with ability to easily extend and customize the harvesting logic

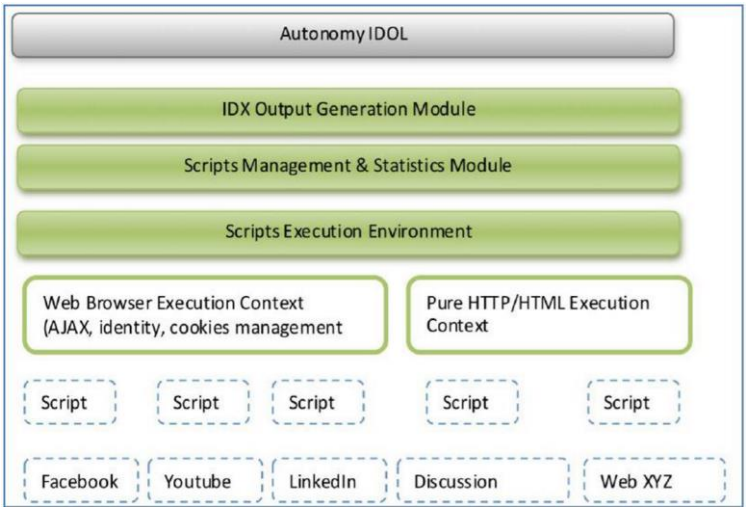


Fig. 1: An overview of the system architecture

WIDA framework is content extraction framework that has proved to be especially efficient for harvesting the social sites and discussion groups with very detail level of granularity and ability to bypass or overcome many usual technical constraints. WIDA consist of set of predefined common scripting functionality and best practices of how to effectively harvest social media sites and external web sites at the very detail level of data granularity. The biggest advantage is its flexibility of harvesting and monitoring virtually every external site. Even sites that have implemented various restrictions like special URL based encrypted / cookie based session management or AJAX based dynamic content etc.

3.1. Capturing and formatting hospitality Data:

Three main investigation that have been launched are: deploying a survey to local hotels, scrawling hospitality web sites and developing a data model for capturing and formatting collected data from social networks, web sites and devoted hospitality web site tools.

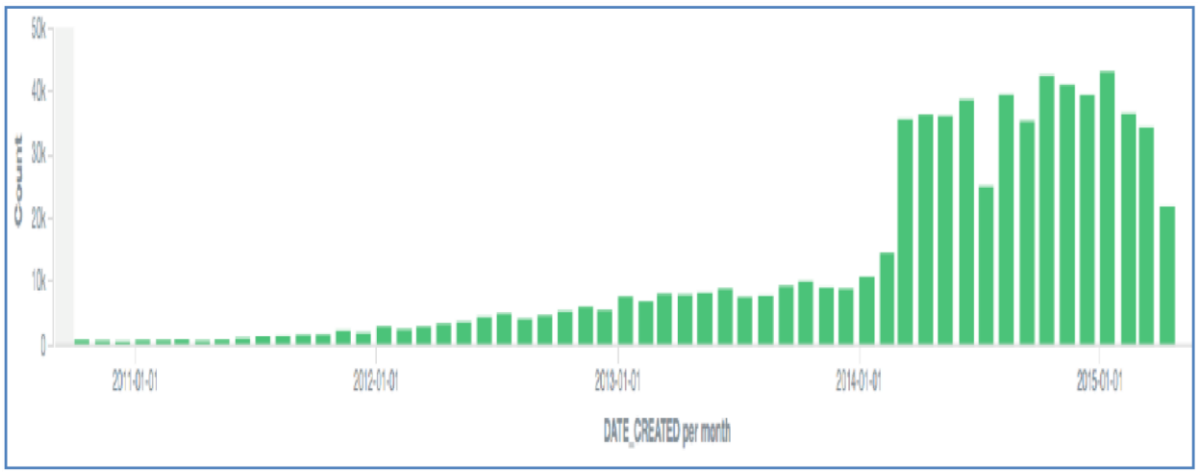


Fig. 2: Distribution of the data crawled during the project

3.2. Data Model for Capturing Social Data

We developed a data model to be used as a base for capturing and analyzing hospitality data. In this model, several parameters or variable have been defined for capturing and contextualizing customers' comments and feedbacks. This includes language, the source, tourist location, negative or positive comments, hotel, date, scoring of services, etc. Fig.3 below show amount of comments from visitors countries.

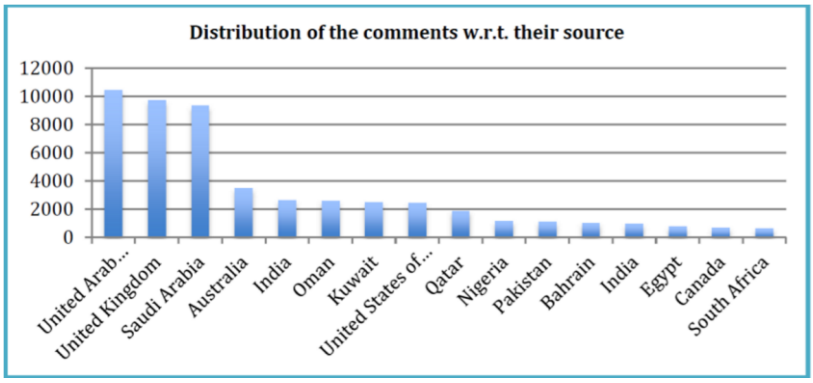


Fig. 3: Distribution of comments amongst countries from which they originated

3.3. The BDAH Analytical Tool

Social Media Management Software is generally understood as a set of tools designed to manage or analyze interactions through multiple social media accounts from a single dashboard. Most systems permit listening for brand mentions, posting to multiple channels, responding to inquiries, and running marketing campaigns. They include analytics packages to measure the relative success of campaigns. Within this broad definition, there are distinct use cases which emphasize different feature sets. Many social media platforms offer a broad feature set encompassing multiple or these use cases, while some excel in one particular area. Many companies use more than one tool to manage their social media efforts. This includes social media monitoring, media publishing, engagement, and marketing. In this project we are we developed an initial tool to use social media as source for hospitality data analytics to measure and optimize its impact on hospitality in Dubai. For this we used the following framework and applications to develop the analytical tool: xdSuite Framework to develop the IDOL based User Interface (can be migrated to ElasticSearch), SearchMiner Application to deliver IDOL based reporting and ElasticSearch and WIDA Framework for data harvesting (HP Autonomy, Attivio,

ElasticSearch)

The key BDAH tool capabilities includes modularity by design of key functional and visualization elements like charts, categories, navigation, ideas cloud etc., drag & drop support, Multilanguage support both from user interface and data processing perspective (ENG, CZ), integrated translation services, trends visualization (Clusters, 2D Maps, Spectrographs), flexible Dashboard definition for user roles, easy documents administration (rating, tagging, deleting), flexible Highlighting capabilities (keyword, names, date, custom defined entities, etc.), built-in HP IDOL Search capabilities (parametric, keyword search, time and relevance restriction, conceptual, field based etc.), support for dynamic drill downs, and data exporting (xlsx, docx, txt; template definitions). The developed BDAH tool can leverage the advanced Audio & Video Analytical Capabilities from HP Autonomy as shown in Fig. 4 below

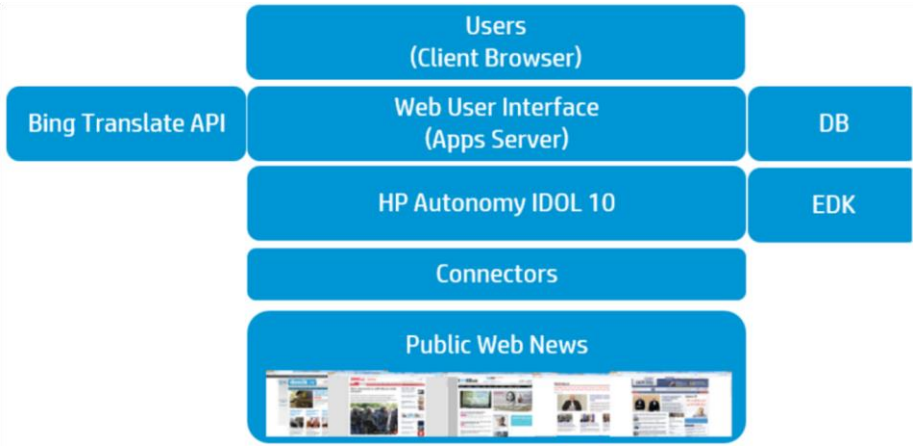


Fig. 4: Audio & Video Analytical Capabilities of BDAH tool

3.4. The BDAH Visualization Dashboard

The BDAH visualization tool was developed as a dashboard which will present the analytical results in a way that support managers in drawing conclusions and developing new products and services that have high factor of success. Two main tools to support hospitality mangers in their decision: (i) Online Analytical Processing tool (OLAP) which will support managers in analysis and understanding the impact of the customers comments and (ii) A Sentiment Analysis Managers' Dashboards (SAMD) predictive analytical tools to support hospitality managers in analyzing their customers' reviews on all aspects so they can identify areas to improves.

3.4.1. Online Analytical Processing (OLAP) Tool

The developed OLAP Tool is a descriptive analytical which supports managers in analysis and understanding the impact of their customers feedback. The OLAP is a powerful for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive "what if" scenario (budgeting, forecast customer numbers) planning. Below in Fig. 5 and Fig. 6 are some queries run by for hotels' managers on the OLAP Tool.

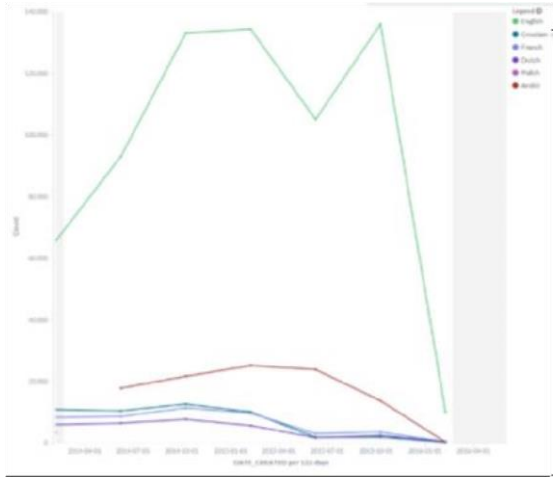


Fig. 5: What are the seven most attractive touristic sites in Dubai? and **Fig. 6:** Seasons per world regions/country visitors to Dubai?

3.4.2. Sentimental Analysis Managers' Dashboard (SAMD) Tool

The Sentiment Analysis Managers' Dashboards (SAMD) tool is used to analyze customers' reviews and take informed decision to improve their hospitality business. Fig. 7 below shows an instance of the dashboard along the legend of its functionalities

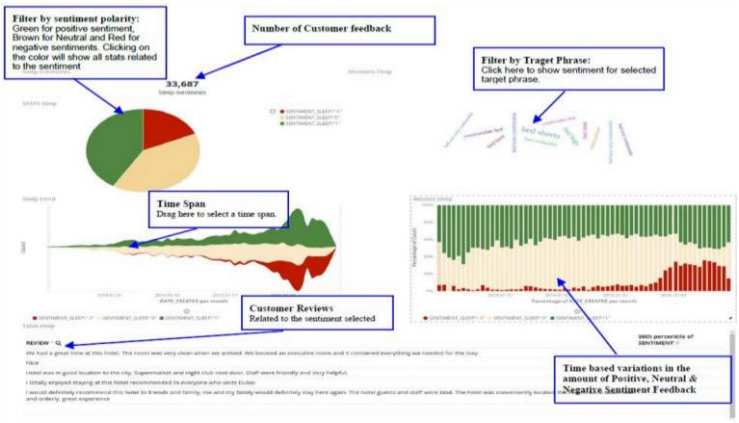


Fig. 7: SAMD Dashboard Legend.

The presentation below is a show case on the use of SAMD tool to analyze genuine customers' reviews during the years 2015 and 2016 based on eight selected aspects: Facilities, Food, Hotel, Room, Location, Sleep, WIFI and Staff. Fig. 8 below are the list of five hotels which received the biggest number of customers' reviews for each of the above aspects.

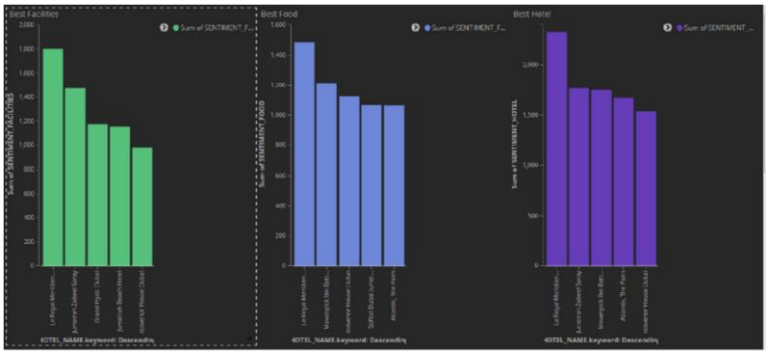


Fig. 8: The best five hotels on facilities, food and hotels aspects.

4. CONCLUSION

In today's knowledge economy, a company's ability to sustain its growth and competitiveness depends on how well it uses intelligently the social Web to efficiently manage its communications with stakeholders including customers, suppliers, and competitors. The DAHL analytical tool has been used by some hotels' managers in Dubai who recognized that such analytical especially its dashboard helped them understand their customers and take instant and long-term decision which showed improvements in the customers feedback. This improvement in customers views will sustain the tourism industry in Dubai especially when mining online booking systems like Booking.com and Velocity.com, social networks, Web and mobile services to establish the profile of people visiting Dubai and understand their needs and requirements is almost inexistent.

ACKNOWLEDGEMENTS

The author acknowledges the contribution of Dr. Zakaria Maamar to the research proposal sponsored by the UAE National Research Funding.

REFERENCES

Ali Thomas, G., Marir, F., Romas, M., & Prreti, P. (2013). A Framework for Designing Knowledge Management Systems- Aggregating the Existing Approaches. *The Fifth International Conference on Information, Process, and Knowledge Management*.

Malleda W., Marir, F., & Vassilev, V. (2013). Algorithms for mapping RDB Schema to RDF for Facilitating Access to Deep Web. The *First International Conference on Building and Exploring Web Based Environments; WEB 2013, 32-41*.

Marir, F. (2014). Mining social networks and patient stories to discover new knowledge about diabetes, *16th International Conference on Data Warehousing and Knowledge Discovery - DaWaK 2014.*

Marir, F. (2013). eRaUI: An adaptive interface for e-Research tools. *The First International Conference on Building and Exploring Web Based Environments, WEB 2013.*

Marir, F. & Ndata, J. (2013). Knowledge Enhanced Framework for designing e-Workflow Systems. *The Fifth International Conference on Advances in Databases, Knowledge, and Data Applications, 143-149.*

Lijun Duan and Hao Tian (2017) Collaborative Web Service Discovery and Recommendation Based on Social Link, *Future Internet 2017, 9, 63 (www.mdpi.com/journal/futureinternet)*

Punita Duhan & Anurag Singh (2014) Enterprise 2.0: a boon or bane for entrepreneurial and innovative expenditures? *Journal of Innovation and Entrepreneurship volume 3, Article number: 15.*

Linda Jane Coleman, Anurag Jain, Nisreen Bahnan, and Douglas Chene (2019) Marketing the Performing Arts: Efficacy of Web 2.0 Social Networks, *Journal of Marketing Development and Competitiveness, Vol 13 No 3.*

Proceedings of the **2019 International Conference on Big Data in Business**
October 29th to October 31st, 2019

Big data, complexity and financial policy

Charilaos Mertzanis
Abu Dhabi University
charilaos.mertzanis@adu.ac.ae
and
Haitham Nobanee
Abu Dhabi University
haitham.nobanee@adu.ac.ae

Extensive abstract of a paper submitted for consideration for presentation in the conference: ***Big Data in Business (ICBIB 2019)***
October 29-31, 2019
London, United Kingdom

Information is a key element of financial decision-making. However, the link between information and financial activity is complex. Hamilton et al. (2007) note that, due to the rapid financial innovation, deregulation and global capital market integration, financial intermediation and market activity have realized tremendous growth and structural change throughout the globe. These developments have profound implications for the riskiness, management and performance of financial institutions and the global financial system.

The speed of change and its impact on financial market activity has been largely the result of advancements in information and data management technology. The latter has enabled the automation and computerization of work processes and business functions, as well as the generation and rapid processing of large volumes of data that have in turn driven innovation in new financial products and strategies. The globalization and integration of financial markets has accelerated changes in financial data infrastructure, leading to a sharp increase in global interconnectedness of financial markets and institutions that mediate financial activity.

Further, following the recent financial crisis, the accurate calculation of system-wide risk has become a core policy concern. Systemic risk modelling uses financial market data. The latter are relatively easy to collect, are public, and are quite objective. However, market data may not reflect the true fundamentals of the underlying financial institutions, and may lead to biased estimates of the probability of failure. This bias may be stronger when referring to the probability of network failures. Idier et al. (2013) show that market data models are not much reliable in predicting systemic risk. Fantazzini and Maggi (2013) similarly show that market data models may be good in very short-term predictions, but not in medium and long-term ones. Indeed, market prices are formed through complex interaction mechanisms that often reflect speculative behavior rather than the fundamentals of the companies to which they refer. Market models and financial network models based on market data may therefore reflect spurious components that could bias systemic risk estimation. This weakness of the market suggests the need to enrich financial market data with data coming from other, complementary, sources. Evaluation criteria for the riskiness of financial institutions include not only market prices but also credit ratings, reports of qualified financial analysts and opinions of influential media, among others. Thus, data capturing diverse signals almost in real time, offer the opportunity to extract new useful evaluation information that can complement market prices and substitute market information when not available (e.g. unlisted financial institutions, shadow banking, etc.).

These developments have revealed the inadequacy of traditional risk identification methods and of functional supervision by sector or institution as well as raised the importance of the effective monitoring and

regulating of the stability of the financial system as a whole. Policy analysis has accordingly moved to a new framework that focuses on risks to the system as a whole and tends to analyze the financial system as a complex adaptive system. Arthur (1995, 1999) argues that the concept and analysis of complex adaptive systems is increasingly used to study complex phenomena in many research disciplines. Farmer et al. (2012) note that the analysis of complex systems uses techniques, which are different from those used in conventional economic theory, such as optimization under constraint. The new techniques include network analysis, agent-based modelling,

non_linear dynamics, catastrophe theory and the theory of critical phenomena, as well as data mining. In particular, agent-based modelling conceives of the financial system as made up of interacting individual agents (people, firms, regulators, governments), each with the capacity to act with purpose and intent, each of which is acting in the context of networks in which the fundamental behavior of the agent is not fixed, but evolves in response to the behavior of others.

Miller and Page (2007) note that the use of computational models as a primary means for exploring the financial system complexity is important for various reasons. First, such models embrace systems characterized by dynamics, heterogeneity, and interacting components and are therefore suited for complexity analysis. Second, they require clarifications about how they can be used to study socioeconomic outcomes and therefore need to be further advanced. Third, given existing trends in the speed and ease of use of computation and the emergence of big data, computation models will become a predominant means by which to explore the world. Anand et al. (2012) argue that the challenge financial regulators are facing today in adopting computational models and frameworks to improve the understanding of risks to the financial system as a whole, is that traditional financial analysis has typically been based on an individual institution, sector, geographical location or some other partial level. That has led to the development of regulatory definitions, tools, and approaches, which are best in analyzing outcomes when studying at that partial level but not at the financial system as a whole. As instruments become increasingly complex and institutions and markets become increasingly interconnected, this segmented approach cannot provide the concepts and tools for supporting analysis of financial markets as an integrated complex network system. Haldane (2009) notes that risk measurement in financial institutions and systems has been idiosyncratic. Risks have been evaluated only by type and by institution. However, in a network, this individual node-focused approach gives little sense of risks across institutions and much less to the overall system comprising many networks, which are also expanding. Mertzanis (2013, 2014a) notes that risk management practice today has to operate within a complex financial environment, that includes complex instruments, processes, institutions and systems. Following financial innovation and changing demand patterns, financial instruments have evolved to include complex and opaque products traded in less transparent market environments. Financial behavior patterns are inadequately approximated by normal distributions thus restricting predictability. The links and interactions among financial institutions and between them and the financial system as a whole are complex and instrumental in determining the system's behavior that feeds back on the performance of institutions. The elucidation of complex instrument valuation, information generation dynamic processes, and system-wide interactions among market actors globally could allow financial institutions to respond more efficiently to events that not only affect their own financial positions but also cause feedbacks between market actions and asset valuation thus affecting performance. Complexity analysis can address these issues and inform risk management practice. In this respect, the quality of information, the integration of diverse and complementary signals and the management efficiency of all that data and information are key challenges in this process. Lin et al. (2018) provide some relevant empirical evidence of different measures of systemic risk under complexity conditions.

The purpose of the paper is to explore how complexity analysis can be used to inform the management of financial risks in modern complex financial environments characterized by system-wide interconnections that give rise to big data. Complexity characterizes financial instruments, financial processes and financial systems. Each type of complexity is examined and the implications for financial risk management are explored. Then, the paper draws the implications of financial complexity for big data management and introduces the global legal identifier initiative designed to deal with granularity and comparability of financial big data globally. The paper aims at highlighting the link between complexity in finance and big data challenges and provide a background for understanding the legal entity identifier initiative used to manage the association. While the merits and application of complexity theory to finance are still a matter of debate, complexity analysis could contribute to the formation of a coherent body of propositions that are capable of better approximating reality in financial systems, i.e. explain the stylized facts in finance and manage financial system risk. In this respect, the proper treatment of big data is essential to this endeavor.

Machine Learning to Predict Accounting Restatements

KIM TROTTIER[7] and VINCENT CHIU

Beedie School of Business

Simon Fraser University

INTRODUCTION

Annual report restatements are costly in terms of the direct cost of audit, regulatory, and corporate effort as well as the indirect cost of reduced investor confidence in the capital markets. It would be useful to be able to identify some red flags that could serve as early indicators of misstated financial information, so as to reduce the negative effect of restatements. Prior research has drawn on economic theory to identify characteristics of misstating firms, and used this information to formulate predictive models meant to identify misstating firms. We expand this effort by exploring whether Machine Learning could be used to improve upon these predictive models.

Machine Learning (ML) is a field of artificial intelligence where algorithms "learn" about the relationship among variables in large data sets through analysis and statistics. Within the ML alternatives our method of choice is Random Forest (RF) because these are more powerful than standard methods such as ordinary least squares regression, and interpretation is easier than other ML techniques such as neural networks. As with any ML, there is some mystery as to how the Random Forest (RF) reaches its conclusion, leaving us unable to discern which financial reporting characteristics load as significant variable, or their relative magnitude. In addition, each RF trial produces a unique outcome that is not replicable and lacks econometric inference on the probability that result would be consistently obtained over many trials. We mitigate this concern by running our RF over multiple trials to ensure stability.

From an academic perspective, the main downside to ML is that it lacks theory to support and guide the analysis. The main benefit is that ML can analyze structures and relationships that are far more complex and numerous than can be captured with standard methods. An ML approach flips the traditional research paradigm on its head since the data informs the theory rather than the other way around. In our opinion, research should be inclusive and iterative. We believe research questions should be examined

[7] Contact: kim.trottier@sfu.ca

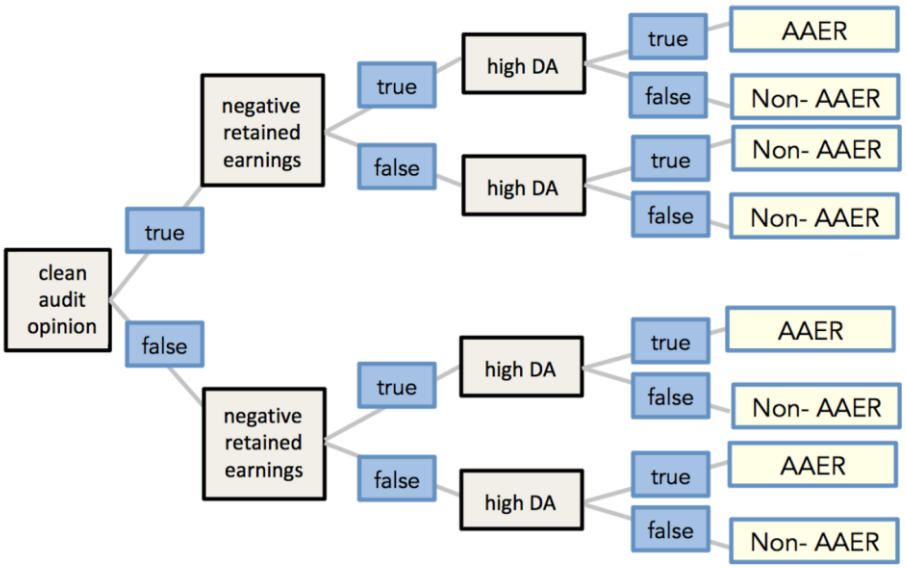
through various lenses, whether it be theory-driven, data-driven, analytical, behavioral, or archival empirical.

We follow Dechow, Ge, Larson, and Sloan, 2011 (DGLS hereafter) for data on accounting restatements, and are grateful they have made their information available for research purposes. DGLS collect data on Accounting and Auditing Enforcement Releases (AAERs) to develop a model to predict misstatements, through a measure called the F-score. They discuss the advantage of this proxy as providing a high level of confidence that restatement was required, while the disadvantage is that identification of instances to investigate is a function of policies and procedures at the U.S. Securities and Exchange Commission (SEC).

Our first test is to compare the relative performance of DGLS's F-score to RF. We expect RF to have better predictive abilities, at the cost of less information on which firm characteristics drove those predictions. We apply a RF technique called "variable importance" to obtain some insight into these effects. We then dig deeper into the forest to perform a novel analysis from the perspective of the trees. Each of our ten trials produces a measure we call V-score, that tells us what percentage of the trees in the forest "voted" to classify the firm as an AAER firm. Where V-scores are low, all or most of the trees "agreed" it *was not* an AAER firm. Where V-scores are high, all or most agreed to classify it as AAER. The mid-values of V-scores indicate uncertainty among the trees, where approximately half the trees wanted to classify the firm as AAER while the other half disagreed. In other words, some firms were more difficult to classify than others. We explore what this uncertainty among the trees can tell us about firm characteristics and financial reporting.

METHODOLOGY

Here we present a simple example to explain Random Forests. The method begins with a decision tree, with TRUE/FALSE decisions at every nodes. The nodes reflect "features" or characteristics that we expect to be useful in classification. For example, suppose we expect three features to be useful in classifying firms as AAER or Non-AAER: (1) audit opinion (2) retained earnings (3) discretionary accruals. Suppose we feed this information to a machine, and it learns to map our data (audit opinion, retained earnings, and discretionary accruals) to outcomes (AAER or Non-AAER) in a period of "training". This might lead to the following decision tree (where DA is discretionary accruals):



Once the learning is complete we can then ask the decision tree to make a classification prediction about a new firm based on its audit opinion, retained earnings, and DA. These predictions will be somewhat accurate but can be improved upon. Suppose we build another decision tree with slightly different features. Perhaps this one learns how to classify firms based on audit opinions, cash flows, and discretionary accruals. Then we add another tree that classifies firms based on retained earnings, cash flows, and leverage. Adding more trees creates a forest; randomizing various features among the trees creates a random forest, which is more accurate at predictions than any single decision tree.

DATA

Outcomes Data

Our Random Forest requires data on classification features as well as the outcomes (AAER vs Non-AAER) related to that data. For outcomes information, we obtain a list of AAER firms from the Center for Financial Reporting and Management (CFRM) at UC Berkely[8]. This dataset, which tracks quarterly and annual SEC enforcement actions, was compiled by DGLS for their 2011 paper, and was kindly shared by the authors. The time period of AAERs spans 1971-2002, which we extend by 5 years in either direction to examine data from 1966 to 2007.

Features data

[8] We used version 20160930 of this dataset.

For features data we focus on *Compustat*, which Tallapally and Luehlfing (2011) see as increasingly similar to the publicly-available Edgar Online system, therefore reflects information used by investors,

Auditor, and regulators. We begin with all for 1,371 *Compustat Quarterly* variables for all firms listed in in *Compustat Quarterly* from 1966 to 2007 (41 years). We retain only numerical variables since the string data type are mostly unique identifiers such as the business name and address, then remove variables that are industry-specific. We assign a value of zero to any missing data.

Principal Component Analysis of the remaining variables reveals that many are highly correlated, therefore we use judgment to identify 250 relatively orthogonal parameters that should appropriately capture firm characteristics. This is known as *features engineering* in machine learning. Among these parameters are the variables used in DGLS, as reported in Appendix A.

We then expand our sample size by collect these 250 variables firms listed *Compustat Quarterly* from 1962 to 1994 (an additional 32 years), creating a dataset of xxx variables for 654,774 observations on 11,932 firms. The firm count by year and quarter in Table A shows the highest number of firm-quarters from 1982 to 1995, but we have no a priori reasons to believe the distribution is problematic.

Table A: Firm Count

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
1962	34 39	34	44	
1963	39 44	40	47	
1964	41 48	41	49	
1965	47 54	47	54	
1966	74 82	74	84	
1967	83 91	82	98	
1968	85 94	91	101	
1969	93 95	93	108	
1970	111	119	114	283
1971	199	242	240	2567
1972	266	277	272	2689
1973	290	300	297	2776
1974	344	350	358	2809
1975	1369	1557	1679	2873
1976	2545	2678	2749	2836
1977	2873	2860	2837	2871
1978	2822	2814	2791	2810
1979	2807	2791	2770	2788
1980	2727	2723	2701	2783
1981	3024	3012	2997	4871

1982	5100	5115	5087	6400
1983	6097	6197	6208	6694
1984	6434	6508	6447	6747
1985	6572	6554	6498	7205
1986	6668	6858	6852	7566
1987	7022	7115	7138	7566
1988	7033	7034	6970	7263
1989	6822	6793	6675	6782
1990	6418	6369	6231	6344
1991	5890	5891	5790	5841
1992	5559	5565	5463	5543
1993	5336	5332	5229	5601
1994	5366	5383	5244	5301
1995	5078	5058	4915	4960
1996	4712	4688	4590	4685
1997	4488	4473	4346	4376
1998	4174	4122	4020	4046
1999	3869	3842	3713	3741
2000	3591	3574	3451	3439
2001	3303	3282	3187	3208
2002	3102	3094	3012	3002
2003	2912	2919	2841	2843
2004	2738	2736	2665	2675
2005	2598	2581	2504	2514
2006	2422	2434	2373	2360
2007	2286	2279	2203	2191
2008	2126	2124	2065	2072
2009	2011	2030	1980	1984
2010	1930	1942	1895	1966
2011	1835	1834	1792	1912
2012	1793	1832	1819	1876
2013	1832	1850	1802	1813
2014	1761	1767	1716	1706
2015	1651	1526	480	114

Outcomes data

Our outcomes information begins with the CFRM dataset (see previous Overall Data) on firms subject to AAERs by the Securities and Exchange Commission. While the CFRM provides various details on AAERs, we use the "Detail" file which lists one observation per firm-misstatement event. There is a total of 1,540 firm misstatement events, where each event can be comprised of multiple violations in the same time period, against more than one party in the firm.

From this set, we remove 57 observations with no time period assigned to them, and a further 415 with missing CIK numbers, resulting in a sample of 1,090 AAER observations. At this point, we do not reduce our sample further, unlike DGLS who remove observations where the violation did not directly affect accounting numbers. Our objective is to identify AAERs first, with further analysis on accounting afterwards.

This sample includes 784 (72%) observations that relate to a violation affecting at least one of: assets, debt, inventories, liabilities, receivables, revenues, other, equity, cost of goods sold, reserves, or accounts payable. The table below tabulates the top 5 accounting areas affected. A further 56 relate to accounting disclosure.

	Affecting quarterly information	Affecting annual information	Accounting				
			Revenues	Assets	Receivables	Inventory	Cost of goods sold
# of observations	803	693	431	203	168	120	97
% of sample	74%	64%	40%	19%	16%	11%	9%

Random Forests

We first explore the performance of the Random Forest (RF) in classifying firms, with a particular focus on accuracy, precision, and recall. While the result is informative, the RF technique lacks measures for statistical inference, preventing us from forming judgments about the parameters of a population and the reliability of the statistical relationship.

To compensate for this, we run the RF multiple times to observe whether results are stable. Based on Wagner, Hastie, and Efron (2010) we re-run the RF 10 more times to obtain a sufficient distribution for inference without overfitting the data.

We design our RF re-runs to explore three sources of variation that could affect results: Non-AAER samples, training period, and allocation of features. In essence we run the RF 30 more times, each time holding two of the three sources of variation constant in order to separate out the effect of each source.

The first source of variation is the subset of Non-AAER firms used in the RF. We randomly select 10 other samples of Non-AAER firms, and re-run the RF keep the training period and allocation of features constant. The second source of variation is the training period. Keeping everything else constant, we vary the training period from 60% to 80% of the data to better understand the learning component of our model.

Finally we explore whether the feature allocation has an effect on results. For this analysis, we keep the list of AAER firms and Non-AAER firms constant, and vary how the features are allocated among the trees. In our program, the features are allocated based on a random seed. We keep this seed constant throughout the analysis except for our last set of results, where we randomize it.

Logistic regressions

We compare the Random Forest performance to another technique for classification: logistic regression. Our objective is to answer two different but related questions. We want to first contrast the two methods when both are run/estimated with the same set of features. This provides insight into the relative power of the two methods. Next, we use the logistic regression method to compare the performance of our feature variables relative to those proposed by extant literature, to get a sense of how well current theories can explain AAER classification.

RANDOM FOREST RESULTS

Main result

To complete

Proceedings of the **2019 International Conference on Big Data in Business**
October 29th to October 31st, 2019

ANALYSING AND COMPARING SAMPLING ALGORITHMS FOR EFFECTIVE FRAUD DETECTION.

Shalini .U. Nair
MSc. Advanced Computer Science – 2019

Abstract

One of the primary difficulties looked by the financial business today is exchange robbery (Fraud), which has brought about huge losses not only to the banks but to the world economy. Organizations are progressively turning towards cutting edge examination and AI calculations to perceive exchange designs that demonstrate burglary. Fraud detection a classification problem, the current methodologies expect the fundamental training set to be consistently appropriated. However, in a class imbalanced characterization, the training set of one class (majority) exceeds the training set of the other class (minority), in which, the minority class is often the more intriguing class and hence has caught the attention of many researchers. The objective of the think about is to implement a fraud detection system (FDS), investigate and overcome the problems of imbalance using the sampling techniques, discover out the qualities and shortcomings of the sampling techniques and to provide an abreast and comprehensive analysis about the subject. The work aims to supply researchers and professionals with learning strategies utilized in fraud detection research for their quick get to and usage. The experiments verify that the selection of appropriate sampling algorithm has prevalent execution on precision, compared to no examining, and it too appears as an advantage within the running time and cost.

Keywords

Fraud; Datamining; Classification; Imbalanced data set; Machine learning (ML); Oversampling; Under sampling; Hybrid methods.

1. Introduction

Gosset et al.,(1999) state that the definition of fraud is difficult to be formed since the distinction between fraudulent and legitimate behaviours is not always obvious. Fraud can be perpetrated everywhere including financial institutions, insurance companies, corporations as well as the government. Sinayobye et al.,(2018) refers to fraud as obtaining services, goods, and money by the unethical way, and is a growing problem in the world today. It is a crime where the purpose is to appropriate money by illegal means. On the other hand, Alexopoulos et al.,(2008) defines fraud as "the deliberate and premeditated act perpetrated to achieve gain on false ground. The consequences of fraud are not restricted to economic losses, but they can also lead to violation of human rights, physical and psychological harms as well as premature deaths."

Frauds can be committed in several ways; this creates huge complexities in designing the detection systems. The major difficulty arises with the fact that, as the fraud detection techniques evolve, so does the fraud execute patterns. The behavioural adaptations are quite rapid due to the evolution and availability of advanced technologies to both researchers and fraudsters. Another major drawback in creating a compelling fraud detection framework is the requirement for confidentiality and the requirement for anonymization of the records prior to the actual pattern mining. This leads to the loss of several behavioural patterns contained in the transactions. Thirdly, the implicit imbalance existing in transactions also provide a huge downside in determining an efficient algorithm for fraud detection (Nadarajan et al.,2016).The design of cost-effective fraud detection solution is the key to reducing the losses using mathematical algorithms. (Sinayobye et al., 2018).

The class imbalance can be inherent property or due to confinements to get information such as cost, security and huge exertion. Class imbalance is one of the challenges of machine learning and information mining areas (Elrahman et al., 2013). Real-world applications such as medical diagnosis, fraud detection (credit card, phone calls, insurance), network intrusion detection, pollution detection, fault monitoring, biomedical, bioinformatics and remote sensing (land mine, under water mine) endure from these phenomena (Elrahman et al.,2013). Hence Elrahman et al.,(2013) suggests that it is significant for a classification model to be able to realize higher recognizable proof rate on the rare events (minority course) within the datasets.

This research discusses the problems of identifying fraudulent behaviour in a Fraud detection system (FDS) and *aims* to supply a few answers by centering on pivotal issues such as: i) *Identifying the datasets*, ii) *why and how sampling is useful* in the presence of *class imbalance* , iii) *improving fraud detection accuracy*, iv) how to *assess performances* in a way which is relevant for detection (Pozzolo, 2015),v) Finally, a design and assessment of a *prototype* of a *Fraud Detection System* to be able to meet the real-world working conditions that can classify fraud transactions accurately. Sampling techniques are analysed compared and used to upgrade or change unequal data distribution into breakeven data samples thereby providing information balancing. The study uses decision tree (DT) which is a supervised machine learning methodology performing the classification task.

The rest of the report is outlined as follows Section 2 contains the Literature review of the related work in the fraud detection and sampling. Section 3 tells about the techniques used to handle classification imbalance problem Section 4 addresses the Methodology used in the study to classify the transactions, overcome the imbalance problem along with the weaknesses and strengths of the sampling techniques chosen, Section 5 details on the evaluation metrics used, Section 6 contains the Results and Discussion Finally, Section 7 concludes the paper.

2. Literature Review of the Related Work

Over and above 90% of the fraud is online and the fraud tracking solutions make use of transaction rules to detect suspicious transactions. The rules are set either by human review, prior knowledge or by experts. Shockingly this conventional approach of utilizing rules or rational explanation to inquiry exchanges is still utilized by a few banks and instalment portals. The "rules" in such systems are a compound of information and horizon-scanning. The result of these forms is by and large two fold naming the exchanges as Bonafide or extortion (legitimate or fraud).

Padmavathamma et.al.,(2012) used the traditional Rule-based method in telecommunication fraud detection while Akhilomen ,(2013) made use of Neural Networks in his Cyber credit card fraud detection system .His findings were that the Major impediment of the Rule-based system was the occurrence of false positives and the Judgment is dependent on individual trainings and exchange guidelines, which vary depending on the trade. Hence, rules ought to be the last line of defence in the fraud detection.

Sorournejad et al., (2016) in his Survey mentions of both supervised and unsupervised ML based approaches involving ANN (Artificial Neural Networks), Decision Trees, SVM (Support Vector machines), HMM (Hidden Markov Models), clustering etc for Credit Card Fraud Detection .Sahin et al., (2013) used C5.0, C&RT and CHAID in his study and concluded that the three DT implementations outperformed the SVM implementation. Decision tree classifiers are the finest conventional techniques for classification. In spite of the fact that it is capable, it lacks in generalization exactness for the concealed information.

Elrahman et al.,(2013) in his review on Class Imbalance Problem mentions of the Imbalance in data that hinder the performance of machine learning and data mining techniques, High overlapped classes and high noise level produced higher complexity, Issues of concept complexity or overlapping, which refers to level of separability between data classes. Few approaches have been proposed to overcome the issues of imbalance class problems- Re-sampling, cost sensitive boosting and feature selection at the data level and other ones at the algorithm level such as Improved algorithm, single class learning, ensemble method and hybrid approach.

Lusa et al.,(2012) studied the Random Forest, K-NN, SVM and PAM (Prediction Analysis of Microarrays) and came to a conclusion that In most imbalance problems, cost of errors for different classes is uneven and usually it is unknown. Feature selection is another critical issue in machine learning and data mining. High dimensional data and irrelevant features may reduce the performance of the classifier and increase the misclassification rate .Evaluation metrics is a critical issue in machine learning-F-measure and Mathew correlation coefficient (MCC) (Lusa et al., 2012) .

Chawla et al.,(2002) proposed a synthetic method called SMOTE, They used Naive Bayes and the decision tree algorithms in their study and stated that Under sampling may cause loss of useful information by removing significant patterns and over sampling may cause over fitting and may introduce additional computational task. To overcome this problem a synthetic minority over sampling technique (SMOTE) can be used by generating a synthetic example rather than replacement with replication.

In spite of various efforts made towards handling class imbalance problems through sampling, a study by Weiss et al., (2003) argued that there is no formal standard to define the most suitable class distribution, and experiments conducted by Visa, S.,(2005) discovered that often, a 50:50 imbalance ratio between minority class and majority class in training set does not always return the optimal classification performance. Despite the disadvantages with sampling, it is still a wellknown approach to handle imbalanced datasets compared to cost-sensitive learning algorithms (Vaishali. G ,2012).Nowadays many class imbalance datasets come in larger volume than before, thus motivating sampling to reduce the data size for a feasible learning task (Vimalraj et al.,2018).

3. Classification Imbalanced Problem Handling Techniques

Learning from unbalanced datasets is a difficult task since most learning algorithms are not designed to cope with a large difference between the numbers of cases belonging to different classes (Batisa et al., 2000). There are several methods that deal with this problem. A few of them are given below.

- Resample the training set.
- Resample with different ratios
- Use the right evaluation metrics
- Using K-fold Cross-Validation
- Distance-based methods
- Algorithm level methods ○ Cost-based methods ○ Ensemble based methods ○ Recognition based methods

4. Methodology

The study uses python programming and Knime tool to design the FDS model. Data pre-processing is done using python. The Workflow for the FDS model is built using the Knime tool. The source of the dataset is -

Credit card fraud – Machine Learning Group - ULB (2019), Insurance claims fraud dataset (Databricks), Financial Payment

fraud dataset-TESTIMON (2017). The credit card dataset didn't need any pre-processing whereas the Insurance and Financial Payment dataset were pre-processed. I have compared the performance of the eight sampling methods on the mentioned datasets .For ease of visualization, I performed Dimensionality reduction via principal component analysis.

4.1 System Flow

The fraud detection method proposed in this paper is based on the data mining techniques. Firstly, the data is loaded into the system from source (csv file) ,then data is pre-processed, and the features are selected according to the specific fraud types. Dataset is split into training and test and based on various sampling techniques the samples are picked. Decision tree is used as a classifier to classify the data as fraud or legal. Training Model of the FDS system is devised by applying the classifier on the training set. The test data is then passed to the model for prediction. Evaluation metrics are used to score the results. Finally, the experimental results are analysed. It is a data mining technique that transforms raw data into a readable format. Raw data(real world data) is always unpolished and that data cannot be sent through a model. That would cause certain errors. That is why we need to pre-process data before sending through a model. There is a lot of redundant information, imbalance and noise, which seriously affects the efficiency of data analysis.

Data pre-processing includes data cleaning, data integration, data selection, data transformation. Data cleaning eliminates noise and inconsistent data, whereas data integration helps to combine multiple data sources together, data selection achieves data extraction from the data related to the task analysis, data transformation transforms or unifies the data into a suitable mining form. These steps need not necessarily be used. One of the most important steps is the data selection. There are several commonly used methods for selecting feature attribute sets, including data cube aggregation, wavelet transform, attribute subset selection, principal component analysis and so on. Principal Component Analysis (PCA) is computationally expensive and can handle sparsely sloped data. It can treat multidimensional data as K-dimensional problems. Principal components are used as inputs for fraud analysis.

The underlying system data flow shown in the Fig 1 briefly describes the nodes and workflow Management, and how the interactive views communicate. The fraud detection is initiated by feeding the number of transactions made by the customers as Data Input; these transactions are fed in a csv file format. Data Table hold the meta-information concerning the type of its columns in addition to the actual data. The data can be accessed by iterating over instances of Data Row. Each row contains a Row ID unique identifier and a specific number of Data Cell objects, which hold the actual data.

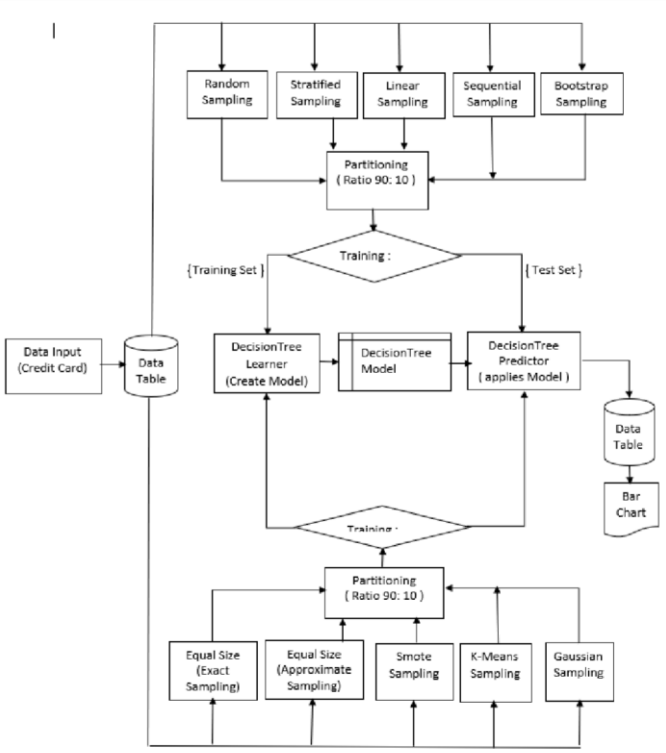


Figure 1. System Flow of the FDS with the Sampling algorithms

Transactions from the Data Table are fed to the sampling algorithm, it is a procedure that allows us to select a subset of units (a sample) from a whole transaction without enumerating all the possible samples of the transactions, at the same time split the whole dataset into two parts Training and Test based on 90:10 ratio. Training set is fed to Decision Tree Learner and then the Model is built. During the training phase, legal transaction pattern and fraud transaction pattern of each customer is created from their legal transactions and fraud transactions, respectively, by using frequent item set mining. During the testing phase, Trained Model and Test sets are passed to predictor for fraud detection, the matching algorithm detects to which pattern the incoming transaction matches more. If the incoming transaction is matching more with legal pattern of the particular customer, then the algorithm returns "0" (i.e., legal transaction) and if the incoming transaction is matching more with fraud pattern , then the algorithm returns "1" (i.e., fraudulent transaction). Prediction results are stored in the Data Table and graph is plotted based on the measuring metrics.

4.2 Sampling Algorithms

With rapid increase of data amount, the database is not a previously small scaled database and it develops into a multidimensional massive one (Chyouhwa et al.,2011). Original data mining method cannot obviously solve massive data problem, so the sampling algorithm of large data sets appears at the same time. The sampling algorithm can effectively and rapidly mine frequent item set under condition that satisfies certain mining accuracy. Thus, data mining algorithm can analyse the sampling data, which is far smaller than original big data set to largely improve the performance. The proposed sampling algorithms are random sampling, stratified sampling, linear sampling, sequential sampling, Bootstrap sampling, Equal size

sampling (Exact /Approximate), Synthetic Minority Over-sampling Technique (SMOTE), K-Means Clustering Sampling. Knime tool has nodes with these sampling algorithms hence they have been included in the study.

4.2.1 Random Sampling (RS)

A simple random sample is a subset of individuals (a sample) chosen from a larger set (a population) (Yates et al.,2008). Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals. **Advantages** - Easy method to use, No need of prior information of population, Equal and independent chance of selection to every element, Represents the target population and eliminates sampling bias.
Disadvantages - If sampling frame is large this method is impracticable, it does not represent proportionate representation.

4.2.2 Stratified Sampling (SS)

This is a method of sampling from a population which can be partitioned into subpopulations. When subpopulations within an overall population vary, it could be advantageous to sample each subpopulation (stratum) independently (Yates et al., 2008). Stratification is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should define a partition of the population i.e. it should be collectively exhaustive and mutually exclusive: every element in the population must be assigned to one and only one stratum. Then simple random sampling is applied within each stratum. (Hunt et al., 2011).
Advantages - Highly representative of the target population and high in statistical efficiency.
Disadvantages - Classification error. Time consuming and expensive. Prior knowledge of composition and of distribution of population required.

4.2.3 Linear Sampling(LS)

This method always includes the first and the last row and selects the remaining rows linearly over the whole table (e.g. every third row). This is useful to down sample a sorted column while maintaining minimum and maximum value. **Advantages** - Easy to Execute and Understand. Useful to down sample a sorted column while maintaining minimum and maximum value.
Disadvantages - Sampling error if there is a hidden pattern. Difficult when the size of a population cannot be estimated.

4.2.4 Sequential Sampling (SES)

Sequential sampling is not a probability-based sampling technique. In this method the researcher picks a single or a group of samples in each time interval, studies the subjects, analyses the results and then picks another group of samples if needed and so on. This technique includes a sampling plan in which an undetermined number of samples are tested one by one, accumulating the results until a decision can be made (Knime, 2019).
Advantages -Due to the repetitive nature of this sampling method, minor changes and adjustments can be done during the initial parts of the study to correct and hone the research method. It is not expensive, not time consuming and not workforce extensive.
Disadvantages - Hardly representative of the entire population. Hardly randomized

4.2.5 Bootstrap Sampling (BS)

Bootstrapping is a sampling technique, which randomly draws rows from the input with replacement. The output table will therefore likely contain duplicate rows while other rows are not present in the output at all. Bootstrap is a data resampling scheme introduced by Bradley Efron in 1973. Stephaniein (2016) defines bootstrap as a sample that is a smaller sample that is "bootstrapped" from a larger sample. Bootstrapping is a type of resampling where large numbers of smaller samples of the same size are repeatedly drawn, with replacement, from a single original sample.
For example, let's say your sample was made up of ten numbers: 49, 34, 21, 18, 10, 8, 6, 5, 2, 1. You randomly draw three numbers 5, 1, and 49. You then replace those numbers into the sample and draw three numbers again. Repeat the process of drawing x numbers B times. Usually, original samples are much larger than this simple example, and B can reach into the thousands. After many iterations, the bootstrap statistics are compiled into a bootstrap distribution. You're replacing your numbers back into the pot, so your resamples can have the same item repeated several times (e.g. 49 could appear a dozen times in a dozen resamples). Bootstrapping is loosely based on the law of large numbers, which states that if you sample over and over again, your data should approximate the true population data (Stephanie, 2016).
Advantages - Single sample method can serve as a mini population, from which repeated small samples are drawn with replacement repeatedly. As well as saving time and money, bootstrapped samples can be quite good approximations for population parameters.
Disadvantages - A bootstrap sample can only tell you things about the original sample and won't give you any new information about the real population. It is simply a nonparametric method for constructing confidence intervals and similar.

4.2.6 Equal Size Sampling (Exact / Approximate)

This sampling technique removes rows from the input data set such that the values in a categorical column are equally distributed. This can be useful, for instance if a learning algorithm is prone to unequal class distributions and you want to downsize the data set so that the class attributes occur equally often in the data set. The technique will remove random rows belonging to the majority classes. The rows returned by this method will contain all records from the minority class(es) and a random sample from each of the majority classes, whereby each sample contains as many objects as the minority class contains(KNIME ,2019) **Advantages** - Minority class is considered, and majority class is downsized.

Disadvantages - Classification error. Information loss. Memory expensive.

4.2.6.1 Exact Sampling (ESS)

If selected, the final output will be determined up-front. Each class will have the same number of instances in the output table. This sampling is slightly more memory expensive as each class will need to be represented by a bit set containing instances of the corresponding rows. In most cases it is safe to select this option unless you have very large data with many different class labels(KNIME ,2019).

4.2.6.2 Approximate Sampling (ASS)

If selected, the final output will be determined on the fly. The number of occurrences of each class may slightly differ as the final number can't be determined beforehand (KNIME ,2019).

4.2.7 SMOTE (Oversample by Minority and Oversample by 2)

SMOTE is a powerful sampling method that goes beyond simple under or over sampling. To overcome the issue of over-fitting and extend the decision area of the minority class samples, a novel technique SMOTE ‘‘Synthetic Minority Oversampling Technique’’ was introduced by Chawla et al.,(2002). This technique produces artificial samples by using the feature space rather than data space. It is used for oversampling of minority class by creating the artificial data instead of using replacement or randomized sampling techniques. It was the first technique which introduced new samples in the learning dataset to enhance the data space and counter the scarcity in the distribution of samples. The oversampling technique is a standard procedure in the classification of imbalance data (e.g., minority class).The pseudo code of the SMOTE algorithm and detail can be found in (Chawla et al.,2002).

Algorithm works roughly as follows: It creates synthetic rows by extrapolating between a real object of a given class and one of its nearest neighbours (of the same class). It then picks a point along the line between these two objects and determines the attributes (cell values) of the new object based on this randomly chosen point (KNIME,2019). The method proposed by Chawla et al., (2002) states that “the minority class records are oversampled not by repetition but based on nearest neighbour method”.

Advantages - Alleviates overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances. No loss of information. It’s simple to implement and interpret.

Disadvantages - While generating synthetic examples, SMOTE does not take into consideration neighbouring examples can be from other classes. This can increase the overlapping of classes and can introduce additional noise. SMOTE is not very practical for high dimensional data.

4.2.8 K-Means Clustering Sampling (KMCS)

According to Bejarano et al.,(2011) this Sampler is a modification of the k-means problem, it is a simple iterativerefinement heuristic due to Lloyd (1982): Start with k arbitrary (or random) centres, compute the clusters defined by them, define the means of these clusters as the new centres, recompute clusters and repeat. Lloyd’s method is fast in practice but is guaranteed to converge only to a local optimum. The purpose of using K-means algorithm clustering is to divide minority class into groups according to the distributing regulation inside it .Gather the similar samples into clusters ,generate new samples in these clusters .Thus ,new samples are produced in the intersection of samples that are as much similar as possible ,which can represent the characteristic of the samples preferably in this cluster ;meanwhile ,every cluster can obtain certain proportion of new samples ,which can guarantee that new samples of whole minority class has better coverage and representation , accord it with the distribute of primary sample space much more(Yang et al., 2012).

In this method we can select groups or clusters, and then from each cluster, select the individual subjects by simple random sampling. You can even opt to include the entire cluster and not just a subset from it. Here the entire dataset is clustered based on the class (cluster_0 / cluster_1).The number of clusters is also to be given here I have taken it as 2.Then perform a random sample and select points from the clusters for 0 / 1 class. The main difference between cluster sampling and stratified sampling lies with the inclusion of the cluster or strata.

Advantages - Cuts down on the cost of preparing a sampling frame.

Disadvantages - Sampling error; overlapping data points. The populations the clusters contain are only representative of that specific group and not the actual population. The number of clusters to be formed should be known.

4.3 Decision Tree Learning for Fraud Detection classification

Machine learning techniques are applied in many areas of research for deriving computational intelligence. It allows for the generalization of specific examples that can be used in modelling, predication, and classification of datasets. A decision tree is one such widely used machine learning technique that has been effective for classification or regression. Its usefulness results from the ability to compensate for missing values and having a highly flexible hypothesis space (Mitchell ,1997).

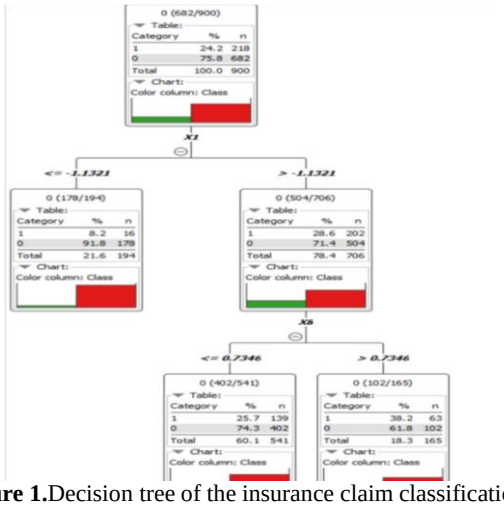


Figure 1. Decision tree of the insurance claim classification.

The ability to derive well defined rules from data set makes Decision tree as one of the most popular classification techniques. The decision tree is a structured tree, having a root node and several internal and leaf nodes. Each leaf node is associated with one class level attribute. In a decision tree, each internal node splits the attribute into two or more branches based on the number of discrete values hold by that attribute. The path from root to a leaf node is included as a classification rule to the rule base. Depending on the type of input features decision tree can be constructed using different induction algorithms like ID3, C4.5 and CART. For instance, ID3 can be applied on discrete values only, C4.5 on continuous values and CART can be applied on both types of data (Sahin et al., 2013).

In this research the knime node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain in two partitions at a given split point. Nominal splits can be either binary (two outcomes) or they can have as many outcomes as nominal values. In the case of a binary split the nominal values are divided into two subsets. The algorithm can be run in multiple threads, and thus, exploit multiple processors or cores .Fig 2 shows the decision tree used (KNIME ,2019).

5. Evaluation Metrics

5.1 Accuracy

Accuracy is the common evaluation standard of classification methods, but the classification result of unbalanced data set can't be evaluated reasonably. This is because the samples of majority class are more than the samples of minority class, if we classify all samples as majority class, the accuracy is still very high, but the recognition rate of minority class is zero(Yong ,2012).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

5.2 Confusion Matrix

All the problems can be classified into two kinds when we tend to solve them. So, the critical point of the research of unbalanced data set classification focus on how to improve the classification performance of minority class, which is one of the two classifications problem (Yong ,2012).Fig 3. shows the mixed matrix of two classes data set. The minority class and the majority class label as positive and negative respectively. TP and TN are samples' amount of minority class and majority class respectively under the condition of right classification of classes. FN and FP are samples' amount of minority class and majority class respectively under the condition of wrong classification of classes (Yong ,2012).For a two-class classifier, a confusion matrix consists of information about actual and predicted classification return by a classifier. Often, a classifier performance is evaluated based on the information obtained from the confusion matrix (Ali et al., 2013).

		Predicted Class: + ve	Predicted Class: - ve
Actual Class	+ ve	True Positive (TP)	False Negative (FN)
	- ve	False Positive (FP)	True Negative (TN)

Figure 2. Confusion matrix

The entries in the confusion matrix shown in Fig 3.are denoted as;

- True Positive (TP) refers to the number of positive examples which are correctly predicted as positives by a classifier.
- True Negative (TN) denotes as the number of negative examples correctly classified as negatives by a classifier.
- False Positive (FP), often referred to as false alarm; defines as the number of negative examples incorrectly classified as positives by a classifier.
- False Negative (FN), sometimes known as miss; is determined as the number of positive examples incorrectly assigned as negatives by a classifier.

However, by analysing the four entries in the confusion matrix is not enough in determining the performance of a classifier. Therefore, several derivatives based from the previously discussed confusion matrix are used in evaluating a classifier in this study (Ali et al., 2013). A few of these are:

5.3 Recall - Recall is the accuracy of properly classified minority class (Yong ,2012) .

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (2)$$

5.4 Precision - Precision is the rate that the properly classified minority class takes up of all classes (Yong ,2012).

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (3)$$

5.5 F- Score -This is a performance score that combines both precision and recall. It is a harmonic mean of these two variables. .If the recall and the precision are

all big, so it can properly reflect the classification performance of the minority class (Yong ,2012). Formula is given as:

$$\frac{(1+\beta 2)\,\text{Recall} * \text{Precision}}{\beta 2\ \text{Recall} + \text{Precision}} \quad (4)$$

Where β is a nonnegative constant and is usually set to 1 so

$$\text{F score} = 2*\frac{(\text{Precision}*\text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4.1)$$

5.6 MCC-Mathew correlation coefficient- Matthews correlation coefficient is used as a measure of the quality of binary classifications. It considers true and false positives and negatives and is a balanced measure which can be used in imbalanced data like CC transaction data. The MCC is a correlation coefficient between the observed and predicted binary classifications and its value is between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction, and −1 indicates total disagreement between prediction and observation. This is a performance score that considers true and false positives, as well as true and false negatives. This score gives a good evaluation for imbalanced dataset. Formula is given as (Seeja and Zareapoor,2014)

$$\text{MCC} = \frac{\text{TP}*\text{TN} - \text{FP}*\text{FN}}{\sqrt{(\text{TP}+\text{FP})*(\text{FN}+\text{TN})*(\text{TP}+\text{FN})*(\text{FP}+\text{TN})}} \quad (5)$$

6. Results and Discussion

The three financial datasets, Credit Card (CC), Insurance Claim (IC) and Financial Payment Services (FPS) were thoroughly analysed at the outset to gain insight into which features could be discarded and those which could be valuably engineered for data distribution of the class. Table 1 shows the percentage of distribution of the fraud and legitimate data in the respective dataset.

Table 1. Percentage wise distribution of the three datasets chosen

Datasets	No Fraud%	Fraud%
CC	99.87	0.31
IC	73.3	24.7
FPS	93.93	6.07

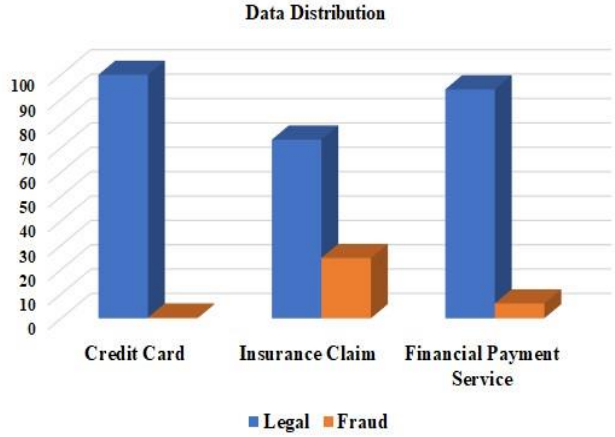


Figure 3. Bar plot of the unbalanced data in the 3 datasets chosen

To deal with the large skew in the data, I choose different sampling techniques along with appropriate metrics such as Accuracy, F Score, MCC and used a machine learning algorithm based on decision trees which works best with strongly imbalanced classes. The study compares each dataset chosen with all the proposed sampling methods: RS, SS, LS, SES, BS, ESS/ASS, SMOTE and KMCS. The Table 2 shows the confusion matrix with the samples tested compares the measuring metrics (Accuracy, F Score, MCC) of different Sampling Methods.

The evaluation metrics values (shown in Table 2) are devised from the confusion matrix values using their respective formula. The Fig 5. Confirms that accuracy is almost similar for all sampling methods for all the datasets. In small dataset (IC) BS performance (78%) and SMOTE sampling (72%) is better than the others. While in larger data set like CC the accuracy result is above 90% for all the sampling techniques. ESS being 92% and the others 99%. This is because with ESS there is undersampling done for the majority class and information is lost.

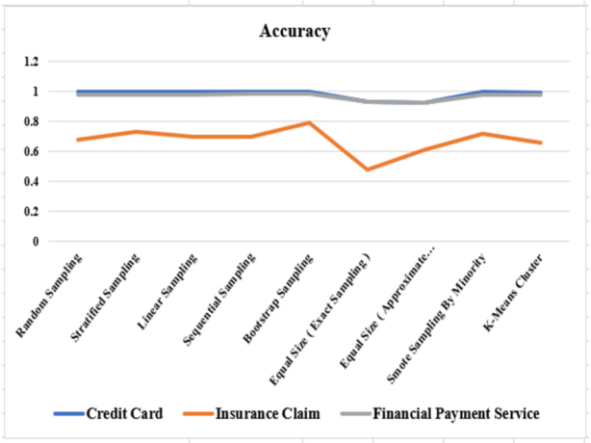


Figure 5. Accuracy score of the sampling algorithms for the 3 datasets chosen.

Similarly, for the FPS dataset equal size sampling is around 93% as compared to the others with 97% accuracy. Even though the accuracy score of these sampling algorithms yield an accuracy (Fig 5.) of 99% which seems impressive.

However, minority class could be totally ignored in case of imbalanced datasets, and this can prove expensive in some classification problems e.g. - in case of a CC fraud, which can cost individuals and businesses lots of money.

The results and the plots for F score (Fig 6.) show that, SMOTE performed better for fraud detection of imbalanced dataset about 99% followed by ESS and SS at 92%. KMCS and RS showed approximately 88% F score accuracy while SES showed the least i.e. 78% correct prediction. Sampling method such as ESS and SS, show promising F1-Score (Fig 6.) but considering the volume of data used for training the model SMOTE provides better results. I also observed that, use of appropriate data exploration technique enhanced the performance of model. With the increase of positive samples (Table 2) in the training sample set the performances of all methods increase. Using K-fold cross validation improves the performance of the model. Moreover, using algorithms which handle imbalance dataset by itself also enhance the model performance.

Figure 6. F Score of the sampling algorithms for the 3 datasets chosen

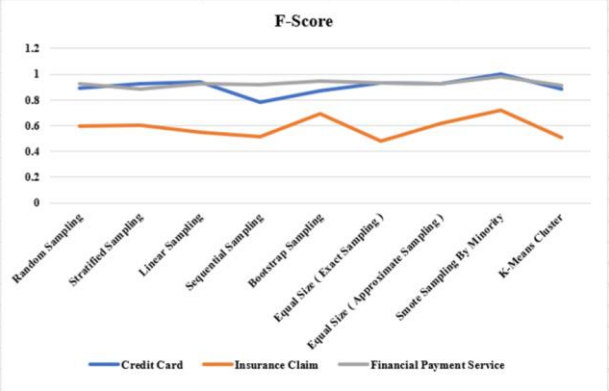


Table 2. Study results for the three datasets for the sampling algorithms proposed and the used metrics

Dataset	Sampling	Training Samples	Test Samples	Confusion Matrix				Accuracy %	F Score %	MCC %	Recall %	Precision %
				TP	FP	FN	TN					
CC (Legal 99.87% : Fraud 0.31%)	RS	256326	28481	28426	4	16	35	0.9993	0.8887	0.7845	0.8431	0.9480
	SS	256326	28481	28429	3	11	38	0.9995	0.9221	0.8476	0.7760	0.9270
	LS	256326	28481	28426	0	12	43	0.9996	0.9387	0.8840	0.8910	1
	SES	256326	28481	28453	6	11	11	0.9994	0.7819	0.5685	0.7500	0.8300
	BS	230693	25633	25576	2	21	34	0.9991	0.8734	0.7637	0.8000	0.9200
	ESS	984	99	42	5	2	50	0.9293	0.9293	0.8594	0.9200	0.9300
	ASS	960	96	39	3	4	50	0.9271	0.9261	0.8524	0.9270	0.9250
	SMOTE	511767	56863	28306	78	38	28441	0.9980	0.9980	0.9959	0.9980	0.9980
	KMCS	256336	28481	28419	2	21	39	0.9900	0.8860	0.7860	0.8250	0.9750
IC (Legal 73.3% : Fraud 24.7%)	RS	900	100	57	16	16	11	0.6800	0.5941	0.1882	0.5048	0.5051
	SS	900	100	65	10	17	8	0.7300	0.6001	0.2104	0.4933	0.4925
	LS	900	100	64	14	16	6	0.7000	0.5479	0.0966	0.5466	0.5500
	SES	900	100	66	13	17	4	0.7000	0.5127	0.0281	0.5130	0.5152
	BS	810	90	61	8	11	10	0.7889	0.6890	0.3809	0.5890	0.5862
	ESS	444	50	17	11	15	7	0.4800	0.4800	-0.077	0.3458	0.3417
	ASS	457	51	16	8	12	16	0.6154	0.6154	0.2381	0.4131	0.4114
	SMOTE	1355	151	54	22	20	55	0.7219	0.7218	0.4439	0.6870	0.6866
	KMCS	900	100	61	17	17	5	0.6600	0.5047	0.0093	0.5200	0.5266
FPS (Legal 93.93% : Fraud 6.07%)	RS	12177	1353	1253	16	13	71	0.9808	0.9257	0.8191	0.5048	0.5051
	SS	12177	1353	1263	8	13	69	0.9763	0.8856	0.8602	0.4933	0.4925
	LS	12177	1353	1259	11	16	67	0.9808	0.9249	0.8221	0.5466	0.5500
	SES	12177	1353	1278	6	14	55	0.9852	0.9192	0.8402	0.5130	0.5152
	BS	10959	1218	1116	16	19	67	0.9877	0.9477	0.7776	0.5890	0.5862
	ESS	1477	165	75	3	6	81	0.9333	0.9333	0.8914	0.3458	0.3417
	ASS	1450	161	86	7	2	67	0.9281	0.9280	0.8892	0.4131	0.4114

	SMOTE	22876	2542	1230	39	42	1231	0.9815	0.9815	0.9363	0.6870	0.6866
	KMCS	12177	1353	1260	10	14	69	0.9778	0.9107	0.8427	0.5200	0.5266

MCC score for all the three data sets with the sampling algorithms seems appreciating (Fig 7). Though the score has been very low (max 44%) with the small dataset (IC) it is high (above 95%) with the larger dataset (CC and FPS). MCC score for SMOTE sampling shows best result in small as well as large dataset. Supervised learning such as decision tree performed well in this study.

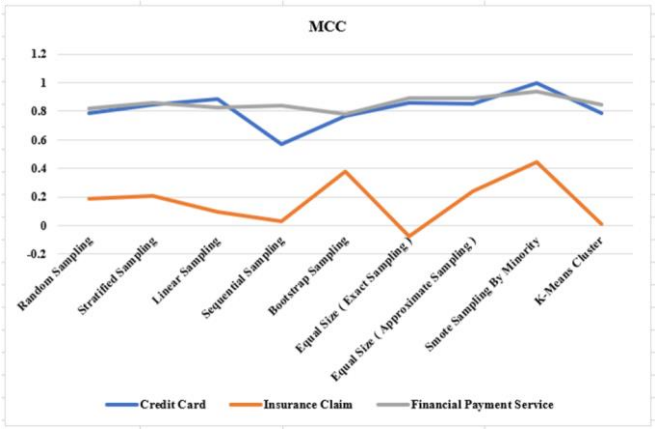


Figure 7.MCC score of the sampling algorithms for the 3 datasets chosen.

As summarized by three datasets, there is lesser data in the IC and FPS datasets. So, when we find anomalies in fraud detection, we obtain a lower of accuracy, F-Score and MCC for all the sampling methods, because we trained the systems for a small number of data and validated the test data for a lesser amount. Conversely, when I applied this model with a large dataset the CC Dataset, I got metrics values of more than 98%. Larger the dataset better is the model prediction and the metric performance. Supervised learning dataset is suitable for history database for credit card fraud detection. In general, there is no single strategy that is currently superior to all the others with respect to data or algorithms. Even if we find one then we need to evaluate it in terms of time, cost, and computation power. This will need testing on several databases and algorithms using all strategies. Experimental results support the idea that the final performance is extremely dependent on the data nature and distribution. It is evident that this study can select few candidates that perform better than others without exploring the whole dataset as in the case of BS and KMCS for smaller dataset. In general, SMOTE is the method returning the larger accuracy when large imbalanced dataset is considered.

7. Conclusion

Ideally, there would be trade-offs between false positives and false negatives in real life. Accuracy is very high, despite the presence of false positives and false negatives. However, the F1 score and MCC score tell a better story of the actual model performance. It is noticed that most fraud detection systems in all areas use supervised approach. In addition, the most commonly used machine learning techniques are Artificial Neural Networks (ANN), Decision tree, Support Vector Machines (SVM), Naive Bayes, Random forest and K-NN algorithms. These techniques can be used alone or combined with an ensemble or metalearning techniques to build strong detection classifiers. Results show that SMOTE methods perform better than any other sampling method. Depending on the resource of imbalanced datasets a range of sampling methods have been used to resolve imbalance. The overall performance of ML models built on imbalanced datasets, will be constrained by its ability to predict rare and minority points. Identifying and resolving the imbalance of those points is crucial to the quality and performance of the generated models. This research study can be extended in future to resolve imbalance between Big data application in various streams. The concept of combining classifiers is proposed as a new direction for the improvement of the performance of individual classifiers. Taking advantage of power of data mining in finding hidden patterns or anomalies in financial data fraud detection we can focus on shared data sets (available to the public and researchers) that contains samples of different malicious behaviour.

Acknowledgement

I would like to thank Dr.Varun Ojha for guiding and supporting me all through the project development tenure. Your discussion, ideas, and feedback have been invaluable.

References

Aida Ali, Siti Mariyam Shamsuddin and Anca L.Ralescu,(2013) " Classification with class imbalance problem: a review",*Int. J. Advance Soft Compu. Appl*, Vol. 5, No. 3, ISSN 20742827, (accessed : 15 June 2019).

Alexopoulos, P., Kafentzis, K., Benetou, X., Tagaris, T., and Georgolios, P. ,(2008) "Towards a generic fraud ontology in e-government", in *Proceedings of the International Conference on Security and Cryptography, Portugal*, (accessed : 3 July 2019)

Andrea Dal Pozzolo , (2015) "Adaptive Machine Learning for Credit Card Fraud Detection",*Université Libre de Bruxelles Computer Science Department Machine Learning Group* , (accessed : 5 July 2019). Databricks, " Data Science - Insurance Claims", Available at : https://databricks-prodcloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4954928053318020/1058911316420443/167703932442645/latest.html , (accessed : 24 June 2019). Efron, B., Rogosa, D., &Tibshirani, R. (2004) "Resampling methods of estimation", *N.J. Smelser, & P.B. Baltes (Eds.). International Encyclopedia of the Social &Behavioral Sciences*, (pp. 13216 – 13220), New York, NY: Elsevier, (accessed : 11 June 2019).

Hunt, Neville; Tyrrell, Sidney (2001). "Stratified Sampling",*Webpage at Coventry University*, Archived from the original on 13 October 2013 (accessed : 12 July 2019).

Janvier Omar Sinayobye ,Fred Kiwanuka ,SwaibKaawaaseKyanda, (2008) "A State-of-the-Art Review of Machine Learning Techniques for Fraud Detection Research" , *ACM/IEEE Symposium on Software Engineering in Africa*,(accessed : 16 July 2019)

Jeremy Bejarano, Koushiki Bose, Tyler Brannan, Anita Thomas, (2011) "Sampling Within k-Means Algorithm to Cluster Large Datasets" , *Technical Report UMBC High Performance Computing Facility (HPCF)* , (accessed : 30 June 2019)

John Akhilomen,(2013) "Data mining application for cyber creditcard fraud detection system," *Lecture Notes in Engineering and Computer Science*, pp. 1537-1542, (accessed : 24 June 2019) .

KNIME ,(2019)Base Nodes version 4.0.1.v201908131444 by KNIME AG, Zurich, Switzerland,URL: https://nodepit.com/node/org.knime.base.node.preproc.equalsizesampling.EqualSizeSamplingNodeFactory , (accessed : 15 June 2019)

K. R. Seeja and Masoumeh Zareapoor, (2014) "FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining",*The Scientific World JournalVolume 2014*, Article ID 252797, (accessed : 20 Aug 2019).

L. Lusa and R. Blagues,(2012) " The Class-imbalance for highdimensional class prediction",*11th International Conference on Machine Learning and Application, IEEE*, (accessed : 10 June 2019).

Machine Learning Group-ULB, Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud , (accessed : 24 June 2019).

N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelemer, (2002) "SMOTE: Synthetic Minority Over-Sampling Technique" ,*Artificial Intelligence Research*, vol. 16, pp. 321357,.

P. Gosset and M. Hyland, (1999) "Classification, detection and prosecution of fraud in mobile networks" ,*Proceedings of ACTS Mobile Summit, Sorrento, Italy*, (accessed : 3 June 2019).

Rajani Padmavathamma,(2012) "A Model for Rule Based Fraud Detection in Telecommunications" , *International Journal of Engineering Research & Technology (IJERT)*, Vol. 1, ISSN: 2278-0181, (accessed : 5 Aug 2019)

Sahin, Y. , Bulkan, S. and Duman, E.,(2013) "A cost-sensitive decision tree approach for fraud detection", *Expert Systems with Applications*, vol. 40, issue 15, pp. 5916-5923, (accessed : 29 June 2019).

Samaneh Sorournejad, Zojah, Atani , Amir Hassan Monadjemi(2016) "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective", *CoRR*,(accessed : 14 Aug 2019)

Shaza M., Abd Elrahman and Ajith Abraham, (2013) "A Review of Class Imbalance Problem," *Journal of Network and Innovative Computing*, ISSN 2160-2174, Volume 1 , pp. 332-340, (accessed : 12 June 2019).

S. P. Lloyd, (1982) "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, pp. 128–137, (accessed : 2 Aug 2019).

Sivakumar Nadarajan, Dr. Balasubramanian Ramanujam, (2016) "Fast and Effective Credit Card Fraud Detection in Imbalanced Data using Parallel Hybrid PSO" ,*International Journal of Advanced Research in Science, Engineering and Technology* ,Vol. 3, Issue 9 , ISSN: 2350-0328, (accessed : 6 July 2019)

Stephanie,(2016)Available at: https://www.statisticshowto.datasciencecentral.com/bootstrap-sample/ , (accessed : 24 June 2019).

TESTIMON, (2017) "Synthetic Financial Datasets For Fraud Detection", (accessed : 13 July 2019), Available at :https://www.kaggle.com/ntnu-testimon/paysim1

T. M. Mitchell, (1997) "Machine Learning",*New York:McGrawHill*,(accessed : 8 June 2019).

Vaishali, G., (2012) "An Overview of Classification Algorithms for Imbalanced Datasets.", *Int Journal of Emerging technology and Advanced Engineering*,(accessed : 18 Aug 2019).

Vimalraj S Spelmen; R Porkodi., (2018)"A Review on Handling Imbalanced Data" in *International Conference on Current Trends towards Converging Technologies (ICCTCT)*, (accessed : 3 Aug June 2019).

Visa, S., Ralescu, A., (2005) "Issues in mining imbalanced data sets -a review paper," in *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference* , (accessed : 24 June 2019) .

Weiss, G.M. and F. Provost, (2003) "Learning when training data are costly: The effect of class distribution on tree induction", *Journal of Artificial Intelligence Research*, Vol 19: p. 315-354, (accessed : 6 June 2019).

Y. Yong, (2012) "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm," *International Conference on Future Electrical Power and Energy Systems, Energy Procedia*, vol. 17, pp. 164 – 170, (accessed : 14 Aug 2019).

Yates, Daniel S.; David S. Moore; Daren S. Starnes , (2008) "The Practice of Statistics," *W.H. Freeman & Company*. ISBN 978-0-7167-7309-2 , (accessed : 2 June 2019)

The effect of residential segregation on social segregation:

Evidence from Flickr

Kirsten Cornelson[9]

Abstract

In this paper, I provide a new measure of inter-racial interactions in U.S. cities, and use this measure to help quantify the impact of residential segregation on social segregation. The measure is based on faces in geotagged photographs from Flickr (a social media and photosharing website), which I run through face detection and race classification algorithms to measure the frequency with which white social media users interact with black friends. An independent survey verifies that the racial breakdown of faces in online photographs corresponds well with actual social behavior. The quantification exercise is based on a theoretical model which provides an important insight: residential segregation can *increase* the number of inter-racial interactions, because it ensures that people live closer to other-race interaction partners that are more similar to them on unobserved characteristics. Efforts to racially integrate neighborhoods may therefore either increase social interactions across race (if the causal effect of neighborhoods is sufficiently strong), or decrease them (if sorting into neighborhoods has a sufficiently positive effect on interracial interactions.) To see which scenario is more plausible, I use the equilibrium conditions of my model and a lower-bound estimate of the causal impact of neighborhoods based on the disutility of travel to simulate the impact of completely desegregating U.S. cities. I find that, in this scenario, desegregation has a negative impact on inter-racial interactions. My model and results suggest that policy designed to reduce desegregation may have unintended consequences.

1 Introduction

There is a high degree of residential segregation between blacks and whites in the United States. In 2010, the average black American lived in a Census block that was 54.1% black, despite the fact that

[9] University of Notre Dame Economics Department. 3051 Jenkins-Nanovic Halls, Notre Dame IN, 46556. Email: kcornels@nd.edu.

blacks made up only 12.2% of the population as whole.[10] Even more striking, Echenique and Fryer (2007) report that, as of 2000, over 60% of Census blocks in most states contained residents of only one race.

A large literature in economics has argued that this segregation is harmful for black economic outcomes (e.g., Cutler and Glaeser, 1997; Card and Rothstein, 2007; Ananat, 2011). Understanding why this relationship holds is important for designing effective policy to combat racial inequality. Sociologists have long argued that the harmful effect of residential segregation operates, in part, through its effect on social segregation (e.g., Wilson, 1987; Massey and Denton, 1993; Krysan and Crowder, 2017). According to this perspective, residential segregation inhibits inter-racial interactions through some causal effect of neighborhoods on the probability or frequency of socializing with particular people. Segregated social networks then help perpetuate racial inequality through channels such as job referrals, social norms, or other kinds of social influence.

While there is a large literature in economics that provides support for the empirical relevance of peer effects (e.g., Duflo and Saez, 2003; Bayer, Ross and Topa, 2008; Dahl, Loken and Mogstad, 2014), there is currently little evidence evaluating the first piece of this argument - the impact of of residential segregation on social segregation. Do segregated neighborhoods lead to less inter-racial contact? If so, how important is this effect?

A key challenge in answering these questions is a lack of data on inter-racial interaction behavior. There are currently no large, publicly available datasets that contain information on inter-racial interaction rates for a broad sample of the population. Previous research on this topic has relied either on Add Health, a survey of teenagers from the 1990's (Patacchini, Picard and Zenou, 2015), or on specialized datasets such as email records from Dartmouth University students (Marmaros and Sacerdote, 2006). A first contribution of this paper is to overcome this challenge by presenting a new measure of inter-racial interactions in American cities. This measure is based on a large sample of geotagged social media photographs, which I run through face detection and race classification software. I validate this measure using survey data, which show that the racial composition of faces in

[10] Calculation by author, using 2010 Census data.

social media photographs corresponds very closely to the racial breakdown of an individuals' friends. Using this dataset, I document a substantial degree of black-white social segregation in the U.S. The typical white Flickr user in my data appears to see black friends for approximately 3.5 % of their interactions. By comparison, this number would be 13.1% in a perfectly integrated world, where interactions reflected the demographics of a user's city.

The second contribution of my paper is to provide a more rigorous theoretical exploration of the relationship between residential segregation and social segregation. I present a model of social interactions that incorporates the impact of race, location, and unobservable partner characteristics on interaction partner choices. The model provides a surprising insight: inter-racial contact may be *increase* with residential segregation. This is true even if moving two individuals closer together makes them more likely to interact, all else equal. The reason is that racial segregation arises through a process of residential sorting, which also ensures that neighbors are similar to each other along other, non-race dimensions. While a white individual may have relatively few black neighbors, the black neighbors she does have are likely to be relatively good matches for her; having these particular other-race partners close by causes her to have more black interactions. If she is induced to move into a more diverse neighborhood, she will have more other-race partners nearby (which should have a positive effect on her inter-racial interaction rate), but those potential partners will be less well-matched to her (which will have a negative effect.) As a result, it is unclear whether her inter-racial interactions will increase or decrease.

Finally, I use the model to evaluate whether segregation is likely to be a positive or negative force for increasing inter-racial contact in practice. I show that my model can be used to answer this question by simulating the inter-racial interaction rate in a world without residential sorting (and therefore, without residential segregation), and then comparing the results of this exercise to the actual inter-racial interaction rate.

The simulation exercise requires an estimate of the causal effect of location on the probability that two individuals interact. Unfortunately, a good estimate of this parameter does not exist. This is due in part to the measurement issues involved in measuring interaction behavior, and in part due to an identification problem: individuals who choose to live in segregated areas may have segregated social interactions for reasons other than the effect of neighborhoods.

While I do not have an estimate of the causal effect of distance on interaction probabilities, I argue that I can place a lower bound on this effect by focusing on a single channel through which neighborhoods may influence interaction behavior: the effect of physical distance itself. Distance should reduce the frequency of interactions between individuals who live far apart, because of the time costs involved in travel. We already have a good sense of how individuals value travel time from a large literature estimating the demand for spatially differentiated goods such as gas stations or coffee shops (e.g., Thomadsen, 2005; Davis, 2006; McManus, 2007). I use estimates from this literature in my model to conduct my simulation exercise. Because this estimate ignores other potential influences of neighborhoods on social interactions (for example, the effects of schools or other local public goods), my estimates can be thought of as a lower bound on the effect of residential desegregation on social segregation.

I find that desegregating U.S. cities would have a substantial and negative effect on the amount of inter-racial contact. In the absence of residential sorting, individuals are less well-matched to their neighbors; as a result, the total amount of social interaction (with both same-race and other-race partners) falls. This effect is particularly strong for other-race partners, leading the proportion of other-race interactions to decline as well. As a result, the absolute number of interactions between black and non-black individuals falls.

As I show in a supplementary exercise, the negative effect of desegregation on social interactions does not arise because the causal effect of physical distance on social interactions is small. In particular, I examine how the inter-racial would change in a different simulation exercise: one where individuals remain in their existing residential locations, but all factors influencing social interactions *except* physical distance are eliminated. I show that physical distance alone can account for a 15-20 % reduction in inter-racial interactions, compared to a world in which neither physical distance nor other factors are present. This is line with previous research, which shows that individuals highly value their time and avoid excess travel. Therefore, while my model does not capture all causal effects of neighborhoods, it does account for an important factor influencing interaction decisions. This causal effect is outweighed, however, by the negative effect of eliminating residential sorting.

In the next part of the paper, I outline the structural model that underlies both of my simulation exercises. This is adapted from the marriage matching model of Choo and Siow (2006). In the following section, I show how the model can be used for each of the exercises I propose. Section 4

describes the data that I require to implement the simulation exercises, including a description of the new Flickr dataset that I have developed for this purpose. Section 5 contains my results, while the last section concludes.

2 Theory

In this section, I present a discrete choice, transferable utility model of social interactions adapted from the marriage matching model of Choo and Siow (2006). The equilibrium conditions from this model will provide the basis for my simulation exercise. In this model, each agent resides at a location in a city, and makes a decision each period about whether to interact, and with whom. The partner characteristics that they choose are race and neighborhood of residence, which may affect the utility of an interaction either directly or indirectly (for example, through the correlation between neighborhood and educational attainment.) The transfer that clears the market for interactions in the model is the choice of meeting point, which affects utility because agents are assumed to dislike travel.[11]

The transferable utility assumption implies that the equilibrium frequency of interactions between any two groups of agents will depend only on the joint surplus that is created by interactions between them, and on population supplies. All else equal, the total surplus created by interactions will be declining in the physical distance between two agents, because they must jointly travel this distance in order to meet.[12] The causal effect of distance on social interactions can therefore be captured by agents' disutility of travel.

While the language I use to explain the model describes individuals' residential locations, the model can equally be interpreted as describing individuals' decisions when they start from work. Empirically, the only difference between the decision to interact when starting from work or home is that the distribution of individuals' starting points (neighborhoods) will differ: an agent's "neighborhood" will

[11] The results of the model will be similar so long as there is any way for agents to transfer utility between them (through choice of activity, who pays, etc.)

[12] In the model, I assume that agents always meet somewhere on the line in between them. In real life, agents may choose to visit locations elsewhere in the city. This can be incorporated into the model by allowing agents to have direct preferences over particular locations, which makes this an imperfectly transferable utility model in the manner of Galichon, Kominers and Weber (2016). This extension will not fundamentally change the predictions of the model, however, because the joint surplus of an interaction will still decline in the distance between any two agents' homes; this is the minimum distance that they must jointly travel.

describe their location of work rather than their location of residence.[13] While I will apply the model to the case where individuals are likely to start from home (i.e. weekend interactions), it would be straightforward to analyze the impact of integrating workplaces in a similar manner.[14]

2.1 Model setup

Agents live in neighborhoods in a city, which are spatially arranged on a line. Agents must decide whether to interact with anyone, and if so, with whom to interact. I assume that each person chooses a single interaction partner.[15] The partner characteristics that the agent chooses are race and location.[16] I refer to agent types as (*i,r*), where *i* is indexes the individual's location on the line (her neighborhood) and *r* indexes their race. I assume that there are a finite number of racial groups *R* and neighborhoods *N*. If two individuals decide to interact, they meet at a third location in the city *m*, which is somewhere on the line in between them.[17]

If an individual agent *g* of type (*i,r*) interacts with an individual *h* of type (*j,s*) at a meeting point *m*, his utility is:

$$U_{ghirjs} = \alpha_{js}^{ir} - 2\delta d(i,m) + \epsilon_{ghirjs}$$

where α_{js}^{ir} is the average utility generated for agents of type (*i,r*) from socializing with an agents of type (*j,s*); *d*(*i,m*) is the physical distance between the agent's home and the meeting point; δ is the disutility of distance[18]; and ϵ_{ghij} is an I.I.D. shock with a type I extreme value distribution. The term α_{js}^{ir}

[13] The overall level of interactions may also differ when starting from work because of "time of week" effects. For example, the individual may get less utility from socializing after work because she is tired. So long as this applies equally to all interactions, it will not affect the results of my model.

[14] In what follows, I treat interactions starting from home as being determined separately from interactions from work. In doing this, I am making use of an implicit assumption in the model: that choices over each interaction are made independently. This rules out, for example, a declining marginal utility of interactions with a given race.

[15] The model should extend to the case where people meet in groups, with a minor extension: the city must now be 2-dimensional. This is so that it is possible to find a location that perfectly compensates every group member.

[16] In practice, agents may care about a wide variety of partner characteristics, such as age or education. As I describe below, preferences over the race and location of interaction partners can be interpreted as reflecting average values of these other characteristics across different groups, as well as any race- or location- specific preferences.

[17] As I describe below, the particular location *m* that is chosen will adjust to ensure that the market clears.

[18] I model the effect of distance as entering linearly into the individual's production function. This implies that each kilometer traveled has the same effect on utility. It is possible to extend the model to capture a non-linear effect of distance, which has been found to be empirically relevant by Davis (2006). In this case, utility is not perfectly transferable through the choice of meeting point, so agents must have some other way to compensate each other (through choice of activity, who pays, etc) in order

will capture average levels of education, age, or other characteristics that affect the match value between a typical member of type (*i,r*) and a typical member of type (*j,s*). The term ϵ_{ghirjs} can be interpreted as reflecting individual partners' deviations from the average level of characteristics for members of their types; for example, ϵ_{ghirjs} may be positive if both partners have higher-than-average education for their tract and race. This term also reflects more idiosyncratic shocks, such as personality characteristics, that affect the valuation of a match.

The agent may also choose not to socialize at all, which I denote as choosing partner type "0". I normalize the intrinsic utility from spending time alone to zero.

2.2 Equilibrium

Following Choo and Siow (2006), the model described above will lead to a quasi-demand function describing the number of (*j,s*) type interactions demanded by all individuals of type (*i,r*), which depends on the meeting point *m* (interpreted as the "price" in this model). The quasi-demand function takes the form

$$ln(\mu^q_{irjs}) = ln(\mu_{ir0}) + \alpha^{ir}_{js} - 2\delta d(i,m) \qquad (1)$$

where $\mu^q{}_{irjs}$ is the total number of type (*j,s*) interactions demanded by agents of type (*i,r*) and μ_{ir0} is the equilibrium number of (*i,r*) agents who choose to spend time alone. In equilibrium, demand for these interactions by agents of type (*i,r*) must equal the "supply" of these interactions by agents of type (*j,s*):

$$ln(\mu^s_{irjs}) = ln(\mu_{0js}) + \alpha^{js}_{ir} - 2\delta d(j,m) \qquad (2)$$

to clear the market. In an earlier version of this paper (Cornelson, 2017), I show that incorporating non-linear utility produces results similar to increasing δ by approximately an order of 2. I show below how the results change when I double δ.

The meeting point *m* will adjust to ensure that this is the case.[19] Setting demand for interactions equal to supply, solving for *m*, and plugging this back into the quasi-demand equation gives:

$$\ln\left(\frac{\mu_{ijrs}}{\sqrt{\mu_{i0}\mu_{0j}}}\right) = \frac{1}{2}[\alpha_{js}^{ir} + \alpha_{ir}^{js}] - \delta d(i,j)$$

To close the model, note that there is an adding-up constraint. If we denote the population of race *r* living at *i* as $f^r(i)$, then $\mu_{ir0} + \Sigma_{t\in R}\Sigma_{k\in N}\mu_{irkt} = f^r(i)$. This lets us rewrite the equilibrium equation entirely in terms of the endogenous terms μ_{irjs} and the parameters of the model:

$$ln\left(\frac{\mu_{irjs}}{\sqrt{(f^r(i) - \Sigma_{t\in R}\Sigma_{k\in N}\mu_{irkt})(f^s(j) - \Sigma_{t\in R}\Sigma_{k\in N}\mu_{ktjs})}}\right) = \frac{1}{2}[\alpha_{js}^{ir} + \alpha_{ir}^{js}] - \delta d(i,j) \qquad (3)$$

This equation says that the frequency of interactions between types (*i,r*) and (*j,s*) (scaled by a function of the total number of partners of each type) is equal to the per-partner surplus created by interactions between these types of individuals.[20] Note that it is only the total surplus that matters here; any difference between α_{js}^{ir} and α_{ir}^{js} is irrelevant, since the agent who gets more utility from the interaction can always compensate the other partner through the choice of meeting point *m*. For the purposes of the simulation exercise, I will rewrite the term $\frac{1}{2}[\alpha_{js}^{ir} + \alpha_{ir}^{js}] = \alpha_{irjs}$, the average utility generated by the interaction.

Equation 3 highlights the role of physical distance in generating social interaction behavior. Because individuals dislike travel, we can increase or destroy the surplus from an interaction by moving agents geographically further apart or closer together, as captured by the term $-\delta d(i,j)$. This is one reason why we should expect interaction rates to change if we alter the geographic distribution of

[19] Specifically, setting Equation 1 equal to Equation 2 and solving for the distance travelled by individual *i* gives $d^*(i,m) = \frac{1}{2}d(i,j) + \frac{1}{4\delta}ln\left(\frac{\mu_{ir0}}{\mu_0}\right) + \frac{1}{4\delta}[\alpha_{js}^{ir} - \alpha_{ir}]_{js}$

[20] Note that this identical to the standard result from discrete choice problems that forms the basis for logistic regression models. The only difference is that we have extended the problem to a case where agents must coordinate

individuals within cities. In what follows, I will explore the consequences of changing this distribution, holding other components of the model - in particular, the average utility terms α_{irjs} - constant.

This equilibrium equation holds for every $(i,r),(j,s)$ pair in the city. If we let N be the number of types, this condition gives us a system of $\frac{(1+N)N}{2}$ equations. Choo and Siow (2006) show that, given values for the right-hand side of the equilibrium equation and a vector of population supplies $f^r(i),f^s(j)$ for each neighborhood, there is a unique vector of social interactions μ_{irjs} (of size $\frac{(1+N)N}{2}$) that will solve this system of equations. I use this fact in my simulation exercises, described in the following subsections of the paper.

3 Simulation exercise

In my main simulation exercise, I explore the potential for policy makers to influence inter-racial interactions by reallocating individuals' residential locations within cities. Specifically, I consider the extreme case where a policy maker is able to completely randomize individuals to neighborhoods. Using the equilibrium condition from my model, I simulate the both the overall interaction rate and the inter-racial interaction rate that would occur after this randomization. The steps of the simulation exercise, which I describe in more detail below, are as follows:

1. Decompose interaction rates into a piece explained by preferences, a piece explained by distance, and a piece explained by residential sorting

2. Randomly reassign residential locations, so that there is an even distribution of each race across tracts.

their decisions. The assumption of transferable utility permits this coordination to take place.

3. Calculate the frequency of inter-racial interactions that would take place with this new distribution of individuals, holding the preference terms estimated in Step 1 constant.

There are two important caveats to this exercise First, the "preference" terms I estimate in step 1 are actually residuals which capture all influences on social interactions other than physical distance. If there are important neighborhood effects other than physical distance (for example, through the

effect of schools on interaction behavior), it may not be appropriate to hold these terms constant when we reallocate individuals to new neighborhoods in cities. My estimates should therefore be thought of as lower bounds on the impact of residential desegregation on social segregation. Secondly, the exercise I consider is quite extreme: while it may be possible for policy makers to decrease residential segregation using tools such as housing vouchers, it is not likely that they will achieve (or want to achieve) a completely random distribution of people across neighborhoods. Despite these caveats, we will see that the exercise provides important insights on the potential unintended consequences of policies altering the spatial distribution of individuals within cities.

A second important feature of this exercise is that it can be used to analyze the impact of desegregation of either residential locations and workplaces together, or residential locations alone. In particular, if we assume that the reallocation of individuals holds at all points in time, we can interpret the model as reflecting a change that integrates workplaces as well as residences. If we instead prefer to think of an experiment that integrates residential locations only, we can assume that the level of interactions only changes during times when individuals are likely to be at home (i.e., weekends). While I will focus on the latter interpretation, it is straightforward to calculate the total effect on interactions by appropriately weighting the (unchanged) interaction behavior starting from work and the altered interaction behavior starting from home.

3.1 Step 1

The first step of my simulation procedure is to estimate the average utility created by different types of interactions. Using the equilibrium condition given in Equation 3 in the last section, and rearranging, we can write the equilibrium number of interactions between types (i,r) and (j,s) as:

$$\mu_{irjs} = e^{\alpha_{irjs} - \delta d(i,j)} \sqrt{\mu_{ir0}\mu_{0js}} \quad (4)$$

Recall here that α_{irjs} is the average utility generated by an interaction between type (i,r) and (j,s), and that μ_{ir0} is the number of partners of type (i,r) that remain unmatched in equilibrium. All of the terms on the right hand side of this equation will have observable counterparts in my data, with the exception of the preference parameters α_{irjs}. Specifically, I can calculate the distance term $-\delta d(i,j)$ using existing estimates of the parameter δ along with geographic information about the distance between pairs of neighborhoods within cities. The terms μ_{ir0} and μ_{0js} can be calculated using information on the population of each neighborhood, combined with information on the average

interaction rate. I describe where I get each of these pieces of information in the next section. Collecting this group of observable terms into a single parameter $\hat{\mu}_{irjs}$, I can rewrite the equilibrium equation as:

$$\mu_{irsj} = e^{\alpha_{irjs}} \hat{\mu}_{irjs} \quad (5)$$

Note that the observable term $\hat{\mu}_{irjs} = e^{-\delta d(i,j)} \sqrt{\mu_{ir0}\mu_{0js}}$ can be interpreted as the interaction frequency that would take place if we eliminated preferences over interaction partner characteristics ($\alpha_{irjs} = 0, \forall ir,js$).[21] The term $e^{\alpha_{irjs}}$ represents the difference between the actual number of interactions and those that occur in this imaginary, preference-less situation.

If I observed μ_{irjs}, the actual interaction frequency for every neighborhood-race pair, I would be able to use Equation 5 to calculate α_{irjs} directly. Unfortunately, however, my data do not allow me to observe the interaction rate between every neighborhood/race pair. Instead, I observe the total frequency with which an individual living in neighborhood i, of race r, interacts with individuals of race s. This is equal to:

$$\mu_{irs} = \Sigma_j \mu_{irjs} = \Sigma_j e^{\alpha_{irjs}} \hat{\mu}_{irjs} \quad (6)$$

Write the average utility generated by an interaction in the following way:

$$\alpha_{irjs} = \alpha_{rs} + \varepsilon_{irs} + \eta_{irjs} \quad (7)$$

The term α_{rs} represents the average utility created by an interaction between someone of race r and someone of race s within the city. The term ε_{irs} represents the deviation from this average for individuals in neighborhood i (for all j). This term would be negative, for example, if race$r$ individuals in neighborhood i were unusually racist against individuals of race s. The terms η_{irjs} represents the additional deviation when someone of race r from neighborhood i interacts with someone of race s from neighborhood j. This will reflect any specific compatibility between these two groups. Note that both the terms ε_{irs} and η_{irjs} arise because of residential sorting, a process that creates systematic differences in observable and unobservable characteristics between individuals living in different

[21] Note that the terms $\hat{\mu}_{irjs}$ are *not* equilibrium outcomes. If we eliminated preferences over partner characteristics, the overall interaction rate (and therefore the terms μ_{ir0}) would change, which is not captured in my construction of $\hat{\mu}_{irjs}$. We can instead think of $\hat{\mu}_{irjs}$ as the initial non-equilibrium result of eliminating preferences.

neighborhoods. To the extent that these characteristics affect the utility from a match, we expect that the term α_{irjs} will vary depending on the location of the two interaction partners (i.e. that both ε_{irs} and η_{irjs} will typically not be zero.) I assume that both ε_{irs} and η_{irjs} are normally distributed, although not in a way that is independent of residential location.

Using this equation in the equilibrium condition and taking logs gives:

$$ln(\mu_{irs}) = \alpha_{rs} + \varepsilon_{irs} + ln(\Sigma_j e^{\eta_{irjs}} \hat{\mu}_{irjs}) \quad (8)$$

Without loss of generality, rewrite the summation in the logarithm term on the right hand side, so that the equation becomes:

$$ln(\mu_{irs}) = \alpha_{rs} + \varepsilon_{irs} + ln(\Sigma_j \hat{\mu}_{irjs} + D) \quad (9)$$

The term D here is simply the difference between $\Sigma_j e^{\eta_{irjs}} \hat{\mu}_{irjs}$ and $\Sigma_j \hat{\mu}_{irjs}$; it captures additional interactions between members of (i,r) that take place with members of race s because of preferences and the associated residential sorting.[22]

Equation 9 guides my estimation of the average racial preference parameters α^{rs}. For this estimation, I start with data on the residential location and inter-racial interaction frequency for a set of individuals. For each individual h in my dataset, I construct the terms μ_{hirs} and $\Sigma_j \hat{\mu}_{irjs}$ and run the following non-linear regression across individuals within a city c, separately for each partner race s:

$$ln(\mu_{hirs}) = \alpha_{rs}^{c} + ln(\Sigma_j \hat{\mu}_{irjs} + D_{rs}^{c}) + \varepsilon_{hirs} \quad (10)$$

This regression relates the interaction frequency for individual h, of race r, living in neighborhood i, with members of race s, to the predicted interaction frequency if we eliminated preferences over interaction partners ($\Sigma_j \hat{\mu}_{irjs}$). The term α_{rs} captures preferences over the average person of race s in the

[22] To see this, note that in expectation, D will be equal to $\overline{\hat{\mu}_{irjs}}(e^{\frac{1}{2}\theta^2} - 1) + cov(\hat{\mu}_{irsj}, \eta_{irsj})$, where θ is the variance of η_{irsj}. θ captures the degree to which preferences for partners vary strongly across neighborhoods, while $cov(\hat{\mu}_{irsj}, \eta_{irsj})$ captures the extent to which individuals prefer their near neighbors. Both of these terms will tend to be more positive in the presence of residential sorting, causing $\Sigma_j e^{\eta_{irjs}} \hat{\mu}_{irjs}$ to exceed $\Sigma_j \hat{\mu}_{irjs}$.

city. The term $\Sigma_j \hat{\mu}_{irjs}$, as discussed earlier, is the predicted frequency of *s*-interactions that would take place based on the effect of physical distance and population supplies alone. The term *D* reflects the deviation from this level of inter-racial interactions that occurs because individuals have preferences over non-race characteristics that may vary by neighborhood. In other words, even without direct preferences over race ($\alpha_{rs} = 0$), we would expect μ_{irs} to differ from $\Sigma \hat{\mu}_{irjs}$. This is because individuals can sort into neighborhoods where they like the particular partners of race *s* relatively better; this will tend to push the inter-racial interaction frequency above the level predicted by $\hat{\mu}_{irjs}$ alone. The entire term $\Sigma_j \hat{\mu}_{irjs} + D$ captures the inter-racial interaction frequency we would predict based on this process. The non-linear estimation process chooses α_{rs} and *D* to minimize the sum of squared error terms ε_{hirs}.

The constant term α_{rs} from this regression, which I interpret as the average utility created by an *r,s* interaction, captures the extent to which the interaction frequency between *r* and *s* in the city deviates from the level predicted by the term within the logarithm, while ε_{hirs} (the error term in the regression) captures the individual's deviation from this average. With sufficient data, I could break ε_{hirs} into a neighborhood by race fixed effect (which I would interpret as ε_{irs} from the model) and an individual error term. In practice, I will often only have one observation per neighborhood, which precludes the use of neighborhood-specific fixed effects. Note, however, that the neighborhood-specific variation will not be important when I move to the next steps of my

simulation exercise.

It is important to emphasize that the goal of this regression is not to estimate any causal relationship. The causal effect of physical distance in my model is already incorporated in the term $\delta d(i,j)$, which appears in $\hat{\mu}_{irjs}$. Instead, this regression is intended to decompose the frequency of inter-racial interactions into a piece that can be explained by this causal effect and into a residual that I interpret as reflecting racial preferences. Of course, the term $\hat{\mu}_{irjs}$ captures only one particular channel through which neighborhoods influence interactions. If this is true, deviations from the predicted level of inter-racial interactions may result from some of these other causal channels. This implies that my "preference" terms $\hat{\alpha}_{rs}$ may actually overstate the extent of racial preferences.

This means that my estimates should be interpreted as lower bounds on the impact of residential desegregation.

3.2 Step 2

The goal of this simulation exercise is simulate the frequency of inter-racial interactions that would take place if we eliminated residential segregation completely. The specific counter-factual I use is one where individuals of different races are evenly distributed throughout the city, with each individual being randomly assigned to a specific neighborhood.

To achieve this distribution of individuals, I assume that every neighborhood has the same total population as it does currently, but that the fraction of individuals of race *r* is equal to the proportion of these individuals in the *city* population for every tract. Random assignment of specific individuals implies that each neighborhood will, on average, contain a representative sample of individuals of each race.[23] I use this fact in the derivation of my simulation results below.

3.3 Step 3

The third step of my exercise is to calculate the inter-racial interaction frequencies that would take place if individuals had the same preferences estimated in the first step of the exercise, but had the geographic distribution simulated in the second step. To do this, I use the equilibrium equation from my model. Letting *N* denote the set of neighborhoods in a city, this is:

$$ln\left(\frac{\mu_{irjs}}{\sqrt{(f^{*r}(i)-\Sigma_{t\in R}\Sigma_{k\in N}\mu_{irkt})(f^{*s}(j)-\Sigma_{t\in R}\Sigma_{k\in N}\mu_{ktjs})}}\right)=\hat{\alpha}_{rs}-\delta d(i,j)$$

In this equation, $f^{*r}(i)$ is the new number of individuals living in neighborhood *i* of race *r*, after the redistribution. The preference term on the right hand side is the average match value for an *r,s* pair, as estimated in step 1 of the simulation. Importantly, the match values do not vary systematically across neighborhood pairs the way they do in real life. That is, the terms ε_{irs} and η_{irjs} will both be equal to zero. This is because we have eliminated any sorting into neighborhoods on the basis of observable or unobservable characteristics. The average value of a match between a person of race *r* in neighborhood

[23] In practice, random assignment will not result in an exactly even distribution of interaction-relevant characteristics across neighborhoods. It is possible for my model to accommodate this random variation. To do this, I can allow my estimated α_{rs} to vary with the observable characteristics of each individual (e.g. education, age) and the average level of these characteristics among individuals of race *s* in the city; this provides an approximation of how interaction values are affected by these characteristics. Then, I could randomly re-assign individuals and construct a predicted α_{ijrs} for each neighborhood-race pair, based on the characteristics of each neighborhood. This procedure should not affect my results on average, however; as a result, I do not implement it here.

i and a person of race *s* in neighborhood *j* will be the same, no matter which *i* and *j* we choose.[24] Using my estimates $\hat{\alpha}_{rs}$, along with calculations of $\delta d(i,j)$ and the simulated population supplies $f^*()$, I can solve this system of equations for μ_{irjs}. These can then be aggregated into predictions about both the overall interaction rate, and the inter-racial interaction rate.

3.4 Supplementary exercise

In a secondary exercise, I attempt to provide an intuitive sense of how important my estimated disutility of travel terms δ are likely to be in influencing individuals' interaction decisions. To do this, I use a version of the model that assumes i) that individuals remain at their existing residential locations, but ii) that physical distance is the *only* factor affecting decisions about whether to interact and with whom. Specifically, the equilibrium condition I use is:

$$ln\left(\frac{\mu_{ij}}{\sqrt{(f(i)-\Sigma_k\mu_{ik})(f(j)-\Sigma_l\mu_{lj})}}\right) = -\delta d(i,j) \quad (11)$$

A key difference between this equation and Equation 3 is that the equation no longer has the subscripts *r* and *s* indicating the race of the interaction partners. If individuals care only about physical distance, then they should draw randomly from the population living in each neighborhood.

Nonetheless, there will be some social segregation induced by individuals' behavior in this model. This is because black and non-black people are not randomly distributed across cities; non-black people will tend to have more non-black people in nearby neighborhoods, and conversely for black people.

This equation says that the log number of interactions between neighborhoods *i* and *j*, once scaled to reflect population supplies, should be fall with the distance between the two neighborhoods, at a rate equal to the disutility of travel. As noted above, given a value for δ and $d(i,j)$ for every neighborhood pair as well as a vector of population supplies $f()$, I can solve this equation for the unique vector μ_{ij} that satisfies this set of equations.

[24] Of course, the match value of specific individuals in these neighborhoods may be higher or lower than this; this is captured by the error terms ϵ_{hgirjs} in each individual's utility function.

To see how the equilibrium terms μ_{ij} are aggregated into inter-racial interactions, consider interaction behavior for individuals living in neighborhood *i* (of any race). The number of interactions between people in *i* and people of race *s* in neighborhood *j* will be equal to:

$$\mu_{ijs} = \frac{f^s(j)}{\Sigma_z f^z(j)} \mu_{ij}$$

That is, the proportion of individuals of race *s* make up a proportion *p* of the total population in a neighborhood, then *p* will also represent the proportion of μ_{ij} interactions that involve with a partner of race *s* in neighborhood *j*. By adding this up over all neighborhoods *j*, we can calculate the total number of interactions that occur between individuals in neighborhood *i* and individuals of race *s*. If we divide this by the total number of interactions for individuals in neighborhood *i*, we have the proportion of interactions for neighborhood *i* that occur with race *s*. Note that this will be the same for individuals of *any* race who live in *i*; individuals who live in the same neighborhood will behave the same way according to this model. However, because of residential segregation, individuals of race *r* will tend to be heavily represented in neighborhoods that are physically close to large numbers of race *r* partners. Individuals living in these neighborhoods will therefore have a disproportionately high number of *r* partners, compared to the city population. This is what will cause social segregation in the model.

We can get a sense of how important the distaste for travel is by comparing the the inter-racial interaction rate generated by this model to two objects. The first is the random interaction rate. This is the rate that would occur if individuals matched randomly within cities. For example, for a typical non-black person in a U.S. city, the random black interaction rate would be around 13% the population frequency of black people in an average U.S. city. If physical distance alone is able to generate a significant reductions in inter-racial interactions relative to the random rate, then we can infer that the distaste for travel is "large" in an absolute sense.

The second object with which we may wish to compare the results of the simulation is an index of existing social segregation. The index I use is simply the absolute difference between the random inter-racial interaction rate and the actual inter-racial interaction rate. As I show below, the typical non-black person in my data appears to have about 3% of their interactions with black people. With random matching, this would be around 13%, meaning that there is approximately a 10 percentage point gap between how individuals would behave in a perfectly integrated world (i.e., no racial preferences or

physical segregation) and how they actually behave. In a similar way, I can construct a measure of how much social segregation there would be in the simulation exercise I describe above.

By comparing simulated to actual social segregation, we can get a sense of how large the distaste for travel is, *relative* to other factors driving social segregation.

4 Data

To perform my simulation, I need four pieces of information. First, I need to know the existing population of each neighborhood, $f^r(l_i)$, separately by race. Secondly, I need to know the geographic distance between neighborhoods within a city; this corresponds to $d(l_i,l_j)$ in my model. Third, I need an estimate of the disutility of travel δ. Finally, I need information on the social interaction behavior of Americans: both how often individuals in different neighborhoods socialize (which, combined with information on population will allow me to estimate the terms μ_{ir0} from my model), and how frequently they socialize with members of different races (which corresponds to the term μ_{irs} in my model). In this section, I describe where I get each of these piece of information.

4.1 Population distribution

Information on the population distribution by neighborhood and geographic distances are available from the U.S. Census Bureau. Throughout the analysis, I will define a neighborhood as a Census tract. The average population in a Census tract in the U.S. is around 4,000 individuals. I restrict the analysis to pairs of Census tracts within the same Core-Based Statistical Area (CBSA.)[16] There are 933 CBSAs in the United States (excluding Puerto Rico), of which I use 202 in my main analysis.[17] These CBSAs account for just over 80% of the U.S. population. The mean number of Census tracts one of these CBSAs is 271, ranging from 31 to 4,701.

Different measures have been used to capture the degree of racial segregation within cities. One popular measure is the Duncan index, which measures the fraction of black or non-black residents within a city that would have to move to produce an even distribution of racial groups over Census tracts.[18] The first column of Table 1 shows that the mean Duncan index in my sample is 0.521 , indicating that about half of all residents in a typical city would have to move to achieve perfect

[16]CBSAs consist of "one or more counties and includes the counties containing the core urban area, as well as any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core." (Census Bureau, 2016.)

[17]The restrictions that lead to the reduced sample of CBSAs are that I must have an estimate of the disutility of travel, which requires information on both travel speeds and hourly wages; and that I must have at Flickr users represented in at least 25 tracts. The latter restriction results in more eliminations from my pool of CBSAs, but is required in order to implement the regressions in Step 1 of Simulation B.

[18]The Duncan is calculated using the formula $D_c = \Sigma_t \left| \frac{\text{Black}_{tc}}{\text{Black}_c} - \frac{\text{Nonblack}_{tc}}{\text{Nonblack}_c} \right|$ where Black_{tc} is the number of black individuals living in tract *t* in city *c*, and Black_c is the total number of black individuals in the city (and similarly for non-black individuals).

integration. The next rows show how the Duncan index varies across the four Census regions. According to this measure, segregation is highest in the Midwest, with an average Duncan index of 0.596, and lowest in the Pacific, with an average Duncan index of 0.457.

Echenique and Fryer (2007) construct an alternative measure of residential segregation that is intended to closely capture the probability of social interactions across different Census blocks. This measure is calculated based on information on the black population of each Census block, the black population of neighboring Census blocks, and those neighbors' degree of segregation, implying that the entire distribution of the city population is incorporated into the final index. The authors show that this index satisfies several desirable properties, including that it is invariant to how the boundaries of neighborhoods are drawn (unlike, for example, the Duncan index.) Their measure of black segregation at the MSA level is provided on the authors' websites.[25] Column (2) of Table 1 shows the mean level of this index (which also varies between 0 and 1) for the CBSAs in my sample, and how it varies across Census regions. Throughout the paper, I refer to this index as the SSI. While the mean level of the SSI is similar to the Duncan at 0.577, this index tells a markedly different story across regions. It shows that there is a much more substantial degree of black segregation in the South and Midwest (with both indices around 0.7) than in the Northeast (around 0.5), and a much lower degree in the Pacific region (around 0.25). As I describe in the results section, the results from my secondary simulation exercise coincide much more closely with the SSI than with the pattern shown in the Duncan index.

4.2 Distance

[25] Specifically, the data were available from from http://www.its.caltech.edu/ fede/segregation/ as of February 2019.

To measure the geographic distance between pairs of Census tracts, I use shapefiles provided by the U.S. Census Bureau. I calculate the great-circle distance between the central latitude and longitude of each pair of tracts within a CBSA. The third column of Table 1 shows that, on average, two randomly selected tracts within the same CBSA are 41.1 km apart. This varies from 37.5 km in the Pacific region to 44.9 km in the Northeast.

Table 2 shows the mean distance to an average black and non-black person within the same CBSA, for black and non-black individuals separately.[26] This is somewhat lower than the average distance between tracts (not population weighted), with estimates ranging between 23 and 28 km.

Somewhat surprisingly, the average non-black person lives *closer* to the average black person than they do to other non-black people. This can occur when the black population is concentrated in the city center, which is a pattern that holds in many U.S. cities. Note, however, that this does not imply that distance doesn't play a role in generating social segregation. While I model the disutility of distance as being linear, my model can produce highly non-linear relationships between distance and the probability of interactions (even without any interaction-specific preferences.) In this case, it will not be the distance to the average black person that matters for interaction behavior, but the frequency of black people in a person's immediate neighborhood.

4.3 Disutility of travel

A number of papers have estimated individuals' disutility of travel in the context of estimating demand for movie theatres (Davis, 2006; Thomadsen, 2005), liquor stores (Seim and Waldfogel, 2013), coffee shops (McManus, 2007), and gas stations ((Manuszak and Moul, 2009; Houde, 2012). The typical strategy of these papers is to examine how much consumers are willing to pay, in terms of price, to avoid extra travel to a location that is further away. A key assumption for identifying the distaste for travel in this way is that consumers otherwise value the competing locations similarly; that is, that there is no correlation between a location's distance from the consumer and its unobservable characteristics. In some cases, such as for gas stations near a consumer's commute path, this seems

[26] Throughout the paper, I use the racial categories "black" and "non-black". This is driven my Flickr photographs data, in which I can distinguish black from non-black people but not different categories of non-black people.

reasonable. In other cases where the assumption is more tenuous, a variety of instruments have been used to try and causally identify the effect of distance.[27]

Table 3 summarizes the findings of these papers. The estimated willingness to pay to avoid a minute of travel varies quite substantially in this literature, both in absolute magnitude (ranging from about $0.10-$0.57 per minute in 2002 dollars) and in relation to average hourly wages (with the hourly valuation ranging from 0.5-2.5 times the average hourly wage.) Broadly speaking, however, the results can be grouped into two sets: one set implying a valuation of time at about the average hourly wage (Davis, 2006; McManus, 2007; Manuszak and Moul, 2009), and another set implying a time valuation of about twice the average hourly wage (Thomadsen, 2005; Houde, 2012; Seim and Waldfogel, 2013). I show estimates using both of these values in my simulation results.

To construct an estimate of the disutility of travel for each city in my sample, I start by estimating the hourly wage for each city. Information on hourly wages is available for some metropolitan areas from the Bureau of Labor Statistics. In order to preserve the majority of CBSAs in my sample, however, I instead impute average hourly wages by using information on state-level hourly wages and the ratio of median income in the CBSA to median income in the state. For each city, I construct a "per minute" disutility of travel equal to either one or two times the wage per minute in that city.

Next, I convert the dollar valuation of time to utility terms using the estimates from Houde (2012), which are available in both units of measure. This gives me an estimate of disutility per minute, which I then convert into "per kilometer" format by using information on travel speeds from the Google Maps API.[28] I have sufficient information on income and travel speeds to calculate the disutility of travel for 862 CBSAs, which include all 202 CBSAs in my main analysis sample. As shown in the fourth column of Table 1, the mean disutility of travel across cities is around 0.458 per kilometer, which corresponds to a dollar valuation of around $0.46 per kilometer in 2010 dollars. The table also shows how the

[27] For example, Manuszak and Moul (2009) use a tax hike near Cook County to estimate consumer's willingness to pay to travel across county lines to purchase gasoline.

[28] Specifically, I choose 10 randomly selected pairs of Census blocks within a CBSA and query the API for a driving time between them on a Saturday afternoon at 3 pm. In my main results, I use Flickr photos taken on weekends to measure social interactions, in order to capture time periods when individuals are likely to be leaving from home.

disutility of travel varies by region. The cost of traveling 1 km is highest in the Northeast and Pacific, and lowest in the South.

A limitation of this procedure is that variation across cities is imposed by assumption, not by revealed behavior. We can get some sense of whether the implied distaste for travel actually corresponds to individuals' travel behavior by using the travel patterns in my Flickr data. As I explain in the next section, the main purpose of my Flickr data is to measure individuals' crossracial interactions. Because the photos are geotagged, however, they also provide some information about how individuals move throughout their home cities. Table 4 shows the relationship between my predicted disutility of travel at the city level and the fraction of photographs that are taken within 1, 3, 5 and 10 km of a Flickr user's estimated home location.[29] The table shows that cities with a higher estimated distaste for travel have a higher proportion of photos taken close to home. For example, the coefficient on δ for photos taken within 1 km of home is 0.158 and is significant at the 10% level. This implies that 1-standard deviation increase in δ (approximately 0.105) is associated with a 1.7 percentage point increase in the proportion of photos taken within 1 km of home, a 10% increase over the baseline of approximately 15%. The coefficients for photos taken within 3, 5, or 10 km of home are also positive, and are significant at the 1-5% level. [30]

4.4 Social interactions

My simulation exercise requires that I have two pieces of information about social interactions: the overall interaction rate (which, along with population estimates, allows me to calculate the terms μ_{ir0} in my model), and the frequency of inter-racial interactions (the term μ_{irs}), by race and by neighborhood. Unfortunately, this information is not available in any large, publicly available dataset. The data used in earlier research on social interactions includes the Add Health dataset

(a survey of teenagers; e.g., Echenique and Fryer, 2007), the Social Capital Community Benchmark Survey (a survey of individuals living in cities that asks respondents how often they participate in different social activities; e.g., Brueckner and Largey, 2008) and the DDB Needham Lifestyle Survey (a

[29] I describe how I infer users' home locations, and provide evidence that I am correctly identifying these locations, in the next section.

[30] In principle, disutility of travel may vary at the tract level. This may occur because individuals value their time

survey that asks similar questions as the SCCBS; e.g., Glaeser and Gottlieb, 2006). Of these, only the Add Health has information on either residential location or cross-racial interactions; however, both pieces of information are available only for a relatively small subsample of respondents.[25] Additionally, this data source only allows us to measure interaction behavior for a very specific group (teenagers), nearly two decades ago.

To measure interaction behavior, I instead rely on a combination of the American Time Use Survey (ATUS) and a novel dataset I have constructed using Flickr photographs. The ATUS is an annual survey of a representative sample of Americans that asks respondents to keep a diary recording what they are doing and who they are with for every 15 minute segment of the day. This can be used to construct an estimate of interaction rates. Unfortunately, the ATUS does not contain geographic information below the state level. To arrive at ATUS estimates for tracts, I use the demographic information present in the ATUS to predict the interaction rate based on tract demographics. I show that this measure predicts the interaction rate well in an independent survey.

While the ATUS is informative about overall interaction behavior, it contains no information about the inter-racial interaction rate. To measure inter-racial interaction behavior, I turn to my Flickr dataset. Flickr is a popular photo-sharing website. As of 2013, the site had around 87 million users uploading approximately 3.5 million photos per day (Jeffries, 2013). I downloaded a large sample of public photographs on Flickr, along with their metadata, and ran them through face

differently, or because travel speeds differ across tracts. In an earlier version of this paper (Cornelson, 2017), I attempt to estimate tract-level disutilities of travel using information on travel patterns from Flickr. While these parameters did a better job of predicting Flickr users' travel behavior, they made essentially no difference to the interaction results. For this reason, I maintain the assumption of a city-level disutility of travel here.

[25] Patacchini, Picard and Zenou (2015) examine the relationship between physical distance and the probability of friendship in the Add Health data, using a sample of about 1500 respondents that have sufficient information on both residential location and social interactions.

detection and race classification algorithms. The racial breakdown of faces in the photographs will provide me with information on individuals' inter-racial interaction rates. The metadata, which contain information such as the make of the camera and the time of the photograph, sometimes include "geotags", which are latitude and longitude coordinates appended by cameras that have access to the internet (smart phones, for example, and higher-end digital cameras.) These geotags allow me to link Flickr users to cities and neighborhoods.

To validate the use of Flickr data to measure inter-racial interaction behavior, I also ran a survey of MTurk workers about their use of social media and their interaction behavior. I show that the racial breakdown of faces in an individual's photographs is an excellent indicator of the racial breakdown of that person's friends. In fact, the faces in a *single* photograph explains approximately 39% of the variation in actual interaction behavior, with the fraction of black faces demonstrating an almost one-for-one relationship with the fraction of the individual's friends that are black.

In the remainder of this section, I provide more information on how I construct measures of interaction behavior from these sources of data, including the construction of the Flickr dataset and the MTurk survey.

4.4.1 Interaction rates

I use the American Time Use Survey (ATUS) to arrive at my estimates of the overall interaction rate for each tract/race pair. My measure of interaction rates in the ATUS will be the probability that a respondent spends any time with friends on his or her diary day. This measure corresponds closely to the decision margin in my theoretical model. The mean of this variable is 22.7% for both races, indicating that about one-fifth of the population spends time with friends on a randomly selected day. This varies somewhat across different times of the week, with an average of 20.6 % on weekdays and 24.5% on weekends. Because I will be focusing on the impact of desegregating residential locations, which are more likely to impact social interactions on weekends, I will focus on weekend socializing in both the ATUS and Flickr results.

Table 5 shows how this probability of weekend socializing varies with age, education, and geographic region, separately for black and non-black respondents. For both racial groups, the interaction rate shows a U-shape in age; the coefficients on age and age squared indicate that interactions decline with age until approximately age 58-61, before starting to rise again. For the non-black sample, the interaction rate shows an approximately linear and increasing pattern in education. For the black sample, interaction rates are similar for all educational groups except for those with postgraduate degrees, who have higher interaction rates. There are no regional differences in interaction behavior for blacks, and only one marginally significant different for whites (with slightly higher interaction rates in the Midwest, compared to other regions of the country.)

I use the results from the model shown in Table 5 to predict the interaction rate for each Census tract based on tract demographics, separately for the black and non-black samples. My predicted interaction rate varies from 18.5%-50.1% for blacks (with a mean of 25.3%), and from 13.7% to 65.2% for whites (with a mean of 24.5%). Because the interaction rate is based on demographic characteristics available in the American Community Survey, I am able to predict it for nearly all of the approximately 70,000 tracts in the U.S.

To validate the use of my predicted interaction rates, I compare the ATUS predictions of interaction rates to actual interaction rates in a survey of MTurk workers. The survey was administered to approximately 1,600 MTurk workers in the summer of 2018. To be included in the survey, a worker had to be living in the United States, and must have posted at least 1 photograph to social media over the past year. This latter restriction was imposed because I will also use the survey to validate the use of my Flickr measure of inter-racial interactions. The survey contained modules asking the workers about basic demographic information; their time use over the previous day, including information on who was present at each moment and the race of those individuals; and their social media use. The demographic module was always presented first, while the other three modules were presented in random order.

Table 6 shows demographic information on the MTurk sample, compared to the U.S. adult population. As might be expected, MTurk workers are not representative of the U.S. population at large: they are significantly younger, more highly educated, and more likely to be white or Asian than other U.S. adults. This is in line with previous work that examines the characteristics of MTurk workers (Berinsky, Huber and Lenz, 2012; Huff and Tingley, 2015). To the extent that these characteristics are reflected in the MTurk workers' residential locations, however, the predicted interaction rates should still be valid for this sample.

I construct a measure of the social interaction rate in my sample using the time use portion of the survey. This module asked respondents what their primary activity was during each 3 hour period of the previous day, and who was with them during that time. I constructed the questions in the time use module to be as similar as possible to the American Time Use Survey questions, using similar question wording, and the same breakdown of activities present in the ATUS lexicon. As in the ATUS, I measure the interaction rate by constructing an indicator for whether a respondent spent any time with friends

on the day in question. 25.8% of the workers in my survey spent time with friends the previous day, which is quite similar to the ATUS mean.[31]

MTurk surveys automatically include information on the respondents' latitude and longitude while taking the survey. Assuming that this location will typically be the user's home, I use these locations to connect respondents to a predicted tract-level interaction rate constructed from the ATUS data. Then, I examine whether the ATUS prediction corresponds well with their actual interaction behavior. Unfortunately, I have too few black respondents (113) to be able to examine the relationship between these variables in the black sample; I therefore restrict myself to the nonblack sample in this exercise.

A regression of the MTurk worker's interaction indicator on his or her predicted interaction rate produces a coefficient of 0.787, which is significant at the 1% level. This coefficient is not statistically distinguishable from 1, which is the coefficient I would expect in this regression. The constant is 0.046 and is not significantly different from 0. When combined with the fact that there is a great deal of error in my "home" assignment processes (which should bias the coefficient in this regression downward), this regression suggests that the estimated interaction rates do a good job of predicting actual interaction behavior.

4.4.2 Inter-racial interactions

To measure the inter-racial interaction rate, I rely on a new dataset I have created using Flickr photographs. In this dataset, I observe the racial breakdown of the individuals in a Flickr user's photos; I use this as my measure of the relative frequency of black interactions. I show below that this measure is quite strongly correlated with actual inter-racial interaction behavior in my MTurk survey. The Flickr photographs also contain geotags, which allows me to infer a home location for each user. While there will doubtless be a great deal of error in the assignment process, I provide evidence below that I am identifying the home neighborhood correctly on average.

To build this dataset, I began by identifying a set of around 170 million geotagged Flickr photographs, all taken within the U.S. between 2006-2015. To do this, I started by pulling a random sample of about 10% of all geotagged photographs taken in the U.S. over this period. Then, I pulled

[31] The survey was run on a Sunday-Tuesday, with the majority of respondents answering on a Monday. Note that the questions referred to the previous day, which means I am capturing interaction behavior primarily on Sundays.

every photograph ever taken by the approximately 365,000 users in this initial sample. In order to remain in the sample, a Flickr user had to i) take the majority of their photographs in the U.S., ii) post photos taken on at least 3 separate days within a single year, for at least one year, and iii) have at least one face in the sample of photos that I use. The second restriction is required in order to infer a home location for each user. The third restriction is required in order to infer something about the user's inter-racial interaction behavior.

I link users to home locations by assigning them to the modal CBSA in which they take pictures, and to their modal Census tract within the CBSA. I do this separately for every year in which a Flickr user posts photographs. In order to abstract from potential moves by Flickr users (some of which may be induced by error in the home assignment process), I assign Flickr users consistently to the home location from the year on which the user posted on the maximal number of days. I also keep only photographs that are taken in a user's home city. I additionally restrict analysis to the 202 CBSAs that contain at least 1 Flickr user in at least 25 separate tracts. This restriction is required in order to run the regression in Step 1 of Simulation B, which identifies the preference parameters α using cross-tract variation. My final sample of Flickr users is comprised of around 87,000 users who post around 18 million photographs in their home cities.

Appendix Tables A1, A2 and A3 provide evidence that I have correctly identified users' home locations. Table A1 shows that the home tracts are visited far more often than any other tract. The table shows the number of unique "visits" (day by tract level observations) to the home location and to other Census tracts the user visits. The average user is observed in her assigned home Census tract on 14 separate days; for any other Census tract that the user visits at least once, the mean number of visits is around 3. For a typical Flickr user, about 43% of all visits are in the home tract.

In Table A2, I show that the surroundings in the home tract are observably different from other tracts the user visits. The table shows the types of venues that appear in the home location and in other visited tracts, using information from the Foursquare database. Foursquare is a service that allows individuals to "check-in" at different locations, providing information to friends and family about where they are. Foursquare maintains a database of venues, which is searchable by latitude and longitude. For a sample of around 20,000 owners, I randomly select one photograph taken in their home tract and one photograph taken outside of their home tract, and search the

Foursquare database for venues within a 25 meter radius around the location where the photograph was taken.[32] I divide venues into five categories: food and drink (e.g., restaurants, bars, and coffee shops), entertainment (e.g., parks, movie theaters, and art galleries), stores, offices, and "other". The latter category is mainly comprised of other commercial buildings that are not designated specifically as office buildings; among the most common types of venues in this category are banks, doctor's offices, and barbers/salons. I compare the number of venues I find of each type when the user is in his or her assigned home tract and when her or she is elsewhere. There are fewer venues of all type near the user when she is in her home tract, and the difference is highly significant for four of the five venue types. When combined with the results for visits, these results show that Flickr users spend substantially more time in their "home" tracts, even though there are fewer commercial venues to visit in these areas.

Finally, in Table A3, I use the one piece of information I have on Flickr users - their names - to examine how city and tract demographics correlate with the user's demographics. For users with last names on their profiles, I use information on the racial distribution of the 1000 most common last names in 2010 (Census Bureau, 2010) to construct a probability that the user is white, black, Hispanic or Asian. I then examine how this probability predicts the proportion of people that are of the same race in a user's CBSA; in the other tracts she visits (aside from the home tract); and in her home tract. Table A3 shows the results for each race in separate panels. The table shows that home tract demographics are more strongly correlated with a user's race than the demographics of other tracts she visits, or than the demographics of the city as a whole. Increasing the probability that a user is white by 1 percentage point increases the proportion white in her CBSA by 0.073 percentage points; the proportion white in visited tracts by 0.099 percentage points; and the proportion white in her assigned home tract by 0.107 percentage points. The results are similar for other races. This provides further evidence that user's assigned home tracts are likely to be strongly correlated with their actual neighborhood of residence.

The information in Table A3 suggests that we can use the characteristics of a Flickr user's assigned home tract as a proxy for their own characteristics. This provides a way to examine the demographic characteristics of the Flickr sample and how these differ from those of the typical American. Table 7 shows the average characteristics of Flickr users' CBSAs/tracts, and compares these to those of an average population living in one of the CBSAs in my sample, and to the population of the entire

U.S. The Flickr users are more concentrated in larger cities, with an average city size of around 5.3 million, compared to 4.7 million for a typical resident of the same cities. This implies that Flickr users

[32] The Foursquare API maintains rate limits which limit the number of searches to their database each day. This is why I use a smaller sample of owners and photographs.

are disproportionately concentrated within the larger cities in my sample. My sample cities are much larger than those inhabited by an average American, which have an average size of 3.8 million. City-level segregation is quite similar for Flickr users and other Americans. However, the two groups live in different areas within cities: Flickr users are concentrated in wealthier and more educated tracts, compared to both other residents of the same cities and the U.S. population in general. Flickr users have a slightly lower proportion of whites, Blacks and Hispanics in their tracts than the typical American (particularly for the latter two ethnic groups) but a higher proportion of Asians. While Flickr users are not representative of the U.S. population, I will attempt to use information from the Flickr sample to predict interaction rates outside of the sample. This process is described in detail below, after I explain my measures of social interactions.

Once I have linked users to home locations, I measure their inter-racial interactions by running their photographs through face detection and race classification software. The face detection algorithm was provided by MIT Information Extraction. Kazemi and Sullivan (2014) report that it has a 95% accuracy rate, with most of the error accounted for by false negatives. However, the rate of false negatives appears to be substantially higher in the Flickr data. In a sample of around 17,500 photographs which I hand-coded, I found 13,560 faces in total, while the algorithm found just 6,861. The lower accuracy rate may be due to the fact that the Flickr photographs are often of relatively low quality, and include many faces that are partially turned away from the camera. The rate of false positives also appears to be elevated relative to that reported in Kazemi and Sullivan (2014), at around 8%; this is due primarily to the fact that Flickr users often take pictures of statues, art with faces, and animal faces. Other non-face objects are rarely identified as faces. 18.6% of the photos in my database are found by the algorithm to have any faces in them. Based on the error rates cited above, I estimate that the actual frequency is about twice as large, in the 35-40% range.

I next run all photographs with faces in them through a race classification algorithm, which classified each face as either black or non-black. The algorithm itself was provided as part of the Scikit Learn machine learning module for Python. I trained the classifier using the faces in a random sample of 20,000 Flickr photographs. Table 8 shows the "confusion matrix" from the testing process, which shows the fraction of non-black/black faces that are categorized as non-black or black. Nonblack faces are correctly categorized 87% of the time, while black faces are correctly categorized 75% of the time. These error rates are in line with the standards in the literature for this type of classification task (Han

and Jain, 2014).)[33] While the classifier is correct in the large majority of cases, the error in the assignment process does create a problem for estimating the fraction of black faces in the sample. This is because most of the faces in Flickr photographs are non-black: in a sample of hand-coded photographs, I estimated the fraction of black faces to be around 5%. As a result, the 13% of non-black faces that are classified as black are numerically a far larger set than the 25% of black faces that are misclassified. This results in a much higher estimated frequency of black faces - around 21% - than is actually the case.

To correct for this error, I fit a model linking the proportion of black faces found by the classifier to an actual frequency of black faces according to a set of hand-coded photographs. I first divided Flickr users into 100 groups based on the fraction of black faces found by the classifier (0-1%, 1-2 %, etc.) I sampled up to 5 users from each percentile range (not all ranges contained 5 users), then downloaded a random sample of 50 photographs with faces for each of these users. I hand-coded the photographs to arrive at an actual frequency of black faces for each user.

The rate of black faces in this sample is very low. The mean number of black faces is around 5 %, and just 14 users have a majority of black faces in their photographs. This is in spite of the fact that users with a high proportion of black faces found by the algorithm were oversampled when I selected photographs to hand-code. It appears then that black individuals are highly under-represented among Flickr users. For this reason, I will assume going forward that my Flickr sample is entirely non-black. As I discuss further in the next section, this does not interfere with my simulation procedure. Every cross-racial interaction for a non-black person represents a cross-racial interaction for a black person as well. I assume that these interactions are generated by the preferences of both interaction partners, and attempt to estimate a parameter capturing this joint surplus.

Figure 1 shows the plot of actual black faces in the hand-coded sample against the number of black faces found by the classifier, with users grouped into 2.5-percent ranges (0-2.5%, 2.5-5 %, etc) based on the algorithm's classification. There is an upward sloping and non-linear relationship between the proportion of black faces found by the detector and the user's actual proportion of black faces. Even for users with a very high proportion of black faces found by the algorithm, the actual fraction of black faces is relatively low, with a maximum of just over 30% for users who are found to have 95% of black

[33] Accuracy rates are much higher in "constrained" classification tasks, where pose and illumination are constant across subjects.

faces by the algorithm. This supports my assumption that the vast majority of users in my sample are non-black. The line on the graph shows the fitted relationship, which I estimate using a linear and quadratic term. I use this relationship to predict the proportion of black faces for each Flickr user, based on the algorithm's results. The mean fraction of black faces in the broader sample is around 3.7 %.

Figure 2 shows how the fraction of black faces found in a user's photographs is related to the percentage of black people in the user's assigned home tract. The relationship between these two variables (shown by the red line) is positive and significant, indicating that individuals living in tracts with more black people have a higher frequency of black interaction. Note, however, that the relationship is quantitatively quite small. For individuals assigned to black-majority tracts, black faces make up an average of 4.2% of all faces in photographs, compared to 3.4% for individuals assigned to black minority tracts. Individuals in all tracts are predicted to have a proportion of black faces under 6%. As indicated by the size of the circles in the graph, the number of individuals living in highly black tracts is quite small; the vast majority of Flickr users are assigned to tracts with fewer than 15% black residents. This raises the possibility that measurement error in either the home assignment or face detection process could be biasing the relationship between tract characteristics for highly black tracts. As these tracts make up a small portion of my overall sample, however, I do not expect this error to affect my estimation procedure.

It is possible that the rate of black faces in a Flickr user's photographs is not representative of the user's actual rate of interaction with black friends. This could arise for two reasons. First, it is likely that users take pictures of family, which means that the racial representation of people in the photographs will tend to overstate the degree of social segregation among friends. Second, it is possible that users "curate" their photographs to show either a higher or lower frequency of inter-racial interactions. To examine the relationship between social media photographs and actual social behavior, I turn to my survey of MTurk workers. In this survey, I ask detailed questions about respondents' social contacts and behavior; I also ask for information about the racial breakdown of faces in a recent social media photograph. I then compare the user's actual social behavior to the behavior depicted in the photo.

Recall from Table 6 that the sample of MTurk workers is disproportionately young and educated, and more likely to be white or Asian than the U.S. population at large. For education and race, this approximately mimics the demographics for the Flickr sample that I estimated in Table 7. The

relationship between social photos and social interactions in the MTurk sample is therefore likely to be similar to that found in the Flickr sample.

For each MTurk respondent, I construct a measure of the inter-racial interaction rate based on the time use portion of the survey. Recall that in this module, I ask users to account for their activities for each 3 hour portion of the day. I also ask who was present during each activity, and for the racial breakdown of the people involved. I use this information to construct the average proportion of black people present when the user spends time with friends alone.[34] This measure corresponds closely to the inter-racial interaction rate in my model, because it captures the relative amount of time spent with black friends, compared to the total amount of time spent with friends. A drawback of this measure, however, is that it can be constructed for a small number of people in my sample: only 461 of the approximately 1500 respondents spent any time with friends on the sample day, and only 217 of these spent time with friends alone (which is required for me to use the fraction of black people present as a measure of the proportion of black friends.)

To connect inter-racial interactions to social media photos, I ask respondents to choose the last photograph of a social event that they posted to a social media site, and to report the racial breakdown of people in the photograph. This exercise is randomly determined to occur before or after the reporting of social behavior. For non-black respondents, the mean reported fraction of black faces is 4%, which is very similar to my Flickr data.

Figure 3 shows a scatter plot of the inter-racial interaction rate for the non-black sample against the fraction of black faces in their social media photographs. The figure shows the fitted line from a regression of the friends measure on the photos measure (with no constant), along with the 45degree line. The two lines are indistinguishable from one another. The regression coefficient is 1.002 and is significant at the 1% level. The fraction of black faces in this *single* photograph can explain approximately 39% of the variation in the proportion of black friends. This suggests that the fraction of black individuals in a user's photographs is an excellent measure of the user's inter-racial interaction rate.

[34] The user may indicate that multiple categories of people (friends, spouse/partner, children, other family, coworkers, etc) were present for a given activity.

My simulation strategy requires that I know the actual black interaction rate for non-black individuals in my 202 CBSAs. To correct for observable differences between my Flickr sample and other individuals in these cities, I regress the black interaction rate on CBSA and tract characteristics, and use these relationships (shown in Appendix Table A4) to predict the black interaction rate for each city. This exercise suggests that the black interaction rate is slightly higher among Flickr users than among non-black Americans more generally: my estimated black interaction rate for all non-black Americans is 3.5%, compared to 3.7% in the Flickr sample.

5 Results

5.1 Main results: the effect of desegregating cities

In this exercise, I randomly reallocate individuals to new locations in their home city, and calculate how inter-racial interaction behavior would change. Step 1 of my exercise is estimating the residual parameters α_{ww} and α_{wb} from my model. Recall from Section 3 that the estimating equation I use

for this is:

$$ln(\mu_{hirs}) = \alpha^c_{rs} + ln(\Sigma_j \hat{\mu}_{irjs} + D^c_{rs}) + \varepsilon_{hirs} \quad (12)$$

In this regression, *h* indexes individuals, *i* neighborhoods, *r* and *s* are racial categories, and *c* is a city. The dependent variable is the estimated fraction of time each Flickr user spends interacting with non-black and black people respectively. This is calculated as the tract-level interaction rate multiplied by the fraction of time the user spends interacting with non-black/black people, as estimated from their photographs. For a typical user, the non-black interaction rate would be around 0.241 (a 25% interaction rate, of which 96.5% is with non-black people) and the black interaction rate would be around 0.01. The key variable on the right-hand side of the equation is $\hat{\mu}_{irjs}$, which I construct based on the disutility of travel, distance, and population supplies for every tract and link to Flickr users based on their home tract locations. I run this regression separately using non-linear least squares for white-white interactions and white-black interactions, and separately for each city. I constrain the coefficient

on $\Sigma_j \hat{\mu}_{irjs}$ to be equal to 1. The constant term outside of the logarithm is my estimate of α^c_{rs}, while a constant term within the logarithm term is interpreted as reflecting

D^c_{rs} - a measure of the strength of residential sorting based on interaction-relevant characteristics.

This procedure produces a separate estimate of α^c_{ww} and $\alpha_{wb}{}^c$ for each city *c*, and, when combined with the error term ε_{hirs}, for each Flickr user. In a next step, I adjust the estimates of α^c_{ww} and $\alpha_{wb}{}^c$ to take account of the observable differences between Flickr users and other residents of their cities. Specifically, I run a regression of ε_{hirs} on the same set of observable characteristics shown in Table A4 and use the coefficients to predict the average level of ε_{hirs} for non-black individuals in the city; I then add this to the term $\hat{\alpha^c_{sr}}$ to arrive at my final estimates.

The fifth column of Table 1 shows the average of the parameters α^c_{ww}, as well as its variation by region. The mean estimate across cities is -3.996. The negative value indicates that individuals prefer not to interact at a randomly selected moment in time; this occurs because individuals choose not to socialize about 75% of the time. Of course, the realized value of interactions *when the individual chooses to interact* are positive. This occurs when the random shock in the individual's utility function is sufficiently positive. The table also shows that individuals in the South appear to enjoy same-race interactions more than individuals in other regions, with individuals in the Northeast enjoying interactions the least.

The sixth column of Table 1 shows the estimates of α^c_{wb}, while the seventh column shows the ratio of α^c_{wb} to α^c_{ww}. The latter measure captures the relative distaste for black interactions, compared to non-black interactions. The average value of α_{wb} is -8.384, indicating that the typical non-black American dislikes interacting with black people about twice as much as she dislikes interacting with non-black people. This ratio ranges from about 1.7 in the Northeast, to 2.4 in the South.

Table 9 shows how this relative distaste for black interactions varies by city characteristics. The first column confirms the regional differences, and shows that relative distaste for black interactions is significantly higher in the South than in the other three regions. The second column shows that the relative distaste for black interactions is higher in more segregated cities, while column (3) shows that it is higher in cities where blacks make up a large proportion of the population. Column (4) shows that there is less distaste for black interactions in larger cities. The fifth column combines these covariates,

and shows that the regional differences disappear once you control for the proportion black in the population, but that all other relationships remain similar.

Although I have estimated the terms α_{ww} and α_{wb} for non-black individuals only due to lack of data on black interaction rates, these parameters can be interpreted as capturing the preference for same-race and different-race interactions respectively. When I implement my simulation exercise, I will impose this assumption, and set $\alpha_{bb} = \alpha_{ww}$.

In the next step of my simulation exercise, I reallocate individuals to new locations in the city. In doing so, I eliminate the term D from the interaction decision. In other words, I eliminate all residential sorting based on observable or unobservable characteristics. Each resident therefore views individuals from every other neighborhood as being (on average) inter-changeable.[35] This allows me to calculate the equilibrium pattern of interactions between every pair of neighborhoods in this world, using the equilibrium condition from the model.

While I have interpreted the parameters α as reflecting preferences for interactions, the more accurate interpretation is that they capture all factors other than physical distance that shape interaction behavior. Importantly, this could include other causal effects of neighborhoods. If this is the case, the terms α_{ww} and α_{wb} will not remain constant when I reallocate individuals to new neighborhoods. This means that my estimates place a bound on the impact of desegregating cities. In other words, if interaction behavior is strongly shaped by neighborhood channels other than the effect of physical distance, then redistributing individuals in a less segregated way may have a more positive effect on inter-racial interaction behavior than that which I estimate.

Table 10 shows the results of this exercise. In the first two rows of the table, I highlight an important and unintended potential consequence of efforts to make cities less segregated: a decline in the overall interaction rate. By removing individuals from the immediate vicinity of interaction partners that they like, the relative cost of socializing rises. My results suggest that this channel is quite strong. Depending on the value of δ that I use and the group I consider (blacks or non-blacks), the simulated interaction rate falls from around 25% to between 5-15%. In Table 11, which breaks the results for non-blacks down by region, I show that this effect is relatively strong in the Northeast, where

[35] Of course, the actual individuals they choose to interact with are not likely to be random; this is captured in the random shock in the utility function.

interaction rates fall to 2.5-5%. The Pacific region also shows a large decline, with interaction rates falling to 4-10%. This is because, as shown in Table 1, these regions have relatively high distaste for travel. The effect is somewhat more muted in the Midwest and South, where individuals are more willing to travel.

In the next rows of Table 10 and Table 11, I show how the proportion of interactions that take place with black people changes when we desegregate cities. The results for the non-blacks in Table 10 show that, for the country as a whole, the relative black interaction rate *falls* when we desegregate cities. Again, this must be related to the fact that individuals are no longer sorted into neighborhoods where they like the black population better. In real life, individuals interact with their black neighbors, both because they are close by and because these neighbors are more closely matched on observable and unobservable characteristics than the typical black person within a city. When we move these desirable black interaction partners further away, the black interaction rate falls. This effect is not mechanical. As shown by the regional breakdown in Table 11, the decline in the rate of black interactions is driven by the South and the Pacific regions, with the black interaction rate actually rising slightly in the Northeast and Midwest. Nonetheless, the potential for black interactions to fall when cities become less segregated is, to my knowledge, a new result in

this literature.

The final rows of Table 10 and Table 11 combine the results on the total interaction rate and the fraction of black interactions to calculate the amount of time that non-blacks spend interacting with blacks. Unsurprisingly, the net effect of a reduction in social interactions and a decline in the relative black interaction rate is to reduce the overall black interaction rate quite substantially. I estimate that non-blacks in the U.S. currently spend about 1% of their non-work time interacting with black people. Under the desegregated regime, this would fall to 0.1-0.3%. Table 11 shows that this is true in every region, even those where the relative black interaction rate rises. This is because the decline in the number of social interactions is much larger in scale than the slight increase in the proportion of black interactions in these regions.

5.2 Supplementary results: is the disutility of travel "big"?

It is possible that the results from my main simulation exercise occur because the disutility of travel is not an important factor in determining individuals' interaction behavior. In this case, the positive causal effect of desegregating neighborhoods in my model is quite limited, and could be outweighed by a relatively small "sorting" effect. In this exercise, I attempt to rule out this possibility by showing that the distaste for travel can explain important deviations in interaction behavior.

Recall that in this simulation, I model a world in which individuals maintain their existing residential locations but make interaction decisions purely based on the distaste for travel. Despite the lack of racial preferences in this world, there will be some social segregation induced by racial segregation in geographic location. By calculating the magnitude of social segregation in this world and comparing it to the perfectly integrated ideal, we can get a sense of whether the parameter δ (the disutility of travel) is quantitatively large; that is, whether it alone can explain significant deviations in interaction behavior from the case of random matching.

The results of this exercise are presented in Table 12. In the first row of the table, I show the actual rate of black interaction for non-blacks. As described in the previous section, I believe this number to be around 3.5%, indicating that non-black individuals have about 3.5% of their interactions with black people. Unfortunately, the lack of black Flickr users in my data preclude me from measuring the fraction of black interactions for black individuals in similar way.

In the second row of the table, I show what fraction of interactions would be with black people under random matching within cities. This is equal to the proportion of black people in an individual's CBSA. For the average non-black person, random matching would imply that 13.1% of their interactions were with black people. This number is higher for black people (19.4%), because there is some segregation present even across cities. The difference between the random and actual proportion of black interactions is a measure of social segregation. This measure is shown in row 3 of the table. This measure of social segregation is 10.4 percentage points for non-blacks, and cannot be calculated for the black population.

I present the results of my simulation exercise in row 4 of the table. In columns (1) and (2), I show the proportion of black interactions that would occur for non-blacks and blacks, respectively, in the world where interaction decisions are made based purely on the distaste for travel and the parameter δ is pegged to be equivalent to the average hourly wage in each city. The simulation shows that, in this world, non-blacks would see blacks for about 11.1% of their interactions, while blacks would see other

black partners for about 37.1% of their interactions. Columns (3) and (4) show the results of the same exercise, with δ increased to be equivalent to 2 times the average hourly wage. In this case, the proportion of black interactions for non-blacks would be around 10.3% and for blacks would be around 42.0 %.

Row 5 of columns (1) and (3) of the table shows the percentage decrease in the proportion of black interactions for non-blacks, compared to the random matching ideal. Based on the existing degree of residential segregation, avoidance of travel alone can account for a 15-20% decline in the relative frequency of cross-racial interactions for non-blacks. As shown in columns (2) and (4) of the same row, the same exercise suggests an increase in the relative frequency of black interactions for black people, on the order of 90-120%. The desire to avoid excess travel, combined with the existing degree of residential segregation, seems to be sufficiently large to induce significant deviations from integrated behavior, even in the absence of racial preferences.

Another way of scaling the effect of the distaste for travel is to compare the social segregation generated by distance alone to the degree of social segregation that exists in real life. This tells us about the *relative* importance of δ compared to other factors (captured in the parameters α in my model) in explaining interaction behavior. As shown in row 6 (columns (1) and (3) only), distance alone can explain 20-30% of the existing level of social segregation. Again, this suggests that the avoidance of travel is quite a strong motivation when it comes to explaining interaction behavior.

Table 13 shows how the results of this exercise vary across Census regions, for non-blacks only. This table shows that both the absolute and relative impact of physical distance are highest in the Midwest and South (explaining a 20-30% reduction in the cross-racial interaction rate from the random ideal, and 20-40% of the total amount of social segregation), moderately sized in the Northeast, and almost non-existent in the Pacific. Examining the regional values of segregation and δ in Table 1 suggests that this pattern is driven primarily by the relatively high degree of black geographic isolation in these regions, rather than by a higher distaste for travel (in fact, the opposite is true; the distaste for travel is lowest in the Midwest and South.)

Of course, these results depend on the existing pattern of residential segregation, which is likely to be generated in part by preferences over social interactions. It is important to emphasize that this sorting does not affect the interpretation of my results. My results suggest that if we eliminated racial

preferences and all other influences on social interaction decisions, but left the pattern of residential segregation the same, we would still see a substantial deviation from the perfectly integrated ideal based purely on individuals' desire to avoid travel. In other words, given the revealed preference for avoiding travel, the existing degree of residential segregation is sufficiently high to induce substantial deviations in behavior. This tells us about the potential importance of travel-avoidance in driving interaction behavior.

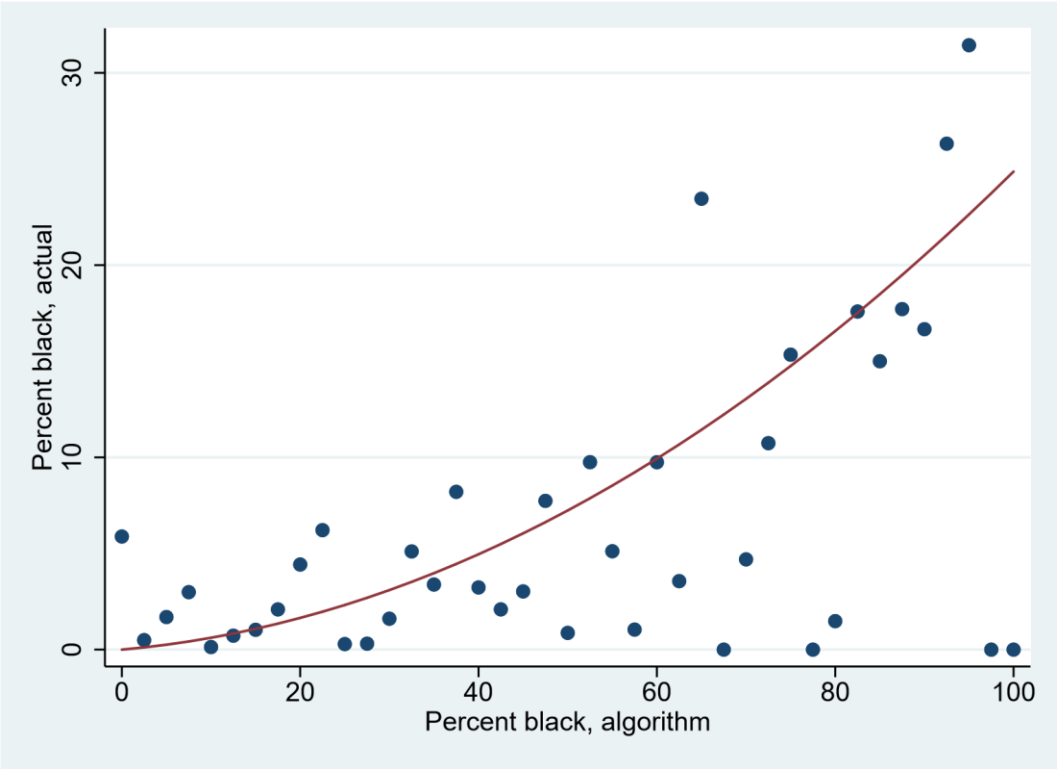
6 Conclusion

This paper highlights an important, and (to my knowledge) previously unrecognized fact: that residential sorting can *increase* inter-racial contact, even when it results in racial segregation. As I argue, and show in my secondary simulation exercise, there is at least one important causal link between neighborhoods and social interactions: the effect of physical distance. This effect is sufficiently important that it can account for a 15-20% reduction in cross-racial interactions from the perfectly integrated ideal, given the existing pattern of residential segregation. Despite this, integrating cities would have a *negative* effect on the number of cross-racial interactions. This occurs because the existing process of residential sorting encourages cross-racial interactions by ensuring that individuals pay a low "price" for interacting with the other-race individuals they like best.

While it is unlikely that policy makers would attempt the type of complete desegregation exercise I consider in this paper, my results are also relevant for assessing the more marginal programs that attempt to change the distribution of people within cities. While these programs may serve many worthwhile goals, including more evenly distributing access to good schools and other public services, they may also have unintended consequences for the pattern of social interactions in cities. As my analysis highlights, these consequences may not be unambiguously positive, and may actually serve to reduce interactions between members of different social groups.

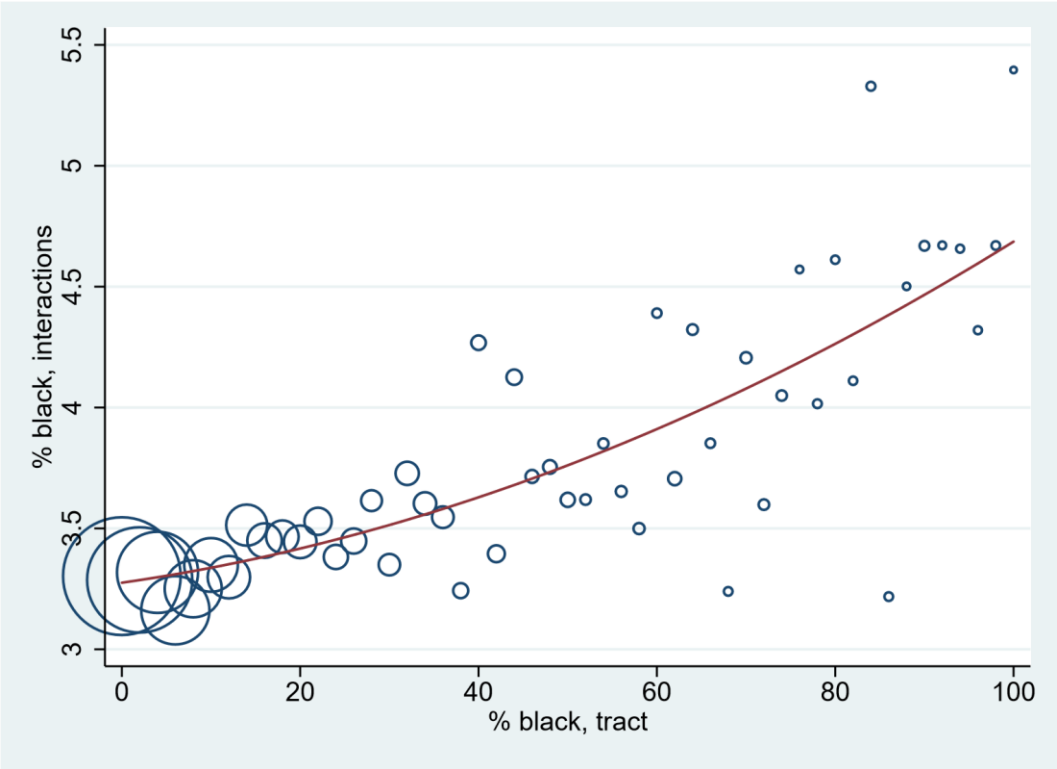
7 Figures

Figure 1: Relationship between black faces, algorithm vs. hand-coded



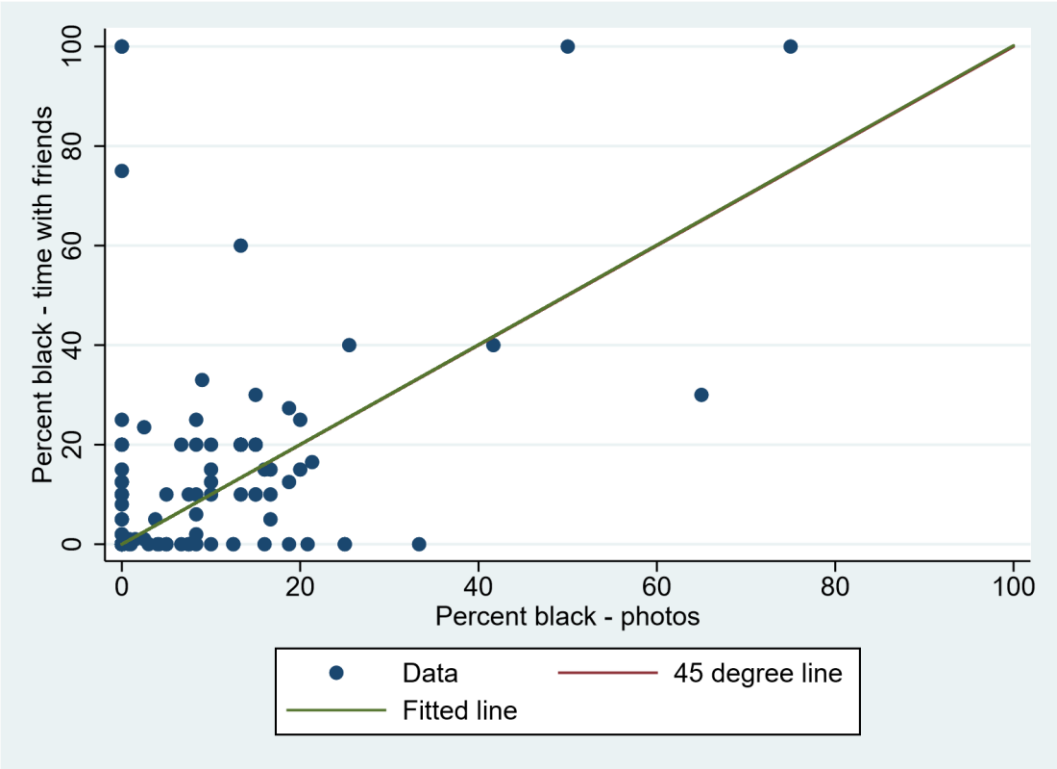
This figure plots the actual fraction of black faces for a set of hand-coded photographs against the fraction of black faces found by the race classification algorithm. Each circle represents a 2.5-percentile group for the fraction of black faces found by the algorithm (0-2.5%, 2.5-5%, etc.); the vertical height shows the mean of the actual fraction black for that group. The size of each circle represents the total number of faces found in the photographs in that group. The red line shows the fitted quadratic relationship between the algorithm's predictions and the actual frequency of black faces.

Figure 2: Relationship between fraction of black interactions and tract demographics



This figure plots the proportion of interaction time spent with black friends against the proportion of the population that is black in a user's assigned home tract. The circles represent the mean within each 2-percentage point cell on the x-asis, with the circle size indicating the number of users in this group. The red line shows the quadratic fit between the two variables. The sample is the set of approximately 93,000 Flickr users living in one of the 202 CBSAs in my main sample.

Figure 3: Relationship between time spent with black friends and black faces in social media photos: MTurk



This figure plots the proportion of interaction time spent with black friends against the proportion of black faces in an MTurk worker's last social media photograph. The sample is a set of 170 non-black MTurk workers who responded to my survey and spent any time alone with friends on the day prior to the survey. The green line is the fitted line from a regression of the interaction measure on the photo measure (with the constant supressed); this overlaps almost precisely with the 45 degree line.

8 Tables

Table 1: Key parameters by region

	Duncan index	SSI	Avg. dist. b/w tracts	δ	α_{ww}	α_{wb}	Relative distaste for black interactions $\left(\frac{\alpha_{wb}}{\alpha_{ww}}\right)$
All	0.521	0.577	41.1 km	0.458	-3.996	-8.384	2.098
Northeast	0.563	0.474	44.9 km	0.523	-4.914	-8.385	1.706
Midwest	0.596	0.642	37.6 km	0.429	-4.597	-9.604	2.089
South	0.502	0.700	39.6 km	0.419	-3.463	-8.409	2.428

Pacific	0.457	0.271	37.5 km	0.506	-3.689	-7.250	1.965

This table shows the mean level of the key parameters in my model, for all 202 cities in my main analysis sample, and separately by region. The Duncan index is computed directly from Census data. The SSI was provided by the authors Echenique and Fryer (2007); note that this variable is available for 167 of my 202 CBSAs only. The average distance between tracts was calculated using shapefiles provided by the U.S. Census Bureau. For details on the estimation of the other parameters, please see the Data section.

Table 2: Distance to the average white/black person, by race

	Distance to the average:	
	Non-black person	Black person
Non-black	27.8 km	26.7 km
Black	26.9 km	22.7 km

This table shows the mean distance to the average black/non-black person within the same CBSA, for blacks and non-blacks separately. These figures were calculated using great-circle distance between Census tracts, based on shapefiles provided by the U.S. Census Bureau, as well as information on the population of each tract by race from the 2010 Census. The sample is the set of all individuals living in one of the 202 cities in my main analysis sample.

Table 3: Previous estimates of the disutility of distance

	Context	Year(s) of observation	Estimated cost of travel, per minute*	Ratio of travel cost to average hourly wage*
Thomadsen (2005)	Fast food, Santa Clara County	1999	0.49	2
Davis (2006)	Movie theatres, 36 cities	1996	0.23&	1
McManus (2007)	Coffee shops, University of Virginia	2000	0.10	0.5-1#
Manuszak and Moul (2009)	Gas stations, Chicago & surrounding area	2001	0.18-0.24	0.68-0.91
Houde (2012)	Gas stations, Quebec City	1991-2001	0.10-0.57@	0.75-2.50@
Seim and Waldfogel (2013)	Liquor stores, Pennsylvania	2005	0.46	1.95

* All dollar estimates are in 2002 USD. Where possible, I use the authors' reported estimates of hourly wages to construct the ratio shown in column (4). Where this is not possible, I use the national hourly wage for the appropriate year, multiplied by the ratio of median income in the relevant geographic area to the median income of the United States.

& Davis (2006) estimates a non-linear function of distance; following Seim and Waldfogel (2013), the reported coefficient is the estimated cost of travelling 3.2 km.

The estimated coefficient is equal to approximately the average wage for students in the relevant geographic market; it is equal to about 0.5 times the average wage for adults in Virginia.

@ The initial estimates reported by (Houde, 2012) are larger than this. His preferred estimates suggest that a time valuation of 4 times the average hourly wage. However, these estimates do not account for traffic. Once I adjust for the average speed of traffic in Quebec City at rush hour (the relevant time, since the estimates examine consumers' willingness to deviate from commute paths), the estimates are reduced to those shown in the table.

Table 4: Relationship between estimated disutility of travel and travel patterns in Flickr

	Fraction of photos taken within indicated distance of home			
	1 km	3 km	5 km	10 km
Coefficient on $\hat{\delta}$	0.158*	0.168**	0.154**	0.089***
	(0.095)	(0.085)	(0.069)	(0.043)
N	202	202	202	202

This table shows the results from a regression of the mean fraction of photos taken within the indicated distance of Flickr users' homes on the estimated disutility of travel. A increase in the disutility of travel implies that users living within that city or tract dislike travel more. The sample is the set of 202 CBSAs in my main analysis sample.

Table 5: Frequency of social interactions: ATUS

	Dependent variable: indicator for spent any time with friends on diary day	
	Non-black sample	Black sample
Age	-0.017***	-0.011***
	(0.001)	(0.001)
Age squared	0.000***	0.000***
	(0.000)	(0.000)
No high school	-0.112***	-0.051**
	(0.008)	(0.020)
High school	-0.083***	-0.065***
	(0.006)	(0.018)
Some college	-0.059***	-0.068**
	(0.006)	(0.018)
Bachelor's	-0.025***	-0.059***
	(0.006)	(0.020)
Region - Northeast	0.004	-0.009
	(0.004)	(0.019)
Region - Midwest	0.010*	0.002
	(0.005)	(0.019)
Region - South	0.002	-0.005

			(0.005)	(0.017)
Constant	0.722***	0.596***		
			(0.005)	(0.038)
N	56,048	9,073		
R^2	0.028	0.019		
Mean of dep. variable	0.245	0.246		

This table shows the results from a regression of an indicator for spending any time with friends on demographic characteristics, in the American Time Use Survey sample. The sample is the set of all diary respondents aged 15-85 who filled out a diary on a weekend day.

Table 6: Demographics: MTurk survey respondents compared to U.S. population

	MTurk	U.S. population (over 18)
% age 18-25	15.5%	15.0 %
% age 26-45	67.9%	36.0 %
% age 46-65	14.5%	33.3 %
% age 65	2.1%	15.7 %
% male	52.1%	48.5 %
% white	73.5%	69.3 %
% black	7.1%	11.8 %
% Asian	6.5%	5.0 %
% Hispanic	11.7%	13.9 %
% No high school	0.1%	13.3 %
% high school	10.4%	38.0 %
% some college	33.3%	23.3 %
% Bachelor's	41.3%	16.4 %
% post-grad	14.9%	9.0 %
N	1,628	11,572,214

This table shows demographic information for my MTurk survey sample, and for the U.S. population over the age of 18. Information on the U.S. population comes from the 2006-2010 American Community Survey.

Table 7: Tract demographics: comparison to U.S. population

	Flickr users	Population - same cities	Population - all
CBSA population	5,348,751	4,688,542	3,834,430
CBSA SSI*	0.702	0.711	0.700
Tract level:			
Density (pop/sq. km)	3,338	2,604	2,197
Median age	37.8	37.0	37.1
Median income	$34,171	$30,461	$29,046
% white	72.5	71.0	73.3
% black	10.3	13.6	12.7
% Asian	8.5	5.8	4.9
% Hispanic	13.3	17.8	16.4
% No high school	11.4	14.8	15.2
% high school	20.5	27.0	28.3
% some college	25.0	28.0	28.2
% Bachelor's	25.2	19.0	17.8
% post-grad	17.9	11.2	10.5
Number of individuals	87,640	231,874,255	285,098,410

Number of tracts	27,279	54,336	66,970
Number of cities	202	202	933

This table shows average home CBSA and tract demographics for users in my sample, compared to the averages for the U.S. population living in the same set of cities (column 2), and to the entire U.S. population (column 3.) Demographic information is taken from the 2006-2010 American Community Surveys. * This variable is calculated for the 167 CBSAs in my main analysis sample for which it is available.

Table 8: Race classification confusion matrix

	Classified as:	
	Non-black	Black
Actual race: Non-black	87%	13 %
Black	25%	75 %

This table shows the proportion of non-black/black faces that were classified as non-black/black by the race classification algorithm. The sample is a subset of faces from 20,000 randomly selected Flickr photographs, 10 % of which are set aside for testing purposes.

Table 9: Relationship between city characteristics and relative distaste for black interactions

	Dependent variable: $\frac{\alpha_{wbc}}{\alpha_{wwc}}$				
	(1)	(2)	(3)	(4)	(5)
Northeast	0.162				-0.130
	(0.309)				(0.315)
Midwest	0.311				-0.141
	(0.290)				(0.332)
South	0.666***				-0.017
	(0.256)				(0.323)
SSI		0.580**			0.512
		(0.291)			(0.470)
% black			2.934***		2.208*
			(0.882)		(1.300)
Log population				-0.247***	-0.345***
				(0.093)	(0.109)
Mean of dep. var.	2.434	2.434	2.434	2.434	2.434
N	167	167	167	167	167

This table shows the results from a regression of relative distaste for black interactions on city characteristics. The relative distaste term is constructed as $\frac{\alpha_{wbc}}{\alpha_{wwc}}$, where α_{wbc} captures non-black people's preferences for interacting with black people and α_{wwc} captures non-black people's preference for interacting with other non-black people. As both terms are negative for all cities in the sample, more positive values of this term indicate a stronger relative *dislike* for black interactions. See the text for details on how the α terms are estimated. The sample is the set of 167 cities in my main analysis sample that have information on the black SSI from Echenique and Fryer (2007).

Table 10: Results: main simulation exercise

			δ low	δ high
	Non-black	Black	Non-black	Black
Interaction rate				
Actual	25.1%	26.5%	25.1%	26.5 %
Simulated	11.4%	14.5%	5.2%	6.7 %
% of interactions with blacks				
Actual	3.5%	Unknown	3.5%	Unknown
Simulated	2.3%	93.2%	2.3%	93.2 %
% of time interacting with blacks				
Actual	0.9%	Unknown	0.9%	Unknown
Simulated	0.3%	13.5%	0.1%	6.2 %

This table shows the results from my second simulation exercise, in which individuals are randomly distributed throughout the city but maintain their racial preferences. The first two columns show the results of the exercise when I peg δ, the disutility of travel, to be equivalent to the hourly wage in the city; the third and fourth columns show the results when I peg δ to be equivalent to 2 times the average hourly wage. Row 1 shows the actual and simulated interaction rate for blacks and non-blacks. Row 2 shows the

actual and simulated proportion of all interactions that take place with black people. Row 3 combines the first two rows and calculates the percentage of time that each group spends interacting with black people.

Table 11: Results: main simulation, by region (non-blacks)

	Northeast	Midwest	South	Pacific
Interaction rate				
Actual	24.1%	25.5%	24.9%	25.7 %
Simulated - δ low	5.3%	15.4%	14.6%	9.6 %
Simulated - δ high	2.4%	6.8%	6.9%	4.1 %
% of interactions with blacks				
Actual	3.6%	3.3%	3.5%	3.5 %

Simulated - δ low	3.9%	3.9%	1.3%	1.2 %
Simulated - δ high	3.9%	4.0%	1.3%	1.2 %
% of time interacting with blacks				
Actual	0.9%	0.8%	0.9%	0.9 %
Simulated - δ low	0.2%	0.6%	0.2%	0.1 %
Simulated - δ high	0.1%	0.3%	0.1%	0.0 %

This table shows the results from my second simulation exercise, in which individuals are randomly distributed throughout the city but maintain their racial preferences, separately by region and for non-blacks only. Row 1 shows the actual and simulated interaction rate. Row 2 shows the actual and simulated proportion of all interactions that take place with black people. Row 3 combines the first two rows and calculates the percentage of time that each group spends interacting with black people.

Table 12: Results: secondary simulation exercise

			δ low	δ high
	Non-blacks	Blacks	Non-blacks	Blacks
Black int. rate:				
Actual	3.5%	Unknown	3.5%	Unknown
Random	13.1%	19.4%	13.1%	19.4 %
Social segregation (Actual - random)	9.6 pp	Unknown	9.6 pp	Unknown
Simulated	11.1%	37.1%	10.3%	42.0 %
% change, random to simulated	-15.3%	91.2%	-21.4%	116.4 %
% explained by distance	20.8%	Unknown	29.2%	Unknown

This table shows the results from my first simulation exercise, where individuals are assumed to maintain their existing location in the city but make interaction decisions based on the disutility of travel only. The first row shows the actual proportion of interactions with black people, based on the Flickr data; because my Flickr data do not contain a sufficient number of black individuals, I report this for non-blacks only. The second row shows the proportion of black interactions that would occur if individuals matched randomly within cities. It is equal to the proportion of black people within an individual's CBSA. The difference between the random and actual rates is my index of social segregation, reported in row 3. Row 4 reports the frequency of black interactions that would occur if individuals made decisions based only on the disutility

of travel. The fifth row reports the percentage difference between the predicted and random rate, while the sixth row reports the % of the total social segregation index that can be accounted for by distance alone.

Table 13: Results: secondary simulation, by region (non-blacks)

	Northeast	Midwest	South	Pacific
Actual	3.6%	3.3%	3.5%	3.5 %
Random	12.7%	13.6%	19.2%	5.5 %
Segregation	9.1 pp	10.3 pp	15.7 pp	2.0 pp
Simulated (δ low)	11.2%	11.1%	15.7%	5.5 %
Simulated (δ high)	10.2%	9.7%	14.6%	5.4 %
% change, random to simulated	-11.8 to -19.7%	-18.4 to -28.7%	-18.2 to -24.0%	0.0 to -1.8 %
% explained by distance	16.5 to 27.5%	24.3 to 37.9%	22.3 to 29.3%	0.0 to 4.5 %

This table shows the results from my first simulation exercise, where individuals are assumed to maintain their existing location in the city but make interaction decisions based on the disutility of travel only, with the results broken down by region. The interpretation of each line is the same as in Table 12, with the exception that I report ranges for the last two rows (based on the simulated results with δ low and δ high).

9 Appendix

9.1 Additional tables

Table A1: Number of visits to home and other locations in CBSA

	Mean number of visits	Fraction of all visits
Home tract	14.0	42.3 %
Other tracts	2.6	4.5 %

This table shows the number of unique visits a Flickr user makes to his or her assigned home tract, compared to other tracts she visits. The sample is a set of approximately 87,000 Flickr users who live in one of the 202 CBSAs in my main analysis sample.

Table A2: Number of Foursquare venues around photo locations: home tract vs other visited tracts

	Number of venues		
	Home tract	Other visited tracts	Difference
Food & drink	0.780	0.929	-0.149***
			(0.016)
Entertainment	0.550	0.567	-0.016
			(0.014)
Stores	0.342	0.448	-0.107***
			(0.011)
Offices	0.138	0.156	-0.018***
			(0.005)
Other	1.962	2.092	-0.129***
			(0.026)
All venues	3.773	4.191	-0.419***
			(0.036)

This table shows the mean number of Foursquare venues within 25 m of a photograph's location, depending on whether that location is within the Flickr user's home Census tract or not. The sample for these calculations is a sample of approximately 40,000 photos taken by about 20,000 Flickr users who live in one of the 202 CBSAs in my main sample. Users were randomly sampled for inclusion in this set, and two photos were randomly sampled for each user: one in the user's "home" census tract, and one in another tract that the user visits.

Table A3: Relationship between demographics predicted by last name and home tract demographics

	Dependent variable:		
	% same race/eth, CBSA	% same race/eth, visited tracts	% same race/eth, home tract
Prob. user is white	0.073***	0.099***	0.107***
	(0.004)	(0.005)	(0.008)

N	10,172	10,172	10,172
$R2$	0.027	0.033	0.016
Prob. user	0.033***	0.053***	0.082***
is black	(0.007)	(0.008)	(0.012)
N	10,172	10,172	10,172
$R2$	0.002	0.004	0.004
Prob. user	0.116***	0.136***	0.140***
is Hispanic	(0.004)	(0.004)	(0.006)
N	10,172	10,172	10,172
$R2$	0.061	0.096	0.057
Prob. user	0.045***	0.099***	0.100***
is Asian	(0.003)	(0.004)	(0.005)
N	10,172	10,172	10,172
$R2$	0.019	0.059	0.036

The table shows the results from a regression of the proportion white, black, Hispanic or Asian in i) a Flickr user's CBSA, ii) all tracts a Flickr user visits (excluding the home tract), and iii) the home tract on the user's probability of being that race, based on his or her last name. For example, Column (1) in Panel 1 regresses the proportion white in a user's CBSA on the probability that she is white based on her last name. The probability distribution over race by last name comes from the Census Bureau (2010). The sample is the set of Flickr users in one of the 202 CBSAs in my main sample who have a last name appended to their profile.

Table A4: Relationship between black interactions and city/tract characteristics: Flickr

	Dependent variable: black interaction rate
Ln CBSA population	-0.002

	(0.031)
CBSA segregation	0.146
	(0.461)
CBSA % black	0.006
	(0.005)
Ln tract population	-0.102**
	(0.041)
Tract % black	0.012***
	(0.002)
Tract % no high school	-0.003
	(0.004)
Tract % high school	-0.001
	(0.004)
Tract % some college	-0.011***
	(0.004)
Tract % college	-0.006
	(0.005)
Tract median age	-0.043*
	(0.023)
Tract median age squared	0.001**
	(0.000)

Ln tract median income	-0.012
	(0.076)
Tract density	0.000***
	(0.000)
Northeast	0.002
	(0.097)
Midwest	-0.168
	(0.114)
East North Central	-0.094
	(0.103)

10 References

References

Ananat, Elizabeth Oltmans. 2011. "The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality." *American Economic Journal: Applied Economics*, 3(2): 34–66.

Bayer, Patrick, Stephen L. Ross, and Giorgio Topa. 2008. "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes." *Journal of Political Economy*, 116(6): 1150–1196.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis*.

Brueckner, Jan K., and Ann G. Largey. 2008. "Social Interaction and Urban Sprawl." *Journal of Urban Economics*, 64(1): 18–34.

Card, David, and Jesse Rothstein. 2007. "Racial Segregation and the Black-White Test Score Gap." *Journal of Public Economics*, 91(11-12): 2158–84.

Choo, Eugene, and Aloysius Siow. 2006. "Who Marries Whom and Why." *Journal of Political Economy*, 114(1): 175–201.

Cutler, David M., and Edward L. Glaeser. 1997. "Are Ghettos Good or Bad?" *Quarterly Journal of Economics*, 112(3): 827–872.

Dahl, Gordon B., Katrine V. Loken, and Magne Mogstad. 2014. "Peer Effects in Program Participation." *American Economic Review*, 104(7): 2049–2074.

Davis, Peter. 2006. "Spatial Competition in Retail Markets: Movie Theaters." *RAND Journal of Economics*, 37(4): 964–982.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics*, 118(3): 815–842.

Echenique, Federico, and Roland G. Fryer. 2007. "A Measure of Segregation Based on Social Interactions." *Quarterly Journal of Economics*, 122(2): 441–485.

Galichon, Alfred, Scott Duke Kominers, and Simon Weber. 2016. "Costly Concessions: An Empirical Framework for Matching with Imperfectly Transferable Utility."

Glaeser, Edward L., and Joshua D. Gottlieb. 2006. "Urban Resurgence and the Consumer City." *Urban Economics*, 43(3): 1275–1299.

Han, Hu, and Anil K. Jain. 2014. "Age, Gender and Race Estimation from Unconstrained Face Images." *MSU Technical Report*, , (MSU-CSE -14-5).

Houde, Jean-Francois. 2012. "Spatial Differentiation and Vertical Mergers in Retail Markets for Gasoline." *American Economic Review*, 102(5): 2147–2182.

Huff, Connor, and Dustin Tingley. 2015. "'Who are these people?' Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research and Politics*, 2(3): 1–12.

Jeffries, Adrianne. 2013. "The Man Behind Flickr on Making the Service 'Awesome Again'." *The Verge*.

Kazemi, Vahid, and Josephine Sullivan. 2014. "One Millisecond Face Alignment with an Ensemble of Regression Trees." *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874.

Krysan, Maria, and Kyle Crowder. 2017. *Cycle of segregation: social processes and residential stratification.* New York:Russell Sage Foundation.

Manuszak, Mark D., and Charles C. Moul. 2009. "How Far for a Buck? Tax Differences and the Location of Retail Gasoline Activity in Southeast Chicagoland." *Review of Economics and Statistics*, 91(4): 744–765.

Marmaros, David, and Bruce Sacerdote. 2006. "How do Friendships Form?" *Quarterly Journal of Economics*, 121(1): 79–119.

Massey, Douglas S., and Nancy A. Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, MA:Harvard University Press.

McManus, Brian. 2007. "Nonlinear Pricing in an Oligopoly Market: The Case of Specialty Coffee." *RAND Journal of Economics*, 38(2): 512–532.

Patacchini, Eleonora, Pierre M. Picard, and Yves Zenou. 2015. "Urban Social Structure, Social Capital and Spatial Proximity."

Seim, Katja, and Joel Waldfogel. 2013. "Public Monopoly and Economic Efficiency: Evidence from the Pennsylvania Liquor Control Board's Entry Decisions." *American Economic Review*, 103(2): 831–862.

Thomadsen, Raphael. 2005. "The Effect of Ownership Structure on Prices in Geographically Differentiated Industries." *RAND Journal of Economics*, 36(4): 908–929.

Wilson, William Julius. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy.* Chicago, IL:University of Chicago Press.

Measuring the Price Predictability of Broker Identities in the Limit Order Book – a Deep Learning Approach

Samuel Ping-Man Choi[1], Yin-Hei Chan[2], Sze-Sing Lam[1], Hie-Yiin Hung[1]

[1] *Lee Shau Kee School of Business & Administration, Open University of Hong Kong*

[2] *School of Science and Technology, the Open University of Hong Kong*

schoi@ouhk.edu.hk, g1140770@study.ouhk.edu.hk , sslam@ouhk.edu.hk, hyhung@ouhk.edu.hk

ABSTRACT

Limit order books (LOBs) have been widely adopted as a trading mechanism in global securities markets and the degree of LOB transparency is one of the most studied topics in market design. This paper investigates the information content of LOB data from 10 selected Hong Kong stocks. The LOB data are first divided into two data sets – the anonymous LOB and broker IDs in the bid and ask queues, and deep learning is employed to evaluate and compare the price movement predictability of both data sets. The result indicates that the overall prediction performance solely based on broker ID data set can reach, on average, 87.65% that of the anonymous LOB data set. The contributions of this paper are two folds. First, a machinelearning based tool for finance researchers is proposed to quantitatively measure the price predictability of LOB features, and an empirical study is conducted on the impact of LOB transparency on traders' profitability. Second, the empirical result strongly suggests that the broker ID queues in the LOB consist of significant information content for price prediction and it provides insights for regulators to determine the appropriate degree of LOB transparency so as to enable a fair market design for all investors.

KEYWORDS

Limit Order Book; Broker Identities; Market Transparency; Microstructure; Deep Learning; Price Predictability

1. INTRODUCTION

Limit order books (LOBs) enable a centralized, order-driven trading mechanism and have been widely adopted in global securities markets. The degree of LOB transparency greatly affects market efficiency and is one of the most studied topics in market design. Various levels of LOB transparency provide different trading insights for investors and thus regulators are concerned with what degree of transparency is best suit for the market.

In addition, the recent advent of high frequency trading (HFT) has profoundly altered the landscape of the stock market by gradually replacing human traders with automated algorithmic traders. In HFT, the decision time frame greatly reduces from minutes to fractions of a second, and LOB, as the main source of information for HFT algorithms in forming trading decisions, has become increasingly important (O'Hara, 2015). The proliferation of HFT poses new research issues in market microstructure, and demands an efficient and effective method for processing and analysing the enormous amount of market data (O'Hara, 2015). This research study empirically measures the impact of two LOB features on the price predictability and provides useful insight for designing HFT strategies.

We propose the use of deep learning framework for evaluating the information content of broker identities in the LOB by adopting and enhancing an existing deep learning model. In recent years, deep learning has been widely adopted in various business applications (Vieira, & Ribeiro, 2018) and is an ideal tool for our study due to its capability in coping with a huge amount of noisy LOB data. Unlike other approaches which often require domain experts to handcraft essential features, deep learning is able to automatically extract useful features from the data. As shown in this study, deep learning is flexible for handling both anonymous LOB and broker ID data. Furthermore, deep learning enables fast response time and thus is well suit for HFT. Due to its superior performance in utilizing LOB for the mid-price predictability, DeepLOB (Zhang, Zohren, & Roberts, 2019) is employed as a basis to measure the information content in the LOB and the bid-ask queue.

This study examines the information content of Hong Kong stock's LOB data, which are available to the public in near real-time at a subscription cost. Note that Hong Kong stock markets do not provide transaction records with complete broker information. It would be

extremely difficult, if not impossible, to identify the initiator of a trade based solely on the real time LOB data line. In addition, the collected LOB data is inevitably noisy since some data ticks are often skipped or duplicated due to the condition of network traffic or data server. In particular, two attributes from the data set – a snap shot view of anonymous LOB and the bid and ask broker ID queues in LOB – are compared in terms of the price movement predictability.

2. DEEP LEARNING MODEL ARCHITECTURE

We constructed two deep learning models (AnonymousLOB model and BrokerID model) to analyze the two data sets mentioned in the previous section. The model architecture for AnonymousLOB is similar to that of DeepLOB (Zhang, Zohren, & Roberts, 2019) but with a major difference in the convolutional layer. In particular, the AnonymousLOB adopts the approach from (Tsantekidis et al., 2017) to direct convolute the entire LOB and its neighbor time step using a 4×40 kernel instead of breaking it down to the order level using 1×2 kernel and convoluting it with 4×1 kernel. The AnonymousLOB model treats the LOB as a whole sequence while DeepLOB is more focused on the interaction of individual orders. Besides, this approach requires fewer parameters – DeepLOB has 9,392 trainable parameters before the inception module as oppose to 3,888 for AnonymousLOB. As such, the training time for AnonymousLOB is substantially reduced while the prediction performance is on par with the DeepLOB.

The BrokerID model employs a similar structure of AnonymousLOB, but an embedding layer is added before the convolutional layer. The embedding layer translates the broker ID into 32-dimensional vectors as a dense representation. Comparing with using a one-hot encoded vector, our model requires much less computational resource and the dimension is reduced from 100x20x10,000 to 100×20×32 before feeding into the convolution model.

3. EXPERIMENTAL SETUP

Our experimental setup is mostly based on the setting of (Zhang, Zohren, & Roberts, 2019). In this research study, six months (from September 2018 to February 2019) of 10 Hong Kong stocks' intra-day LOB data are collected in the form of price, volume, and broker IDs. Among the 10,000 valid broker ID tokens, the LOB data show at most the first 30 broker IDs in the bid and ask queues.

In this study, we first divide the LOB data into two data sets – the anonymous LOB and broker IDs in the bid and ask queues, and employ the deep learning models to evaluate the price movement predictability of both data sets. The prediction performance of AnonymousLOB and BrokerID models are then compared so as to determine the effectiveness of using broker ID queues alone as a feature. The LOB data collected from the trading days between September 2018 and January 2019 are used as training data, between 1st and 10th of February as test data and the rest of February as validation data. To facilitate deep learning, all prices and volumes of orders placed in the bid and ask queues are normalized by z-score transformation using the previous trading day's mean and standard deviation.

The model is implemented in Keras using Tensorflow 1.13 backend with a fixed random seed. Both models are trained using Intel XEON E5-2630V4 with 64GB RAM and a GPU NVIDIA 2080 and optimized by using ADAM with a learning rate and a batch size obtained by the grid search. The AnonymousLOB and BrokerID models are trained for 50 epochs and 30 epochs respectively and the prediction performance using out-sample test data is reported with respect to the best precision, recall, and F1 score.

4. EXPERIMENTAL RESULTS

Table 1 reports the prediction performance of AnonymousLOB, BrokerID and the performance ratio between the two models in terms of the best precision, recall and F1 scores. The result shows that the AnonymousLOB model consistently outperforms the BrokerID model in all three measures. However, it should be noted that, overall speaking, the prediction performance solely based on broker ID queues can reach 79.49% to 96.19% that of the anonymous LOB, and 87.65% on average. Recap that only 20 broker ID numbers in the bid and ask queues (without the price and order volume) are used in the data set. It is rather surprising that using merely the broker ID queues in the LOB for price movement prediction can achieve such a close performance as oppose to using the anonymous LOB information.

	AnonymousLOB			BrokerID			BrokerID / AnonymousLOB		
StockID	Best precision	Best recall	Best F1	Best precision	Best recall	Best F1	Best precision	Best recall	Best F1
700	72.08%	66.56%	68.15%	62.16%	58.23%	59.17%	86.24%	87.48%	86.82%
1299	75.00%	66.27%	68.54%	62.52%	57.22%	58.47%	83.36%	86.34%	85.31%
2318	71.60%	64.13%	66.03%	62.06%	58.33%	59.39%	86.68%	90.96%	89.94%

27	67.40%	61.89%	62.64%	57.73%	55.00%	55.24%	85.65%	88.87%	88.19%
941	76.93%	65.77%	68.70%	61.15%	57.79%	58.87%	79.49%	87.87%	85.69%
2800	68.28%	65.66%	66.84%	62.89%	63.16%	62.76%	92.11%	96.19%	93.90%
386	74.80%	63.06%	66.94%	64.25%	54.57%	57.32%	85.90%	86.54%	85.63%
388	66.81%	62.10%	62.95%	55.55%	53.36%	53.70%	83.15%	85.93%	85.31%
1093	60.01%	57.65%	57.63%	56.12%	55.41%	55.40%	93.52%	96.11%	96.13%
3968	73.93%	61.58%	64.96%	62.80%	51.26%	53.21%	84.95%	83.24%	81.91%

Table 1: The comparison of the prediction performance of AnonymousLOB and BrokerID. The best and worst ratios between the two models are highlighted in green and yellow respectively.

5. CONCLUSION

This paper proposes the use of deep learning as a new instrument to measure the information content of broker ID queues in the LOB. With deep learning, it becomes feasible to quantitatively evaluate the effect of anonymity by comparing the prediction accuracy of price movement using the LOB data with and without broker identities.

Our empirical result indicates that using only the broker ID queues in the LOB can achieve, on average, 87.65% overall prediction accuracy comparing to using the anonymous LOB information. Since there is no information regarding the institutional brokers and retail brokers provided in the data set, it implies that the broker ID queues in the LOB contain significant information content and are valuable to traders. Our study also provides a measurable value of the broker information in the bid-ask queue for the market data provider to set an appropriate price for the level 2 data.

The contributions of this study are two folds. First, we propose a machine-learning based instrument for finance researchers to quantitatively measure the price predictability of LOB features, and conduct an empirical study on the impact of LOB transparency on traders' profitability. Second, our empirical result strongly suggests that the broker ID queues in the LOB consist of significant information content for price prediction and it provides insights for regulators to determine the appropriate degree of LOB transparency so as to enable a fair market design for all investors.

ACKNOWLEDGEMENTS

We are grateful for CASH Algo Finance Group Limited to provide the stock data for our research. This research work is supported by Faculty Development Scheme (FDS) UGC/FDS16/B10/17.

REFERENCES

O'Hara, M. (2015). High frequency market microstructure. Journal of Financial Economics, 116(2), 257-270.

Ortiz-Catalan, M., Rouhani, F., Brånemark, R., & Håkansson, B. (2015, August). Offline accuracy: a potentially misleading metric in myoelectric pattern recognition for prosthetic control. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1140-1143). IEEE.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2017, July). Forecasting stock prices from the limit order book using convolutional neural networks.

In 2017 IEEE 19th Conference on Business Informatics (CBI) (Vol. 1, pp. 7-12). IEEE.

Vieira, A., & Ribeiro, B. (2018). Introduction to Deep Learning Business Applications for Developers. Apress.

Zhang, Z., Zohren, S., & Roberts, S. (2019). DeepLOB: Deep convolutional neural networks for limit order books. IEEE Transactions on Signal Processing, 67(11), 3001-3012.

Prediction of loan default risk with an ensemble model – A two-step approach

Zhibin Xiong
South China University
Email: zxiong3@sohu.com

Jun Huang
Angelo State University
Email: jun.huang@angelo.edu

Haibo Wang
Texas A&M International University Email:
hwang@tamiu.edu

Abstract

Loan default risk prediction is a critical and challenging topic in credit risk management. A main stream to evaluate the risk is to develop classification models based on traditional statistical methods or machine learning algorithms. In recent years, ensemble models have been widely used to improve the prediction performance as multiple learning algorithms in an ensemble model usually provide higher classification accuracy than that from any of the individual learning algorithms alone. This paper proposes a two-step approach developing an ensemble model to improve predictive performance in loan default risk based on a number of real datasets. In the first step, a Kappa statistic based feature selection method is employed to select the constituent algorithms for the ensemble model. In the second step, a meta-learning strategy is applied to integrate the outputs from the selected constituent models based on the stacking method. Three credit datasets are used to test the validity and robustness of the proposed method and the results show that the proposed method performs competitively well.

Can bank interaction during rating measurement of micro and very small enterprises *ipso facto* Determine the collapse of PD status?

Marco Desogus and Beatrice Venturi

Department of Economics and Business

University of Cagliari

Italy

JEL Codes: C02; C18; G24

KEYWORDS: *credit big data, rating models, MSE (Micro Small Enterprises) lending, financial system stabilit*

Abstract

This paper begins with an analysis of trends - over the period 2012-2018 - for total bank loans, non-performing loans and the number of active, working enterprises. A review survey was done on national data from Italy with a comparison developed on a local subset from the Sardinia Region. Empirical evidence appears to support the hypothesis of the paper: can the rating class assigned by banks - using current IRB and A-IRB systems - to micro and very small enterprises, whose ability to replace financial resources using endogenous means is structurally impaired, *ipso facto* orient the results of performance in the same terms of PD assigned by the algorithm, thereby upending the principle of cause and effect? The thesis is developed through mathematical modelling that demonstrates the interaction of the measurement tool (the rating algorithm applied by banks) on the collapse of the loan status (default, performing or some intermediate point) of the assessed micro-entity. Emphasis is given, in conclusion, to the phenomenon using evidence of the intrinsically mutualistic link of the two populations of banks and (micro) enterprises provided by a system of differential equations.

Introduction: empirical evidence

Firstly, trend data (for the years 2012-2018) was collected and organised on the number of enterprises in the Sardinian regional and Italian national production sectors and, over the same interval, on their performance trends from the credit and financial system. Necessarily, a selection of the graphs on which the analyses were undertaken that drove the reasoning for this thesis are included below.

Section figures 1 and 2 - Number of micro enterprises (normalization from ISTAT and Chambers of Commerce sources[36])

Dataset: number of micro enterprises in Sardinia

Year	2012	2013	2014	2015	2016	2017	2018
B: extraction of minerals from quarries and mines	117	112	97	93	94	102	102
C: manufacturing	7,492	7,283	6,886	6,814	6,828	7,023	7,031
D: supply of electricity, gas, steam and air conditioning	85	109	107	113	132	109	109
E: supply of water, sewerage, waste management and environmental remediation services	190	198	199	214	207	201	201
F: construction	14,340	13,773	13,121	12,619	12,639	13,227	13,243
G: wholesale and retail trade, repair of motor vehicles and motorcycles	30,252	30,137	29,205	28,653	28,993	29,290	29,326
H: transport and storage	3,090	3,038	2,910	2,807	2,853	2,924	2,927
I: accommodation and food service businesses	9,409	9,521	9,424	9,499	9,795	9,478	9,490
J: information and communications services	1,856	1,835	1,783	1,801	1,821	1,809	1,812
K: financial and insurance service businesses	1,613	1,612	1,634	1,637	1,668	1,624	1,626
L: real estate businesses	2,776	2,946	2,860	2,863	3,016	2,877	2,880
M: professional, scientific and technical businesses	15,880	15,445	15,375	15,552	15,880	15,542	15,562
N: rental and travel agencies, business support services	3,313	3,165	3,135	3,141	3,187	3,171	3,175
P: education	528	537	532	533	535	530	531
Q: healthcare and social services	6,161	6,218	6,422	6,612	6,713	6,391	6,399
R: arts, sports, entertainment and amusement businesses	1,217	1,193	1,178	1,146	1,207	1,182	1,183

[36] The methodological normalization of the two databases was required, since the data from *Unioncamere* (Chambers of Commerce), collected during the first phase of the project, were in fact redundant. This was due to the fact that the companies, which actually closed, were not extinguished in real time due to a lack of Chamber of Commerce notification of their cancellation from the Register of Companies. The ISTAT data cited also had not yet been updated to the most recent periods.

S: other service businesses	4,528	4,552	4,536	4,604	4,737	4,567	4,572
TOTAL	**102,847**	**101,674**	**99,404**	**98,701**	**100,305**	**100,045**	**100,169**

years/number of micro enterprises in Sardinia vectors	2012	2013	2014	2015	2016	2017	2018
	102,847	101,674	99,404	98,701	100,305	100,045	100,169

Dataset: number of micro enterprises in Italy

Year	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*
B: extraction of minerals from quarries and mines	1,907	1,850	1,775	1,712	1,796	1,795	1,795
C: manufacturing	345,293	338,015	328,486	321,837	330,613	330,526	330,459
D: supply of electricity, gas, steam and air conditioning	8,380	9,610	9,916	10,205	9,448	9,445	9,443
E: supply of water, sewerage, waste management and environmental remediation services	6,485	6,688	6,748	6,816	6,628	6,626	6,625
F: construction	548,709	528,592	509,648	492,388	515,477	515,341	515,237
G: wholesale and retail trade, repair of motor vehicles and motorcycles	1,124,546	1,116,087	1,086,631	1,068,659	1,089,768	1,089,481	1,089,262
H: transport and storage	119,126	117,430	113,241	110,756	114,173	114,143	114,120
I: accommodation and food service businesses	288,119	294,007	292,996	295,706	290,253	290,177	290,119
J: information and communications services	91,274	89,895	91,020	92,279	90,353	90,329	90,311
K: financial and insurance service businesses	88,998	90,637	92,831	93,799	90,799	90,775	90,757
L: real estate businesses	234,738	242,874	238,492	237,637	236,437	236,374	236,327
M: professional, scientific and technical businesses	702,053	683,778	698,154	707,020	691,902	691,720	691,581
N: rental and travel agencies, business support							

services	132,452	128,082	128,721	128,394	128,327	128,294	128,268
P: education	25,239	25,957	27,351	27,781	26,359	26,352	26,347
Q: healthcare and social services	253,160	254,655	270,894	278,646	262,123	262,054	262,001
R: arts, sports, entertainment and amusement businesses	60,658	60,382	62,001	63,011	60,997	60,981	60,969
S: other service businesses	198,593	196,542	199,755	200,185	197,103	197,051	197,011
TOTAL	**4,229,730**	**4,185,081**	**4,158,660**	**4,136,831**	**4,142,556**	**4,141,465**	**4,140,633**

years/number of micro enterprises in Italy vectors	2012	2013	2014	2015	2016	2017	2018
	4,229,730	4,185,081	4,158,660	4,136,831	4,142,556	4,141,465	4,140,633

Dataset: number of macro enterprises in Sardinia

Year	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*
B: extraction of minerals from quarries and mines	35	32	30	30	28	33	34
C: manufacturing	658	607	566	521	537	619	625
D: supply of electricity, gas, steam and air conditioning	12	14	12	11	11	13	13
E: supply of water, sewerage, waste management and environmental remediation services	96	95	84	87	80	95	96
F: construction	514	448	375	365	371	444	448
G: wholesale and retail trade, repair of motor vehicles and motorcycles	792	774	712	712	729	796	805
H: transport and storage	266	259	251	254	274	279	282
I: accommodation and food service businesses	469	437	438	460	527	499	504
J: information and communications services	67	69	76	73	68	76	76
K: financial and insurance service businesses	29	31	27	25	31	31	31
L: real estate businesses	19	12	7	9	7	12	12
M: professional, scientific and technical businesses	79	78	74	74	84	83	84
N: rental and travel agencies, business support services	259	244	250	233	251	265	268
P: education	31	29	33	34	38	35	36
Q: healthcare and social services	284	294	302	301	318	321	324
R: arts, sports, entertainment and amusement businesses	69	62	62	62	71	70	71
S: other service businesses	72	72	71	65	72	75	76
TOTAL	**3,751**	**3,557**	**3,370**	**3,316**	**3,497**	**3,746**	**3,784**

years/number of Macro enterprises in Sardinia vectors	2012	2013	2014	2015	2016	2017	2018
	3,751	3,557	3,370	3,316	3,497	3,746	3,784

Dataset: number of macro enterprises in Italy

Year	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*
B: extraction of minerals from quarries and mines	544	486	482	474	454	522	537
C: manufacturing	72,013	69,329	67,936	67,480	68,845	74,003	76,049
D: supply of electricity, gas, steam and air conditioning	546	559	543	570	567	596	613
E: supply of water, sewerage, waste management and environmental remediation services	2,482	2,433	2,398	2,415	2,432	2,604	2,676
F: construction	23,703	21,254	19,455	19,017	19,347	22,007	22,616
G: wholesale and retail trade, repair of motor vehicles and motorcycles	38,867	37,553	36,503	36,568	38,349	40,221	41,334
H: transport and storage	12,629	12,435	12,447	12,869	13,478	13,674	14,052
I: accommodation and food service businesses	19,759	19,200	19,017	19,758	21,747	21,301	21,890
J: information and communications services	6,006	6,094	5,977	6,102	6,580	6,586	6,768
K: financial and insurance service businesses	2,436	2,394	2,378	2,374	2,400	2,566	2,637
L: real estate businesses	696	690	642	636	700	720	740
M: professional, scientific and technical businesses	7,964	7,922	7,741	7,914	8,566	8,588	8,825
N: rental and travel agencies, business support services	11,318	11,280	11,177	11,201	11,632	12,121	12,456
P: education	1,651	1,720	1,737	1,785	2,001	1,904	1,957
Q: healthcare and social services	6,240	6,401	6,401	6,585	7,047	6,996	7,190
R: arts, sports, entertainment and amusement businesses	2,396	2,322	2,168	2,011	2,168	2,369	2,435
S: other service businesses	3,472	3,360	3,425	3,495	3,728	3,743	3,846
TOTAL	**212,722**	**205,432**	**200,427**	**201,254**	**210,041**	**220,523**	**226,621**

years/number of macro enterprises in Italy vectors	2012	2013	2014	2015	2016	2017	2018
	212,722	205,432	200,427	201,254	210,041	220,523	226,621

Italy: quarterly surveys of loans to micro enterprises - stock values in thousands of euro			
31/03/2012 **196,104,283**	31/12/2013 **185,086,851**	30/09/2015 **177,848,224**	30/06/2017 **164,846,165**
30/06/2012 **194,108,548**	31/03/2014 **186,303,319**	31/12/2015 **176,092,382**	30/09/2017 **158,686,183**
30/09/2012 **191,839,155**	30/06/2014 **183,362,737**	31/03/2016 **173,619,544**	31/12/2017 **157,869,518**
31/12/2012 **192,277,603**	30/09/2014 **181,991,189**	30/06/2016 **172,330,425**	31/03/2018 **157,294,801**
31/03/2013 **189,075,414**	31/12/2014 **180,982,518**	30/09/2016 **170,408,469**	30/06/2018 **151,569,685**
30/06/2013 **187,226,274**	31/03/2015 **180,320,433**	31/12/2016 **168,192,831**	30/09/2018 **148,124,634**
30/09/2013 **186,463,427**	30/06/2015 **179,629,410**	31/03/2017 **167,930,164**	31/12/2018 **144,068,531**

Sardinia: monthly surveys of loans to macro enterprises - stock values in thousands of euro

31/01/2012	20,245,576	31/10/2013	18,626,870	31/07/2015	19,243,498	30/04/2017	18,641,130
29/02/2012	20,178,507	30/11/2013	18,294,030	31/08/2015	19,175,138	31/05/2017	18,783,067
31/03/2012	19,755,188	31/12/2013	18,049,343	30/09/2015	19,032,800	30/06/2017	17,918,440
30/04/2012	20,148,404	31/01/2014	19,254,395	31/10/2015	18,949,734	31/07/2017	17,910,313
31/05/2012	19,926,114	28/02/2014	19,358,613	30/11/2015	19,003,366	31/08/2017	17,738,670
30/06/2012	19,240,685	31/03/2014	18,982,302	31/12/2015	19,099,174	30/09/2017	17,243,961
31/07/2012	19,482,460	30/04/2014	19,752,931	31/01/2016	19,112,635	31/10/2017	17,560,839
31/08/2012	19,184,105	31/05/2014	19,601,713	29/02/2016	19,013,982	30/11/2017	17,261,327
30/09/2012	19,126,712	30/06/2014	18,825,923	31/03/2016	19,080,784	31/12/2017	16,993,650
31/10/2012	19,253,365	31/07/2014	18,969,477	30/04/2016	19,130,911	31/01/2018	16,970,089
30/11/2012	19,423,965	31/08/2014	19,579,825	31/05/2016	19,202,102	28/02/2018	17,048,694
31/12/2012	18,902,468	30/09/2014	18,596,942	30/06/2016	19,226,436	31/03/2018	16,844,715
31/01/2013	19,268,629	31/10/2014	19,050,007	31/07/2016	18,919,625	30/04/2018	17,179,620
28/02/2013	19,238,602	30/11/2014	18,619,485	31/08/2016	18,724,079	31/05/2018	17,632,867
31/03/2013	18,896,487	31/12/2014	18,416,271	30/09/2016	18,765,656	30/06/2018	16,543,379
30/04/2013	19,209,658	31/01/2015	18,800,787	31/10/2016	18,805,788	31/07/2018	15,872,481
31/05/2013	18,891,720	28/02/2015	18,647,498	30/11/2016	18,629,860	31/08/2018	15,763,331
30/06/2013	18,818,546	31/03/2015	18,962,928	31/12/2016	18,339,360	30/09/2018	14,822,036
31/07/2013	18,726,947	30/04/2015	18,963,265	31/01/2017	18,799,061	31/10/2018	15,383,655
31/08/2013	18,832,875	31/05/2015	18,900,447	28/02/2017	18,727,914	30/11/2018	15,390,497
30/09/2013	18,288,769	30/06/2015	19,195,611	31/03/2017	17,952,327	31/12/2018	14,189,483

Italy: quarterly surveys of loans to macro enterprises - stock values in thousands of euro

31/03/2012	1,646,994,366	31/12/2013	1,521,713,568	30/09/2015	1,501,913,382	30/06/2017	1,426,588,448
30/06/2012	1,645,185,055	31/03/2014	1,553,078,446	31/12/2015	1,483,096,671	30/09/2017	1,366,124,098
30/09/2012	1,622,500,182	30/06/2014	1,547,273,266	31/03/2016	1,475,216,267	31/12/2017	1,374,039,081
31/12/2012	1,611,501,427	30/09/2014	1,541,493,794	30/06/2016	1,480,433,998	31/03/2018	1,370,734,809
31/03/2013	1,595,250,768	31/12/2014	1,509,099,932	30/09/2016	1,465,279,100	30/06/2018	1,316,758,282
30/06/2013	1,570,164,850	31/03/2015	1,514,608,082	31/12/2016	1,450,690,989	30/09/2018	1,301,863,430
30/09/2013	1,550,624,325	30/06/2015	1,514,910,903	31/03/2017	1,454,317,652	31/12/2018	1,268,272,037

Sardinia: net non performing loans to micro enterprises - in thousands of euro

31/03/2012	415	31/12/2013	424	30/09/2015	488

30/06/2012	**421**	31/03/2014	**434**	31/12/2015	**548**
30/09/2012	**429**	30/06/2014	**436**	31/03/2016	**557**
31/12/2012	**442**	30/09/2014	**443**	30/06/2016	**551**
31/03/2013	**405**	31/12/2014	**453**	30/09/2016	**523**
30/06/2013	**412**	31/03/2015	**464**	31/12/2016	**524**
30/09/2013	**408**	30/06/2015	**479**	31/03/2017	**514**

Italy: net non performing loans to micro enterprises - in thousands of euro					
31/03/2012	**10,473**	31/12/2013	**13,253**	30/09/2015	**14,774**
30/06/2012	**10,825**	31/03/2014	**13,624**	31/12/2015	**14,859**
30/09/2012	**11,126**	30/06/2014	**13,869**	31/03/2016	**14,683**
31/12/2012	**11,743**	30/09/2014	**14,214**	30/06/2016	**14,603**
31/03/2013	**11,967**	31/12/2014	**13,680**	30/09/2016	**14,754**
30/06/2013	**12,350**	31/03/2015	**14,076**	31/12/2016	**15,229**
30/09/2013	**12,692**	30/06/2015	**14,425**	31/03/2017	**14,845**

Sardinia: net non performing loans to macro enterprises - in thousands of euro					
31/03/2012	**1,455**	31/12/2013	**1,609**	30/09/2015	**2,209**
30/06/2012	**1,549**	31/03/2014	**1,681**	31/12/2015	**2,814**
30/09/2012	**1,627**	30/06/2014	**1,774**	31/03/2016	**2,834**
31/12/2012	**1,618**	30/09/2014	**1,908**	30/06/2016	**2,832**
31/03/2013	**1,334**	31/12/2014	**2,057**	30/09/2016	**2,815**
30/06/2013	**1,405**	31/03/2015	**2,045**	31/12/2016	**2,952**
30/09/2013	**1,499**	30/06/2015	**2,137**	31/03/2017	**2,938**

Italy: net non performing loans macro enterprises - in thousands of euro					
31/03/2012	**69,899**	31/12/2013	**104,258**	30/09/2015	**134,512**

30/06/2012	**74,349**	31/03/2014	**111,726**	31/12/2015	**136,564**
30/09/2012	**77,502**	30/06/2014	**116,406**	31/03/2016	**133,185**
31/12/2012	**81,677**	30/09/2014	**119,310**	30/06/2016	**135,080**
31/03/2013	**85,363**	31/12/2014	**122,643**	30/09/2016	**136,487**
30/06/2013	**91,292**	31/03/2015	**126,021**	31/12/2016	**138,805**
30/09/2013	**96,203**	30/06/2015	**131,237**	31/03/2017	**135,640**

Description of the data

The scenario, presented here among the pairs of figures shown above, begins from the progressive reduction in the disbursement of loans - which we assume to be an independent variable - to which a generalised increase in net nonperforming loans can be reasonably correlated [37] (average Bravais-Pearson correlation coefficient = -0.57). Nevertheless, the repercussion of this manifestation had different results on the size of the productive sector. On the one hand, micro enterprises tend to leave the market more quickly than *new-co* companies of a similar size. At the same time, macro enterprises seem to not be affected by this phenomenon. Based on this observation, a theoretical model that explains the basis of these findings was developed.

In fact, from the data-set adopted, the correlation index between the reduction in the volumes of credit disbursed to the production system and the number of enterprises operating on the market in Italy was, +0.73 for micro enterprises and -0.71 for macro enterprises[38].

Hypothesis

First, it should be taken into account that more than 90% of the makeup of Italian and Sardinian production comprises micro enterprises. In the overall statistical data, the granularity of even the largest enterprises is insufficient to compensate and support the general system, as the figures below show clearly.

	years/number of enterprises in Sardinia vectors							
	2012	2013	2014	2015	2016	2017	2018	
Micro enterprises	102,847	101,647	99,404	98,701	100,305	100,045	100,169	
Macro enterprises	3,751	3,557	3,370	3,316	3,497	3,746	3,784	
Total enterprises	106,598	105,204	102,774	102,017	103,802	103,791	103,953	

[37] The rhythmic change occurring between 2016 and 2017 was mainly due to the sale of bank NPLs to companies specialising in recovery.

[38] Regardless, it is believed that not even an inverse correlation is extant: the macro enterprises appeared to be rather indifferent to the quantitative turbulence of financial leverage, at least over the short-term. Hence, the indicator shows a non-correlation, and obviously cannot be considered reliable in the perception of a negative interdependence.

years/number of enterprises in Italy vectors								
			2012	2013	2014	2015	2016	2017	2018
Micro enterprises	212,722	205,432	200,427	201,254	210,041	220,523	226,621	Macro enterprises	4,229,730
4,185,081	4,158,660	4,136,831	4,142,556	4,141,465	4,140,633	Total enterprises	4,442,452	4,390,513
4,359,087	4,338,085	4,352,597	4,361,988	4,367,254				

Therefore, for all the reasons detailed above, stability, or rather growth, in a real economy founded on micro and very small enterprises, is a function - though not univariate - of the credit leverage disbursed. Less available credit significantly impacts the real economy, especially in countries where the financial system is centred on banks. Then, bond and equity markets are insufficiently developed and therefore cannot provide alternative resources to businesses, especially the smallest ones[39]. This idea is consistent with the hypothesis that even small shocks can cause significant effects: the worsening of credit access conditions, especially for the smallest, most indebted and vulnerable subjects, leads to a reduction in consumption and investment, which accentuates the negative phases of the cycle[40]. Currently, once the regulatory bodies' certification for adoption have been obtained, the IRB approaches applied by banks, whether "basic" or "advanced", will be administered *erga omnes* to their customers, without distinctions made due to the size of the operator requesting financial support. Clearly, the rating class assigned determines whether a subject can or cannot access credit and how much it will cost. This then orients the performance results for economic micro-entities, which are structurally incapable of replacing financial resources with endogenous means, in the same terms as the PD assigned by the algorithm, thus overturning the principle of cause and effect. The next section endeavours to formalize this thesis mathematically.

Theoretical modelling

Banks implement their strategies for making loans based on commercial and profit-based rationales. Loans are represented by risk portfolios whose overall weighting is derived from the sum of segmented portions, which in turn are based on precise evaluations of each individual counterparty. In the model below, we will proceed by approximating the portfolio average and imagining that all the *n* positions comprising it[41] [42]have the same PD (Probability of Default), the same *LGD* (Loss Given Default) and the same *EAD* (Exposure at Default): the total expected

losses TEL_n will be[7]:

[39] Cf. Panetta, Signoretti (2010).
[40] Cf. Bernanke et al. (1996).
[41] Cf. Conti (2016).
[42] Portfolio losses measured *ex post* with minimum offsets.
[8] Cf. Conti (2016).

$$[1] \quad TEL_n = n * PD * EAD * LGD$$

Because these are expected losses, they represent a cost and should be recorded in the bank's Income Statement. A separate rationale should be applied to unexpected losses, which represent a capital constraint for banks, and which determine the caution that they more frequently adopt when making loans.

At this point, we have chosen here to apply the "Vasicek model"[8], about which we are offering an integrative emanation, which describes the different mathematical effects of credit intermediation, with the subsequent disaggregation of the peculiarities of the micro and very small enterprises from the total. Our portfolio comprises the *n* positions (financed enterprises) described above and the observation we conduct is over a time *t* = 1 year:

$$[2] \quad TL_n = \sum_{i=1}^{n} U_i LGD_i EAD_i$$

Where:

TL_n is the total loss on the portfolio over the time *t*;

LGD_i and EAD_i are the LGD and the EAD of the i-th company respectively;

U_i is the Boolean indicator, which takes on a value of 1 if the i-th enterprise has reached default within time *t*, or the value 0 if the i-th enterprise has continued to be performing in the range considered.

Now, let us suppose that the default is essentially determined, for all the companies in the portfolio, by the size of the budget surplus (S_i for the i-th enterprise): below a certain size, the business does not hold up and goes into default. S_i represents a random variable, making it necessary to proceed with the description of its evolutionary process. Note that this is a Markovian process: at the instant t_0 the future random value of S_i in t_1 depends only on the state of S_i in t_0 and not on what occurred previously.

Now, let us introduce the amount *X* – to which we assign great significance for the progress of our thesis – which represents a variable of macroeconomic context and which indicates the further scenario in which the enterprises in the portfolio operate. From this there will be derived an additional random element *Y*, whose variation □*Y* in an interval □*t* is the result of the sum of the variations of the elementary states of *Y* (independent among themselves) and whose variance is analogously the sum of the elementary variances: if $\Delta Y(\delta t)$ has a variance σ^2, over the entire interval □*T,* we will see a total variance of □*Y* equal to σ^2□*t* and, with the same probability distribution of the elementary variations and a corresponding standard deviation of $\sigma\sqrt{\Delta t}$. Therefore:

$$[3] \quad dY = X\sigma\sqrt{\Delta t}$$

With *X* being a random Gaussian variable.

The overall process is described by the stochastic differential equation:

$$[4] \quad dS_i = S_i(t)r_i dt + S_i(t)X_i\sigma_i\sqrt{dt} \quad [9]$$

Supplementing with Ito's lemma, we find:

$$[5] \quad S_i(T) = S(0)e^{r_i T - \frac{1}{2}\sigma_i^2 T + \sigma_i\sqrt{T}X_i}$$

It emerges that the randomness of the process $S_i(T)$ is determined solely by X_i.

By analysing X_i, we can break it down into:

$$[6] \quad X_i = \alpha K + \beta e_i$$

where *K* is the random variable strictly interpreted from the macroeconomic context, while the e_i element indicates the random idiosyncratic factor of the individual counterparty. It should be mentioned that a certain component of e_i is however dependent on *K*.

Assuming that both of these are normal Gaussian variables (with the average = 0 and the variance = 1) that are independent among themselves, then:

$$[7] \quad \alpha^2 + \beta^2 = 1$$

[9] The equation can be rewritten: $dS_i = S_i(t)r_i dt + S_i(t)\sigma_i dW_i$. Then $X_i\sqrt{dt} = dW_i$: "Wiener process".
and

$$[8] \quad X_i = \alpha K + \sqrt{1-\alpha^2}e_i$$

With α known a priori, the above formula describes the process that leads to the default: that is, for X_i being less than a minimum M_i

$$[9] \quad PD_i = Pr(X_i < M_i)$$

hence,

$$[10] \quad PD_i = F(M_i)$$

and

$$[11] \quad M_i = F^{-1}(PD_i)$$

with *F* being the cumulative probability function of the normal Gaussian distribution.

Then, based on the above, a default state will be verified when:

$$[12] \quad \alpha K + \sqrt{1-\alpha^2}e_i < F^{-1}(PD_i)$$

The e_i factor is peculiar and intrinsic to each enterprise. Nevertheless, using it we can apply the reasoning and isolate micro enterprises with respect to medium-large enterprises. In fact, this factor can be more easily defined on the basis of deterministic considerations, which can be made on the general and accounting structures of those same mediumlarge enterprises, since they are obviously less dependent on exogenous components of their resources and therefore less sensitive to any turbulence in the economic situation and system.

Instead, for micro and very small enterprises, we might imagine the state of *e* being described by a wave function $\psi(r, t)$, which indicates the amplitude of probability to find the state, at time *t*, of the enterprise *i* in *r*. Where "*r*" is a generic point of the [κ, λ] interval, $\kappa = 0$ is the performance state and $\lambda = 1$, which represents default. Even though the *r* state is also "deterministic", in this case, the result of the measurement becomes probabilistic.

$$[13] \quad P(r,t) \propto |\psi(r,t)|^2$$

Explaining further, for simplicity's sake, let us consider a generic state $|\psi\rangle$ [43] expressed by the linear combination of the two limit states $|\kappa\rangle$ and $|\lambda\rangle$ (Eigenstates):

$$[14] \quad |\psi\rangle = \frac{1}{\sqrt{2}}\{|\kappa\rangle + |\lambda\rangle\}$$

The measuring instrument is the algorithm used to calculate the rating by a generic bank *B*, with which the microenterprise *i* requests a loan. The idea in the thesis is that the idiosyncratic component e_i precisely of micro and

[43] The formalism and conceptualism of quantum mechanics are being borrowed for a theoretical analogy.

very small enterprises, due to their own dimensional characteristics, can be found "simultaneously", at any time, in all points of the interval [κ, λ]. We do acknowledge that the measurement of a "macroscopic" rating may result in three positions:

- (0) indicated by the $|B_0\rangle$ state, if the instrument reads that the micro-enterprise's state is $|\kappa\rangle$;
- (1) indicated by the $|B_1\rangle$ state, if the instrument reads that the micro-enterprise's state is $|\lambda\rangle$;
- (0.5) characterizes the adjustment of the instrument before it is measured.

The interaction between the phenomenon and the measuring produces:

$$[15] \quad |\psi(0)\rangle_1 = |\kappa\rangle|B_{0,5}\rangle \Rightarrow |\psi(t)\rangle_1 = |\kappa\rangle|B_0\rangle$$

$$[16] \quad |\psi(0)\rangle_2 = |\lambda\rangle|B_{0,5}\rangle \Rightarrow |\psi(t)\rangle_2 = |\lambda\rangle|B_1\rangle$$

For the hypothesized linearity, we would have a generic initial state of $|\psi(0)\rangle$, at time t = 0, resulting from the overlapping of $|\psi(0)\rangle_1$ and $|\psi(0)\rangle_2$:

$$[17] \quad |\psi(0)\rangle = \frac{1}{\sqrt{2}}\{|\kappa\rangle|B_{0,5}\rangle + |\lambda\rangle|B_{0,5}\rangle\}$$

which will evolve, following the measurement evidently carried out at a time $t > 0$, in the state:

$$[18] \quad |\psi(t)\rangle = \frac{1}{\sqrt{2}}\{|\kappa\rangle|B_0\rangle + |\lambda\rangle|B_1\rangle\}$$

Clearly, this is not possible, since it is not practical to verify a contextual overlapping of the measurement (macroscopic), which simultaneously is marked $|B_0\rangle$ and $|B_1\rangle$. To understand what is actually happening, we need to contradict the hypothesis of measurement linearity and conclude that, following the measurement of the rating, the state of the micro enterprise will no longer be an overlapping of Eigenstates of the measured quantity, but will "collapse" in κ or in λ (or, more correctly, at any point r of the interval [κ, λ]) according to whether the instrument measures 0 or 1 (or an intermediate quantity, to which r, will correspond, of the expected degree of performance)[44].

[44] The empirical demonstration of the above is work in progress. In an independent manner, an experiment was started on a sample of 100 Sardinian micro-enterprises, simulating a reversal of the rating obtained and predicting its effects on the accounting. The first tests carried out so far have shown that the denial of custody to those who have obtained it (and have continued to perform), would have led to signals of default on the 300th day on average. Similarly, a loan to excluded micro enterprises and reaching non-performing enterprises, would have allowed a condition of regularity throughout the first year.

Summary and conclusions

Now, let us look again at [12]:

$$\alpha K + \sqrt{1-\alpha^2}e_i < F^{-1}(PD_i)$$

This formula indicates the probability that enterprise *i*, based on its own idiosyncratic e_i factor and the conditions of its macroeconomic context.

Looking again at [2]:

$$TL_n = \sum_{i=1}^{n} U_i LGD_i EAD_i$$

We can now explain U_i:

$$[19] \quad U_i = \begin{cases} 1 & if\ \alpha K + \sqrt{1-\alpha^2}e_i < F^{-1}(PD_i) \\ 0 & otherwise \end{cases}$$

and we will have

$$[20] \quad TL_n = \sum_{i=1}^{n} U_i(K, e_i) LGD_i EAD_i$$

Since, according to the hypothesis, PD_i, LGD_i and EAD_i are known *a priori*, the need to maximize e_i and K [45] is clear, limiting ourselves to the bank-business relationship, by both of these entities in cohabitation within the economic system: with the banks needing to minimize losses and reduce capital absorption, and the enterprises needing to obtain credit and do business sustainably and profitably.

Clearly, there are even more macroscopic components inside of *K*, including monetary policy, international economic influences, credibility and trust of the country, etc. This is why we are focusing our attention on the dynamics of the bank-enterprise system (and, in turn, families), since we cannot overlook the positive influence, including macroeconomic effects, generated by an efficient banking sector in the transmission of liquidity to the productive complex so that it can remain healthy and performing (and, on the contrasting condition, where there are the negative effects produced by the disappearance of this virtuous cycle). Let us call the population of banks Z_1 and the population of micro enterprises (portfolio) Z_2.

[45] A targeted analysis of the *K* component, based on the same initial postulate, was conducted by Desogus, Casu (2019).

Now, we should imagine two possible bank strategies:

a) continuing to manage loans using generalist algorithms;

b) reactivating qualitative forms of relationships, trust and credit cooperation.

In the first case, the result seems to be that of saving on allocations, however, the greater perceived solidity would only be apparent and over the short-term, as exposed by the empirical calculations shown in the section above. Both *e* components – with reference to micro and very small enterprises – and *K* are evidently conditioned, negatively, by the

a) strategy. In fact, a correlation between the reduction of credit to the productive system, the greater mortality of the micro enterprises (default of the Z_2 population) and the increase of the non-performing bank loans (default of the Z_1 population) has been observed. Let us now look at the formal bases that support recourse to the b) strategy: whether $z_1 = z_1(t)$ and $z_2 = z_2(t)$ are manifestations of performing loans[46] at time t of the two populations Z_1 and Z_2 respectively:

$$\begin{cases} \frac{dz_1}{dt} = z_1 f_1(z_1, z_2) \\ \frac{dz_2}{dt} = z_2 f_2(z_1, z_2) \end{cases}$$

This system of differential equations indicates the dependence of the number of performing loans of each population on that of both; the link between the two populations is in fact intrinsically mutual.

$$\frac{\partial f_1}{\partial z_2} > 0 \quad \text{and} \quad \frac{\partial f_2}{\partial z_1} > 0$$

A growth in the performance of Z_2 corresponds to a growth in Z_1 (therefore, it is systemic). The mathematical zeros of the model are,

- in the total absence of performance and the consequent extinction of the two populations
- within the limits imposed by the *K* and *e* components, noted upon initial reflection on each enterprise in the Z_2 population, which provide a logistical trend to z_2 with a high point that is always lower than the totality of Z_2.

High values of z_2 are conditioned by $\alpha K + \sqrt{1-\alpha^2}e_i > F^{-1}(PD_i)$, for which U_i – and therefore TL_n – tends toward zero.

[46] The rationale for default levels could be developed in contrasting terms: the final result would be the same.

The model's results indicate a need to make a paradigm shift in the evaluation of the creditworthiness of micro and very small enterprises. From the data collected and from the analyses carried out, current methods have proved to be ineffective in adequately supporting the productive system and, in concert with the local and national economic growth.

Some points for future research seem to be:

- regulation on the incompatibility of commercial and speculative businesses in the hands of the same bank;
- strengthening studies on qualitative rating systems;
- creation of divisions, inside banks, specialized in the analysis of small economic entities with a relationship banking approach;
- consolidation of the cooperative credit system and credit consortiums.

Bibliography

Aghion, P., Howitt P., Mayer-Foulkes D. (2005) *The effect of financial development on convergence: theory and evidence*. Quarterly Journal Economics 120(1), 173-222.

Andreozzi P., Angelini P., Di Salvo R., Ferri G. (2004), *Piccole imprese e credito cooperativo: relazioni più intense e stabili che con le altre banche? I risultati di un'indagine*, in 'Cooperazione di Credito' 183/184.

Beck, T., Levine R., Loayza N. (2000), *Finance and the sources of growth*. Journal of Financial Economics 58(1-2), 261300.

Bencivenga V., Smith B. (1991), *Financial intermediation and endogenous growth*. Review of Economic Studies 58, 195-209.

Benhabib J., Spiegel M. (2000), *The role of financial development in growth and investment*. Journal of Economic Growth 5(4), 341-360.

Bernanke B.S., Gertler M., Gilchrist S. (1996), *The Financial Accelerator and the Flight to Quality*, in 'The Review of Economics and Statistics', 78/1, pp. 1-15.

Bertuglia C.S., Vaio F. (2005), *Nonlinearity, Chaos and Complexity: The Dynamics of Natural and Social Systems.* University Press: Oxford.

Bucci, A., Marsiglio S. (2018), *Financial development and economic growth: long-run equilibrium and transitional dynamics*. Scottish Journal of Political Economy [doi:https://doi.org/10.1111/sjpe.12182].

Bucci, A., Marsiglio S., Prettner C. (2018), *On the (Nonmonotonic) Relation Between Economic Growth and Finance*. Macroeconomic Dynamics [doi:https://doi.org/10.1017/S1365100518000305]

Capinski M., Zastawniak T. (2003), *Mathematics for Finance*, Springer: London.

Conti G. (2016), *Matematica e rischio di credito*, in http://www.mathisintheair.org/wp.

Cosma S. (2002), *Il rapporto banca-impresa: le variabili relazionali e comportamentali nella valutazione del rischio di credito*, Giappichelli: Torino.

Demetriades P., Hussein K. (1996), *Does financial development cause economic growth? Time series evidence from 16 countries*. Journal of Development Economics 51, 387-411.

Desogus M., Casu E. (2018), *Essays in innovative risk management methods based on deterministic, stochastic and quantum approaches*, Anaphora Literary Press: Quanah.

Desogus M., Casu E. (2019), *A contribution on relationship banking. Economic, anthropological and mathematical reasoning, empirical evidence from Italy*, in press, forthcoming.

Duffie D., Singleton K.J. (2003), *Credit Risk: Pricing, Measurement, and Management*, University Press: Princeton.

Fernando C., Chakraborty A., Mallik R. (2002), *Relationship banking and credit limits,* report presented in 'Board of Governors of the Federal Reserve System', Washington D.C.

Ferri G., Messori M. (2000), *Bank-firm relationships and allocative efficiency in northeastern and central Italy and in the south*, in 'Journal of banking and finance' 24.

Greenwood, J., Smith B.D. (1997), *Financial markets in development, and the development of financial markets*. Journal of Economic Dynamics and Control 21(1), 145-181.

Jarrow R.A., Lando D., Turnbull S. (1997), *A Markov model for term structure of credit risk Spreads* in Review of Financial Studies No. 10 (pp. 481-523).

Levine R., Loayza N., Beck T. (2000), *Financial intermediation and growth: causality and causes*. Journal of Monetary Economics 46(1), 31- 77.

Mankiw N.G., Romer D., Weil D.N. (1992), *A Contribution to the Empirics of Economic Growth*, Quarterly Journal of Economics, vol. 107, pp. 407-437.

Meyn S.P., Tweedie R.L. (1993) *Markov Chains and Stochastic Stability*, Springer: London.
Pagano M. (1993), *Financial market and growth: an overview*. European Economic Review 37, 613-622.

Panetta F., Signoretti F.M. (2010), *Domanda e offerta di credito in Italia durante la crisi finanziaria* , Bank of Italy - Occasional Paper - Questioni di Economia e Finanza.

Perko L. (2001), *Differential Equations and Dynamical Systems*. New York: Springer.

Perroux F. (1950), *Economic Space: Theory and Application*, "Quarterly Journal of Economics", 21.

Petersen M.A., Rajan R.G., (1995), *The Effect of Credit Market Competition on Lending Relationship*, in 'Quarterly Journal of Economics', May, pp. 407-443.

Petrulli M. (2007), *Basilea 2. Guida alle nuove regole per le piccole e medie imprese*, Halley Editrice: Macerata.

Ram R. (1999), *Financial development and economic growth: additional evidence*. Journal of Development Studies 35(4), 164-174.

Rioja, F., Valev N. (2004), *Finance and the sources of growth at various stages of economic development*. Economic Inquiry 42, 127-140.

Schena C. (2004), *Il ruolo prospettico dei Confidi nel rapporto banca-impresa: mitigazione del rischio e supporto informativo*, in 'Quaderni della Facoltà di Economia dell'Università dell'Insubria'.

Scott J.A., Dunkelberg W.C. (1999), *Bank consolidation and small business lending: a small firm perspective*, in "Business access to Capital and Credit": a Federal Reserve System Research Conference.

Shaw E. (1973), *Financial Deepening in Economic Development*. New York: Oxford University Press.

Solow R.M. (1956), *A Contribution to the Theory of Economic Growth*, Quarterly Journal of Economics, vol. 70, pp. 6594.

Venturi B., Casula G. (2014), *Lecture notes in Mathematics*, University of Cagliari.

Venturi B., Pirisinu A. (2013), *Mathematics for economists-exercises, problems, models*, Lambert Academic Publishing: Saarbrücken.

EI in a World of AI
In the UAE Context

By

Maya AlHawary

To

Prof. Abtar Singh
Hamdan Bin Mohammed Smart University

What is Emotional Intelligence (EQ)?

Since its discovery in the 1990's Emotional intelligence is one of the most discussed and debated topics in literary and educational circles these days. In fact, some educationists believe that 21century society is based on emotional intelligence. The field of Emotional Intelligence has witnessed profound development since its birth both at macro, and micro as well as personal and professional level. Emotional

Intelligence was defined as "the ability to perceive, use, understand and manage emotions in one self and others" (Mayer &Salovey, 2000, p.196)[1] Golman describes EI as a set rules of social and emotional skills. He defines EQ as "the skills or competencies to be able to know one's own emotions, manage one's own emotions, self-motivate as well as recognize others' emotions and handle relationships".

(Goleman (1998, p.93)[2]

In addition, Mayer, Salovey and Caruso defined emotional intelligence (EQ) as "the capacity to reason about emotions, and of emotions to enhance thinking. It includes the abilities to accurately perceive emotions, to access and generate emotions so as to assist thought, to understand emotions and emotional knowledge, and to reflectively regulate emotions so as to promote emotional and intellectual growth"(Mayer, Salovey and Caruso (2000, p.396).[3]

People with higher intelligent quotient were considered to be successful in the past but this theory has changed now. Now, people with improve emotional quotient are considered more successful as they have the ability to fill the missing links in peculiar situations. In 70% of cases people with EQ perform much better than people with IQ. This evidence and proof has provided us with a completely new dimension in educational and other fields of life. Tones of research stretching on decades prove and support the fact that now emotional intelligence is the key factor in deciding outstanding performances. Every person is bestowed with a bit of emotional intelligence though it is abstract in nature. Emotional quotient is demonstrated in managing behaviours and attitudes, handling complicated social situations and taking personal decisions and directions. The stronger the EQ, the sounder the decision is. "Emotional intelligence is made up of four core skills that pair up under two primary

competencies: personal competence and social competence" (Talent Smart Inc., 2017, p.1).[4]

Dulewicz and Higgs argues comparing IQ with EQ, that "EQ is a key component of effective leadership and leaders with high EQ are able to recognize, appraise, predict and manage emotions in a way that enables them to work with and motivate team members" (Dulewicz and Higgs 2003, p. 2).[5] Another study, Sadri, mentions that "El is essential to effective team interaction and productivity" and that the "emotional intelligence of the team leader is important to the effective functioning of the team. The leader serves as a motivator toward collective act, and facilitates supportive relationships among team members. The emotionally intelligent team leader also provides a transformational influence over the team" (Sadri 2012, p. 535)[6]

Conversely the critiques of EQ theories are of the view that there is an inflated publicity on the topic. These researchers believe that staunch supporters and promoters of the emotional quotient do not clarify the design defined by its chief founders, which eventually make people land in misunderstandings and bizarre

claims.

Zeinder and Roberts state that "on the negative side, writings on EQ place greater emphasis on the emotional abilities than on intellectual intelligence – an outcome that isagreeableto the personal profiles and worldviews of many" (Zeinder and Roberts

2002, p.6).

The construct of Golman on has been under criticism for its obscure design. Opponents of his construct also believe that there is no precise and tangle published evidence to support his claims.

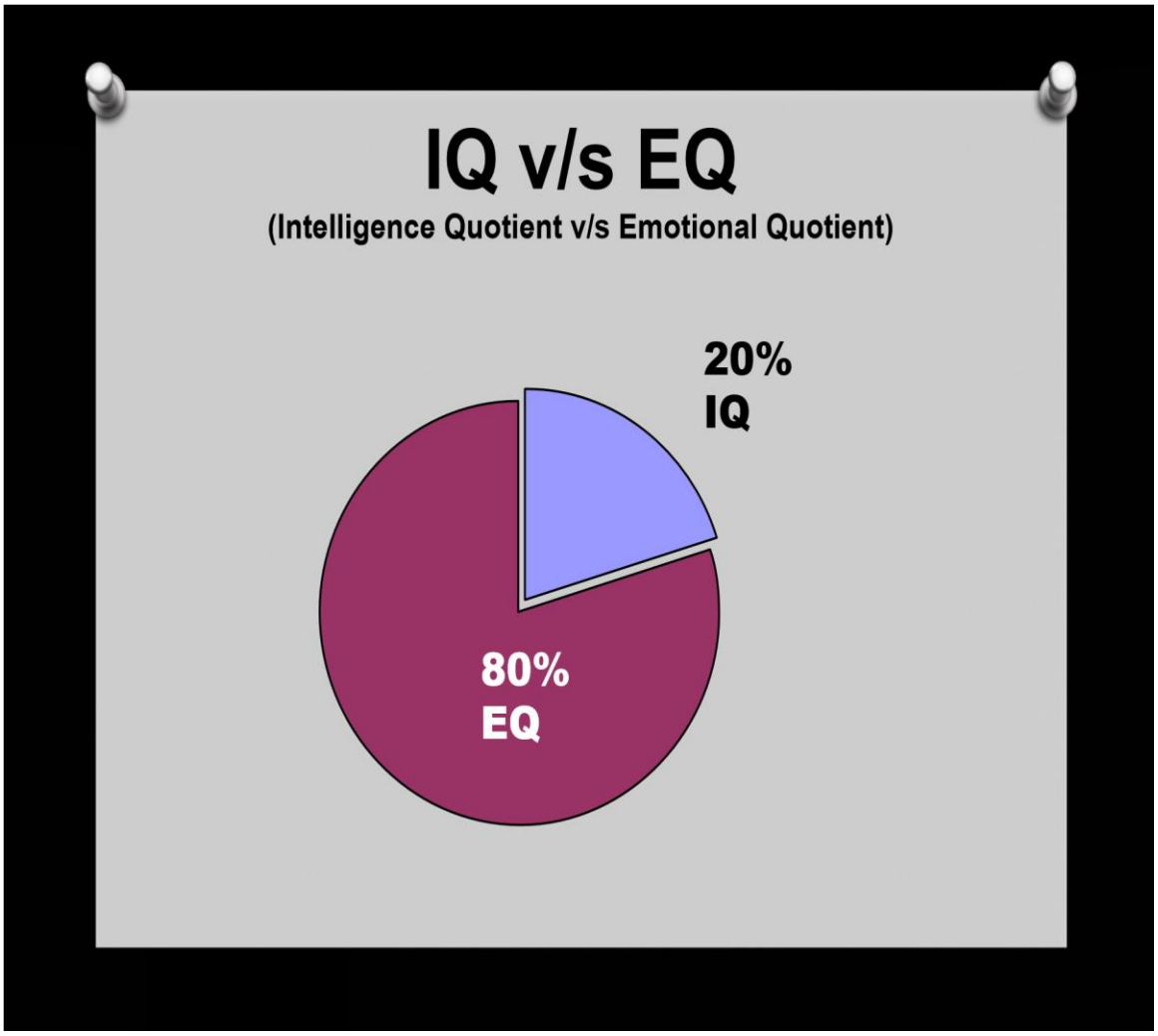
"there is no published study indicative of this trend and the commissioned and unpublished investigation that Goleman (1998) cites in support of this claim, does not include any measures of EQ ()" (Mathews, Zeinder and Roberts, 2002, p.13).

Although, it is believed that there is not clear model or construct which can clearly define and explain EQ/ EI to a lay man, there are many models available in the field of EI/EQ. However, two models that are relied on by the researchers for data collection and measuring emotional intelligence are;

Ability Model

Mixed Model

Ability Model primarily focus on the establishing a link and bridge between mental ability and emotional intelligence. Researchers who support this model do not consider emotional intelligence a separate entity. They count it as an integral part of intelligent quotient or mental ability. The diagram given below clearly demonstrates percentage wise difference of the star performers. Out of 100 times, 80 percent of star performers are people with enhanced emotional intelligence.



IQ versus EQ

Major Distinctions between IQ and EQ

Emotional Quotient	Intelligence Quotient
EQ is tested and evaluated EQ oriented standardized tests	IQ is tested and evaluated IQ oriented standardized tests
EQ level of a person is based on the scores in EQ tests.	IQ level of a person is based on the scores in IQ tests.

A person's real life success is based on intelligent quotient.	A person's academic success is based on intelligent quotient.
EQ measure a person's attitude, social and emotional competence and ability to maneuver in socially complex situations	IQ measures a person's reasoning ability and academic competence
EQ is an ability which is acquired and improved through exposure to real life situations.	IQ is an inborn trait and can be polished with training but cannot be acquired.
EQ is the ability to understand one's own emotions and steer its expression according to the situations.	IQ helps a person to learn and implement rational and reason based learning.
People with high IQ can very good leaders, managers and other key post holders.	People with high IQ can demonstrate exceptional academic and intellectual abilities but cannot be necessarily good leaders.

This Chart compares different aspects of emotional quotient and intelligence quotient.

Comparison Chart

Basis for Comparison	Emotional Quotient	Intelligent Quotient

Definition	Refers to the scores of an individual in specifically designed tests to measure emotional intelligence, it measure emotional competence of an individual	Refers to the scores of individual in specifically designed tests to measure reasoning and academic competence of an individual
Origin	Acquired and polished in life time	Inborn trait of every human being
Measure/ Tests	Emotional Competence / Intelligence	Academic competence/ Intelligence
Ability	Understand, appreciate and steer personal emotions and emotions of other people	The capacity to understand problems and content which require logical reasoning
Success	Helps a person to be successful in real life	Helps a person to be academically successful
Recognise	EQ recognizes a person's ability to live a successful life and take leadership role	IQ recognizes a person's ability to handle numbers and problems that need reasoning.

https://keydifferences.com/difference-between-iq-and-eq.html

Emotional Quotient/ Emotional Intelligence in the United Arab Emirates

As discussed earlier that since its birth in 1990's EI or EQ has created huge debates and discussions across the globe including the United Arab Emirates. In the UAE especially, people involved in leadership and management roles have been taking keen interest in EQ/EI. It has been one of the main areas of professional development trainings in the recent decade or so. It is believed that the leadership with higher EQ is more innovative, adaptive, productive and effective as compared to mechanical management of firms. We see that researchers like Richard E. Mayer and John Sweller have a handsome amount of literature in the field of EQ.

Although a lot has been read and discussed about EQ and IQ in the United Arab Emirates we come to know that with exception for a few very little research has been conducted in the field of Emotional Quotient.

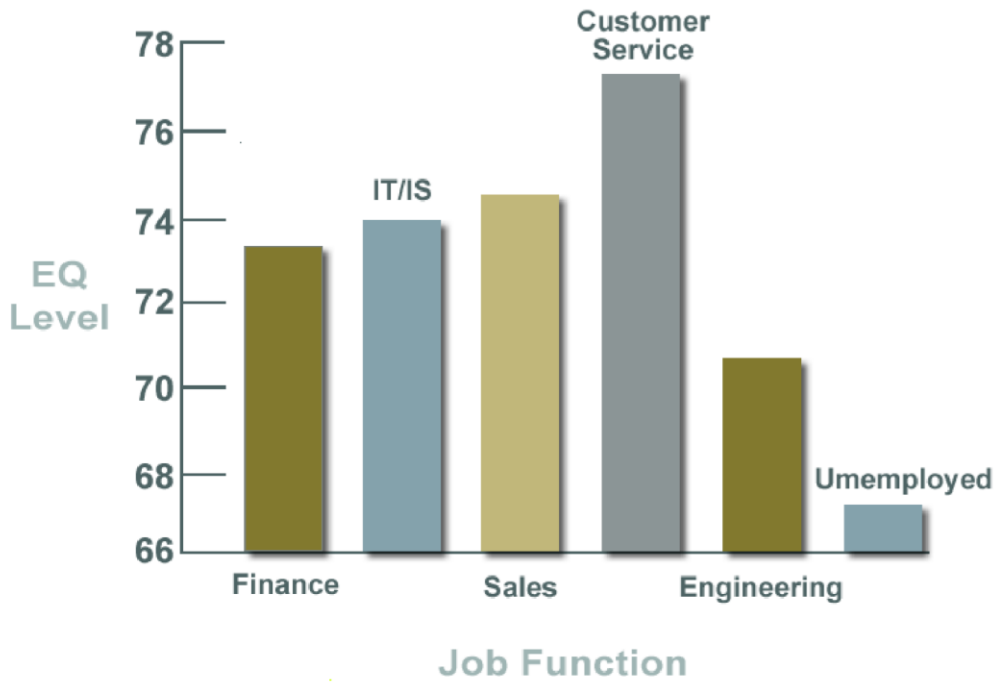
Madsen (2012) carried out a research project to investigate and inquire the impact that contemporary Arab women leaders consider were most significant in guiding them in their lives to get them ready for their future management jobs and tasks within the United Arab Emirates.

Stephenson (2010) from Zayed University conducted a study under the title ' *Researcher Development in UAE Classrooms: Becoming Teacher-Leaders'* added to the information pool meant for professional learning and analyzing teacher professional learning; the improvement and execution of a new model of teacher certified education; the recognition of a structure that has a prospective application in

other contexts; and assessment of an interdisciplinary application of teacher professional learning.

Statistics collected from different spheres of life approves the fact the Emotional intelligence has been one of the key element of professional development in recent year. Firms, companies and organizations want their employees and leaders to be emotionally intelligent so that they can understand and appreciate their own emotions and feelings and the feelings and emotions of their coworkers and subordinates.

Figure given below give us statistical information on the matter in hand.



'In *The Impact of Emotional Intelligence on Employee WorkEngagement Behavior: An Empirical Study'* the researcher has tried to prove that emotional intelligence has become very famous and effective instrument in organization to design a transparent

however, the design is obscure which leads us to the fact that there is ample need for conducting research in this field.

This study contributes to the literature by providing more information about "emotional intelligence, which may alleviate Work Engagement Behavior. It does this by building on the small existing pool of knowledge in order to extend the research on EQ. The expected outcome of this study was an increased understanding of how

EQ impacts on Work Engagement Behavior" (Ravichandran, ArasuandKumar, 2011, p.157).

In 2012, EI training was applied for police organizations by the criminal psychology department representing police organization under the tag *'An Exploration of the*

Relationship Between Emotional Intelligence and Job Performance in Police Organizations' which explains that, after the controlling of general mental abilities and personality traits, "EQ has been found to explain additional incremental variances in predicting police job performance" (Al-Ali, Garner and Magadley, p 1-2).

Importance of Emotional Intelligence:

Emotional Quotient or Intelligence has been one of the most important factor in different areas of life. The EQ concepts have been applied in training and development in these sectors;

- Education
- Health
- Police and Criminology
- Sales
- Customer Services
- Finance

- Engineering
- Information Technology

Considering the impact of applying emotional intelligence at a work place and real life can have impact on the following;

Professional and Personal Relationships. Emotionally intelligent people usually enjoy good relationship with people around them as they have the ability to under and appreciate of their colleagues and fellows around them. Emotionally intelligent people apply their information and learning to understand the emotions of people and steer and handle them accordingly. Such people always enjoy good rapport with their peers and bosses and have greater chances of escalation in their work place.

Performance at work. People with higher and active emotional intelligence have greater and better understanding of the feelings and emotions of their bosses and colleagues and thus have the chance to avoid conflicts and focus their potential on their official tasks. This results in improved performance in the work place.

Work Environment: Organizations have been spending money to get their staff trained in EQ to ensure smoother work environment at their work place.

Mental Health: People with improved emotional intelligence enjoy good health and do not allow people and environment around them to affect them negatively. In this way they avoid problems like anxiety and depression etc.

Physical health. Understanding of their own emotions can help people correctly manage them, and not being able to manage stress, for example, can have direct repercussions

for our physical health. Increased risk of heart attack, a speeded aging process, and higher blood pressure are just some of the physical consequences of stress. A way to start improving your EQ is learning how to identify and relieve tension.

Medical Field: An emotionally educated doctor is always a very successful one in treating his patients by appealing to their emotions.

Career Coaching: Career coaches with sound knowledge of emotional intelligence always leave greater impacts on people in deciding their careers.

There is a lot more that can been discussed, however, the researcher assumes that these examples will do the job. Having discussed all of the above we can clearly and boldly claim that emotional intelligence is one of the most important areas that can be researched into and applied in real life. This will help us develop a mentally and emotionally healthy society in the United Arab Emirates.

Mixed Emotions Towards Artificial Intelligence in UAE and the world.

The world at large and the UAE have mixed approach towards the idea of emotional intelligence. Same stands true for artificial intelligence. Although, developed by human beings these robots are feared and afraid of by human beings. They think that robots with artificial intelligence may take their jobs and control them like slaves.

Some people are against the idea of feeding artificial intelligence to robots.

Luc de Brabandere, a mathematician and senior advisor at the Paris office of Boston Consulting Group, argues thinks that the possibility of having artificial intelligence

working for us and executing our daily chores does not exist and should not exist at all. "He says the fact that machines can recognize faces does not mean they can find them beautiful. Also, a computer may have memory, but it won't be able to remember.

It could produce images, but that doesn't mean it has an imagination. He states that the uniqueness of human intelligence relies on the fact that it can not be copied or reproduced." (Luc de Brabandere 2000, P. 235)

Moreover, a lot of people are afraid that machines will substitute working folk in the industry. Their apprehensions may be right to some extent but they ignore the fact that if human intelligence is supplemented with artificial intelligence work output can increase tremendously. The best example of this virtual/ artificial nurses which do not really replace nurses practically. Virtual nurses actually help nurses and doctors help see more patients with enhanced attention, better treatment procedure and valid diagnosis. In this way doctors and nurses are able to take care of both physical as well emotional healths of the patients.

In the teaching field, teachers are assisted by artificial intelligence in teaching of ICT and IT which helps them reach out to different students at a time.

Call centres use artificial intelligence to help customer reach out to their respective services by following simple instruction given through machine or artificial

intelligence.

These is a just a single example that show that even though it is barely possible for human emotions and reactions to be replicated in machines, up until now, they have come closer to their understanding and replication of human emotions, which is a big step on the way to improving our quality of life with the help of AI.

The UAE govt. has shifted to the use of artificial intelligence in many fields including driverless cars and public transport taxies. The future of artificial intelligence is very bright in the United Arab Emirates.

The Future of Artificial Emotional Intelligence

AEI is relied upon to change numerous aspects of our lives, and there are a wide range of suppositions of what kinds of gadgets or projects could be created. They go from ones that may appear to be totally unimaginable, to others that are really being grown right currently despite the fact that they sound insane.

Innovation magazine WIRED distributed an article on their forecasts on 4 creations of what they call "enthusiastic innovation". These incorporate a Mood Reader that will give official rundowns of the genuine sentiments of an individual, saving us various battles and mistaken assumptions in our connections; a Spouse Finder, which would probably coordinate us with our best accessible fit anyplace on the planet; a versatile self-information holistic mentor gadget by the name of Socrates that will push us to be simply the best; lastly a Career Locator that will comprehend our actual work potential, coordinating it with the prerequisites of the economy. Every one of these gadgets are relied upon to be created in the following 50 years or somewhere in the vicinity, in the

event that we center our endeavors around the greater wellsprings of discontent: our passionate life.

In the movement business, AI has likewise had some advancement over the most recent couple of years. Hitlist, for instance, has the mission to help individuals travel more for less. The organization's CEO, Gillian Morris needs to help individuals travel more, and make their treks progressively effective and pleasant, giving the best costs, times and places to travel. Like Hitlist, a lot more organizations are attempting to make the movement experience better, with increasingly more customization. Counterfeit Emotional Intelligence would be an incredible in addition to here, by perceiving individuals' suppositions and feelings with respect to various goals and exercises, so the AI can offer the most reasonable touring plans for every individual.

There is much potential for AEI in various businesses, and despite the fact that it isn't in all respects likely that we will almost certainly duplicate people's passionate insight, getting machines to perceive feelings is an immense advance towards giving better administration and encounters to clients in an assortment of enterprises.

Conclusion:

There is a squeezing requirement for preparing in this field in the UAE setting and new wave. There are only a great deal of things that machines can show improvement over people, and we shouldn't be too glad to even think about admitting it. Numerous talented employments pursue a similar general work process:

Gather information

Analyze the information

Interpret the outcomes

Determine a prescribed game-plan

Implement the strategy

Those that need to remain significant in their callings should concentrate on abilities and capacities that man-made brainpower experiences difficulty duplicating — getting, propelling, and connecting with individuals. A savvy machine may almost certainly analyze a disease and even prescribe treatment superior to a specialist. It takes an individual, be that as it may, to sit with a patient, comprehend their life circumstance (funds, family, personal satisfaction, and so forth.), and help figure out what treatment plan is ideal.

It's these human abilities that will turn out to be increasingly more prized throughout the following decade. Aptitudes like influence, social comprehension, and compassion will progress toward becoming differentiators as man-made brainpower and AI assume control over our different errands. Lamentably, these human-situated abilities have by and large been seen as second need regarding preparing and instruction thus we have to feature it for the UAE to continue pushing ahead and ahead to what's to come.

References

- Measuring emotional intelligence with the Mayer-Salovery-Caruso Emotional Intelligence Test (MSCEIT) Mayer &Salovey, 2000, p.196
- An EI-Based Theory of Performance From the book The Emotionally Intelligent Workplace Edited by: Cary Cherniss and Daniel Goleman Now available through Amazon.com CHAPTER THREE By: Daniel Goleman (Goleman (1998, p.93)
- Emotional Intelligence John D. Mayer, Peter Salovey, David R. Caruso, and Lillia Cherkasskiy Mayer, Salovey and Caruso (2000, p.396).[3]
- https://www.talentsmart.com/articles/Why-Attitude-Is-More-Important-Than-IQ-982658569-p-1.html Talent Smart Inc., 2017, p.1).[4]
- A new approach to assessing leadership dimensions, styles context Sadri 2012, p. 535
- Dick, W., Carey, L. & Carey, J.O. (2005). The Systematic Design of Instruction (6th ed.). MA: Pearson Education.
- Donmez Mehmet & Cagiltay Kursat (2013) A Review and Categorization of Instructional Design Models, Middle East Technical University Turkey P.1
- De Paolis, L., Mongelli, Antonio, &SpringerLink. (2015). Augmented and Virtual Reality [electronic resource] : Second International Conference, AVR 2015, Lecce, Italy, August 31 – September 3, 2015, Proceedings (1st ed. 2015. ed., Lecture Notes in Computer Science, 9254).
- Gagne, R.M. (1962) Introduction. In R.M Gagne (Ed). Psychological principles in system development . New York: Holt, Rinehart & Winston

- Joeng.(2002) Fundamentals of Photonics. Basic Principals and Application of Holography Lake Forest College Lake Forest, Illinois P. 381
- Hattie, John. (2009) Visible Learning: A Synthesis of over 800 meta-analyses relating to Achievement. New York: Routledge.
- Khalil KM 2016) Applying learning theories and instructional design models for effective instruction University of South Carolina AJP Advances in Physiology Education 40(2):147-156
- Reiser,R.A. (1987) Instructional technology: A history. In R.M. Gagne (Ed). international technology foundations. Hillsdale.NJ Lawrance Erbaum
- Rachid Benlamri Can E-Learning Be Made Real-Time? August 2006 *with*34 Reads DOI: 10.1109/ICALT.2006.1652432 · Source: IEEE Xplore Conference: Sixth International Conference on Advanced Learning Technologies, ICALT'06, 2006.
- Şimşek, A. (2013). Öğretim Tasarımı ve Modelleri. In K. Çağıltay & Y. Göktaş (Eds.), Öğretim Teknolojilerinin Temelleri: Teoriler, Araştırmalar, Eğilimler (1st ed., pp. 99–116). Pegem Akademi.
- Skinner, B.F. (1953) Science and human behaviour. NewYork: Mcmillan
- Smith, P.L. & Ragan, T.J. (2004). *Instructional Design* (3rd ed.) Wiley Jossey-Bass Education, NY: Merill.
- Mehmet Donmez &Kursat Cagiltay (2016) A Review and Categorization of Instructional Design Models *E-Learning Washington DC 2016 P.370*
- Robin A. Walker Program (2012) Hologram as Teaching Agents, 9th International Symposium on Display Holography (ISDH 2012) Teachers College, Columbia University, New York, USA P. 4

TITLE: Evaluation of **Technest** Workforce Skills Program at San Jose City College

DESCRIPTION: There is an imperative to fill the need for effective randomized controlled trials (RCT) evaluations in community colleges. We present a well-designed, collaborative RCT to address the need of community college program evaluations in the effectiveness database. As San Jose City College currently does not have the resources to become familiar with these standards without decreasing their priority to other valuable initiatives, they have partnered with Excalibur Education Group and the Bellwether College Consortium to support the effectiveness evaluation for the **Technest** workforce development training program.

Importance.

Silicon Valley continues to be an engine of economic growth and job creation. The stark reality, however, is that many local communities have no viable options for acquiring the skills to thrive in the high-tech economy. This situation has led to a growing economic inequality that is evidenced by indicators such as long-term unemployment, income inequality, and outright poverty. Due to the current climate of under-qualified local workers for high-paying Information Technology (IT) jobs, more than 68 percent of the region's employers report significant difficulties in securing the talent they need, a situation that has led to soaring foreign worker visa applications. Although the region's community colleges offer computer science and IT degrees and certificates, it may take students several years to complete these programs, due to the need to work part- or full-time. This obstacle can easily lead to graduates entering the job market with outdated skills. There is a pressing need for accelerated certification programs, such as **Technest**, for community college students, and for job seekers who need to update their skills to enter or re-enter the workforce.

Program Information.

To address these issues, San Jose City College (SJCC) has implemented and completely funded **Technest** – a workforce development training program that offers affordable, high-quality training for in-demand jobs in technology. To date, SJCC has collaborated with the MIT Exchange program (MITX) to develop an 18-week coding academy that focuses on Python, the primary coding language requirement for more than 22,000 jobs in the 12-county Bay Area region. SJCC has also received full articulation for a 16-week, 3-unit UC Berkeley course in Foundations of Data Science that teaches critical concepts and skills in computer programming and statistical inference, with hands-on analysis of realworld datasets. SJCC has joined Integrated Device Technology (IDT) to offer a 10-week, 2-unit directed study course that is focused on the Internet of Things (IoT) – the new world of devices that can communicate with each other. The course trains students to program and gather data using single-board computers to drive IoT-related devices.

To help its graduates find jobs while their skills are still current and fresh, SJCC has formed partnerships with IDT, the San Jose Silicon Valley Workforce Development Board, Work2future Foundation, and the Workforce Innovation and Opportunity Act to offer career counseling, job development, and on-the-job training placements. Students can also choose to receive entrepreneurial training through **Technest** to prepare themselves to participate in the growing "Gig Economy" of short-term and temporary freelance jobs. More than 200 students are presently enrolled in three **Technest** program tracks: coding, data science, and IoT, with a projected 85% completion rate – more than double the present industry pass rate for accelerated coding programs. **Technest** will increase access to IT careers for underrepresented groups by providing low-cost training, thereby expanding economic opportunities and financial stability for its graduates. Additionally, the **Technest** program has been recognized with great potential by Microsoft and the Gates Foundation to maintain, enhance, and scale up their program to address. To this end, the SJCC administration dedicates itself to the support of a full and highest quality RCT effectiveness evaluation to document program success, impact, areas for improvement, and opportunities for scale up.

Study Design.

This evaluation will determine the effectiveness of the **Technest** program at San Jose City College, as well as document and describe the transition from the college's standard practice of traditional workforce skills training. The primary research questions to be answered are:

1. Do students who receive workforce training through the **Technest** program persist from the fall to spring semester at higher rates than students who do not go through the **Technest** program?

2. Do students who receive **Technest** workforce training pass more of their classes than students who do not go through the **Technest** program?
3. Do students who receive **Technest** workforce training successfully pass their coding certifications at a higher rate than students in traditional coding education who do not go through the **Technest** program?
4. Do students who receive **Technest** workforce training enter the workforce with coding jobs at a higher rate than students in traditional coding education who do not go through the **Technest** program?

Secondary descriptive data will be examined to measure the impact on the subsamples of interest, including firstgeneration college students, racial/ethnic minorities, males versus females, students from low socio-economic backgrounds, and academically underprepared students. Additionally, instructors will be asked about the unique features of the **Technest** program, their satisfaction with the training and skills received by **Technest** graduates and their readiness for a career in information technology.

Participants.

The study will take place with one community college located in the western United States. The community college has an average yearly academic enrollment of 11,000 full-time students. In 2016, 45% of students were registered as full-time students. Just over half of the students were female (60%). Students are racially/ethnically varied with 49% being white, 35% being Black or African American, and 6% being of unknown ethnicity. The average age of students at the college was 24.8 (SJCC, 2019). Students included in the study sample will be first-time students who have applied to the **Technest** program to the college of fall 2019. The **Technest** program accepts more than 200 new students each fall.

Assignment.

The study will employ a randomized controlled trial (RCT) design. Students who are accepted as first-time, fulltime students into the **Technest** workforce training program for the fall 2019 semester will be randomly accepted or delayed entrance into the **Technest** program. Students will be informed of the research study and asked to participate during his or her orientation session. If the student agrees to participate they will be asked to sign the consent form. If a student does not provide consent, he or she will be dropped from the study sample. Students who agree to participate and provide consent will be entered into an Excel spreadsheet and given a number using the random number generator. All students who receive an even number will be accepted into the **Technest** intervention group and those who receive an odd number will be assigned to the delayed-entrance control group.

Power.

The effective sample size is adequately powered at .80 or higher at α = .05 two-tailed test. Power estimates for the evaluation study were done with the model 1.0: MDES Calculator for Individual Random Assignment (IRA) DesignsCompletely Randomized Trials in the Dong and Maynard PowerUp! Program (Dong & Maynard, 2013). PowerUp! is a tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasiexperimental designs. Conservative power estimates, taking into account attrition, producing the minimum detectable effect size was computed using the above discussed sample size. Half of the sample will be randomized to the intervention and the other half to control. The percent of variance explained by covariates was estimated based on prior research using a similar sample and outcomes (Jamelske, 2009). With this sample the minimum detectable effect size is found to be 0.147. In similar studies examining the effectiveness of programs for first-year students significant effects on credit accumulation and academic achievement were found with an effect size of .23 and .14, respectively (WWC, 2016). See Appendix # for power details.

Data.

The primary outcome variables will include fall-to-spring persistence, percent of classes passed each semester, semester and cumulative grade point averages, coding certification pass rates, and workforce/successful employment rate.

Persistence is defined as a student who is still enrolled at the college in the spring after completing the fall semester. These outcomes will be assessed from data received from the community college and follow-up data collection from the program graduates.

In order to assess the outcomes **student class records** and **enrollment information** will be collected from the college. The data set will be built from data extracted from the college's student information system, by the Office of Research, Planning, and Institutional Effectiveness. Demographic data will be requested for all students applying to the **Technest** program in the fall of 2019. Baseline information, including a measure of student socioeconomic status (Pell grant eligibility) and a continuously-scaled measure of previous academic achievement (student's high school GPA), will be collected from the college in order to assess baseline equivalence of the analytic sample to ensure successful randomization. Additional demographic data such as gender, ethnicity, first-generation student status, and age will also be collected.

Table 1. Student demographic and educational variables

Variable	**Description**
Male indicator	Gender: 1 = male, 0 female
White indicator	Race/ethnicity: 1 = white, 0 else
Age	Age in years
Low income	Low income: 1 = yes, 0 else
First-generation student	First-generation student: 1 = yes, 0 = else
High school GPA	Cumulative high school GPA
Persisted	Persisted: 1 = yes, 0 else
Percentage of classes passed	Percent of classes passed
Semester GPA	GPA after fall 2018 or spring 2019
Cumulative GPA	Cumulative GPA after 2018-2019 academic year

Secondary. Descriptive data will be collected through brief semi-structured interviews with all or a purposive subsample depending on number and cooperation of students and employers. After transcription, interview responses will be compared to find common themes or ideas, which will then be grouped into categories (Strauss & Corbin, 1998). This process will continue until all responses are coded. College administrators and **Technest** instructors can be provided with the interview transcripts, and a summary of the results for member checking.

At the end of two years of the evaluation, **Technest** instructors and students will be contacted to schedule a short interview about the **Technest** workforce development program. Interviews should take no more than 30-40 minutes. A preliminary interview protocol is included in appendix #.

All researchers on this study have received training on best practices for research with human subjects. This study has received approval from the SJCC Institutional Review Board (IRB) to ensure compliance with the institution's policies and procedures on ethics for human subject research. (Attached in Appendix ??.)

Data Analytic Strategy.

The primary focus of the study is to determine the degree to which the **Technest** workforce preparation program impacts student persistence and academic achievement. Because some of the outcomes are dichotomous and others are continuous both logistic regression and ordinary least squares (OLS) regression models will be estimated.

Persistence, *P*, is a dichotomous outcome meaning that the dependent variable is an indicator equal to 1 if the students persisted from fall 2019 to spring 2020 and 0 if not. Because the dependent variable is a dichotomous outcome logistic regression will be used to estimate the effects. The maximum likelihood logit estimation is defined as: $PPPP_{RRii=1} = \frac{e^{P_i}}{1+e^{P_i}}$; $P_i = \alpha + T_i\beta + ED_i\delta$; *R*= 1 signals that student *i* persisted, while *T* indicates participation in the **Technest**

program during the 2019-2020 academic year, and *ED* represents educational and demographic variables. From this equation, the odds ratio will be determined. Similarly, the coding certification pass rate, C, will be calculated using the same statistical methodology. The Wald statistic will also be calculated to test for statistical significance.

The impacts of the **Technest** program will be estimated for both semester and cumulative grade point average using ordinary least squares regression where: $GGPPGG_{ii} = \alpha\alpha + TT_{ii}\beta\beta + EEEE_{ii}\delta\delta$. *GPA* will represent the grade point average (either semester or cumulative) for student *i* while *ED* represents the educational and demographic variables. The intervention coefficient, standard deviation, and statistical significance will be examined.

Cohen's *d*, a measure of the size of the difference between the two groups, will be calculated and reported for each outcome. All analyses will be run using IBM SPSS Statistics 24 software. ***Implementation Fidelity.***

Though the most direct way to determine whether advisors were using appreciative advising would be to observe advisors and see whether they were using the components of appreciative advising or not. However, student privacy concerns make this an inappropriate option. Alternatively, advisors will be asked about their implementation of appreciative advising in the post-interview. Advisors will be asked about how they changed their advising practice, what a

DURATION OF PROJECT AND PROJECT TIMELINE:
The entire research project for the randomization and effectiveness evaluation of two cohorts will span 36-38 months.

YEAR 1:

Tasks / Personnel Effort in Hours	PI	Co PI TW	Co PI LT	Co PI/OIC	SJCC
Research Design Presentation to SJCC Administration and Program	40	60	24	60	16
Participant Data Collection – Initial Cohort		60	8	80	40
Participant Assignment – Initial Cohort	80	60	16	60	24
Participant Data Collection – Cohort 1, Semester 1 (Fall 2019)	32	52	16	80	8
Methodology Fidelity Check	40	28	16	12	8
TOTAL EFFORT IN HOURS	240	312	96	384	128

YEAR 2:

Tasks / Personnel Effort in Hours	PI	Co PI TW	Co PI LT	Co PI/OIC	SJCC
Participant Data Collection – Second Cohort		20	12	40	20
Participant Assignment – Second Cohort	40	32	8	60	16
Participant Data Collection – Cohort 1, Semester 3 (Fall 2020)	16	28	8	52	8
Participant Data Collection – Cohort 2, Semester 1 (Fall 2020)	24	36	8	60	6
Methodology Fidelity Check	8	15	8	14	8
Participant Data Collection – Cohort 1, Semester 4 (Spring 2021)	16	28	8	60	8
Participant Data Collection – Cohort 2, Semester 2 (Spring 2021)	24	40	16	52	12
Methodology Fidelity Check	10	16	4	16	8
TOTAL EFFORT IN HOURS	138	215	72	354	86

YEAR 3:

Tasks / Personnel Effort in Hours	PI	Co PI TW	Co PI LT	Co PI/OIC	SJCC
Preliminary Analysis and Report – Cohort 1 (Fall 2021)	80	100	40	120	8
Participant Data Collection – Cohort 2, Semester 3 (Fall 2021)	32	32	8	72	8
Methodology Fidelity Check	8	16	8	16	8
Participant Data Collection – Cohort 2, Semester 4 (Spring 2022)	48	68	8	72	16
Preliminary Analysis and Report – Cohort 2 (Summer 2022)	80	100	40	120	8
Methodology Fidelity Check	16	20	8	20	8
Final Analysis and Report (Summer 2022)	88	145	40	172	24
TOTAL EFFORT IN HOURS	352	481	152	592	80

Dissemination of Results.

Upon completion of the data analysis of the first year, the research team plans to write a preliminary research report indicating the intervention, research design, and first year outcomes. The report will be submitted to ERIC for publication to the higher education community at-large, the Community College Journal, and the Community College Journal of Research and Practice to reach the leaders of the community college field. Additionally, this research report will be made available on the Arnold Foundation and Excalibur Education Group, LLC websites for open source dissemination and links to this report will be sent to online publications, such as the Chronicle for Higher Education, InsideHigherEd, the American Association of Community College's CCDaily, and other major higher education online sources.

This initial report will be the basis for presentations to the National Council for Continuing Education and Training (NCCET) national conference, the Council for the Study of Community Colleges (CSCC) national conference, the American Association of Community Colleges national conference, the Achieving the Dream conference, and the Community College League for Innovation conference.

The completion of the final report will initiate an update to the reporting outlets mentioned above for the dissemination of the written report in publications, journals, and online media for higher education audiences. This final report will also initiate a presentation to the San Jose City College community at-large and a follow-up presentation to the national conferences listed above. Additionally, presentations will be proposed to the Society for Research in Educational Effectiveness (SREE), the Association for the Study of Higher Education (ASHE), and the American Education Research Association (AERA) to reach other educational research methodologists to promote quality research and open consideration in for those researchers to the community college field, which is often overlooked for research designs.

Finally, both the initial report and the final report will be submitted to the WhatWorksClearinghouse helpdesk to request a review of the research. Upon review, this study will be added to the WWC database for dissemination to anyone looking for RCT research on community college workforce program interventions to use as evidence in future grant proposals. It will be requested that this research be reviewed as a Single Study Review or as part of an Intervention Report on Workforce Preparation Programs under the Supporting Postsecondary Success Review Protocol. Upon review and the receipt of a WWC rating, this research report will also be disseminated to the online association of community college resource development officers for use in future grant proposals as the highest level of evidence. ***Appendix #. Power Analysis.***

Model 1.0: MDES Calculator for Individual Random Assignment (IRA) Designs—Completely Randomized Controlled Trials

Assumptions		**Comments**
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
P	0.50	Proportion of the sample randomized to treatment: $n_T / (n_T + n_C)$
R^2	0.04	Percent of variance in outcome explained by covariates
k^*	6	Number of covariates used
n (Total Sample Size)	1400	
M (Multiplier)	2.80	Computed from T_1 and T_2
T_1 (Precision)	1.96	Determined from alpha level, given two- or one-tailed test
T_2 (Power)	0.84	Determined from given power level
MDES	**0.147**	Minimum Detectable Effect Size

Preliminary Interview Protocol

All questions preceded by a *, are prompts that could be used when necessary.

Thank you so much for taking the time to meet with me. As you know, we are interested in the effects of the Technest program on workforce development coding students at San Jose City College. I will audiotape our conversation today. Do I have your permission to do that?

How long have you been involved in the Technest program? In what capacity?
Were you given specific information about the Technest program when you started?
How different do you think coding program instruction under the Technest program was to what you did before?
Did you find anything difficult about the transition to the Technest program?
Did the college or your colleagues provide any helpful resources for the transition? Or is there anything that should have been provided?
What does a typical Technest coding course meeting look like?
*If no typical session: Can you give me a couple examples of recent sessions?
What are the most important and unique components of the Technest program?

Companies and Customers Co-Creation Value: A Conceptual Measurement Model Creating and Using Big Data

Dr.Muneer Abbad, *Community College of Qatar, Qatar*

Dr.Faten Jaber, Regent's University London, UK

ABSTRACT

The main aim of this study is to build a conceptual model of measuring value cocreation through big data creators (customers) and users (companies) and its outcomes (benefits) for both players based on service-dominant (S – D) logic. By developing a conceptual model uses big data that was the main factor to transfer the concept of value in marketing to value co-creation, the proposed model of this study comes to connect the two players (customer and company) as creator and user of big data to confirm the S -D logic concept. Co-creation as construct in the proposed model could be measured using different theoretical and practical models that exist in literature. The model proposed in this study is ready to be empirically tested by collecting data in different contexts to validate the model and test the hypotheses that proposed in the model.

Key words: value co-creation; conceptual framework; marketing; big data, S – D logic.

Introduction

Organizations have the opportunities to use advanced information and communications technologies (ICTs) to reach and communicate with their customers. Also, customers have the same opportunities by using the advanced ICTs to influence and communicate with their organizations. Customers are playing big role now in participating their organizations in providing products and services which results to change the traditional role of customers (buyers) in value-in-exchange concept to be co-creator of value with their companies (Prahalad & Ramaswamy, 2000). According to this view, Grönroos (2008, p299) mentioned that "value is not created by the provider but rather in the customers' value-generating processes". So, the starting point is customers who buy goods or service and create value for themselves (value-in-use) and the company is considered as creator of the value (Grönroos, 2008). In co-creation context, the value-in-use is not just important for the customer, but it is very important to the companies and it is difficult to manage and measure by providers. In summary, customers have been given the role of co-creation of value and companies facilitate the value creation process by providing the value foundation required and the better companies manage to facilitate value, the more value-in-use will be created, and eventually, the higher the value-inexchange will be (Grönroos, 2008).

Prahalad & Ramaswamy (2000) define co-creation as "a form of market or business strategy that emphasizes the generation and ongoing realization of mutual firm-customer value. It views markets as forums for firms and active customers to share combine and renew each other's resources and capabilities to create value through new forms of interaction, service and learning mechanisms" (p.1). McColl-Kennedy et al. (2012) define value co-creation as "a benefit achieved from integration of resources through activities and interaction with collaborators in the customer's service network" (p.1). Accordingly, goods-dominant (G – D) logic that see customers as passive assets was changed to service-dominant (S – D) logic that suggest both companies and customers can co-create value (Vargo & Lusch, 2004; Lusch and Vargo, 2014). Payne, Storbacka and Frow (2008) develop a conceptual framework for understanding and managing value co-creation and provide a structure for customer involvement that takes account of key foundational propositions of S-D logic and places the customer explicitly at the same level of importance as the company as cocreators of value (Figure 1).

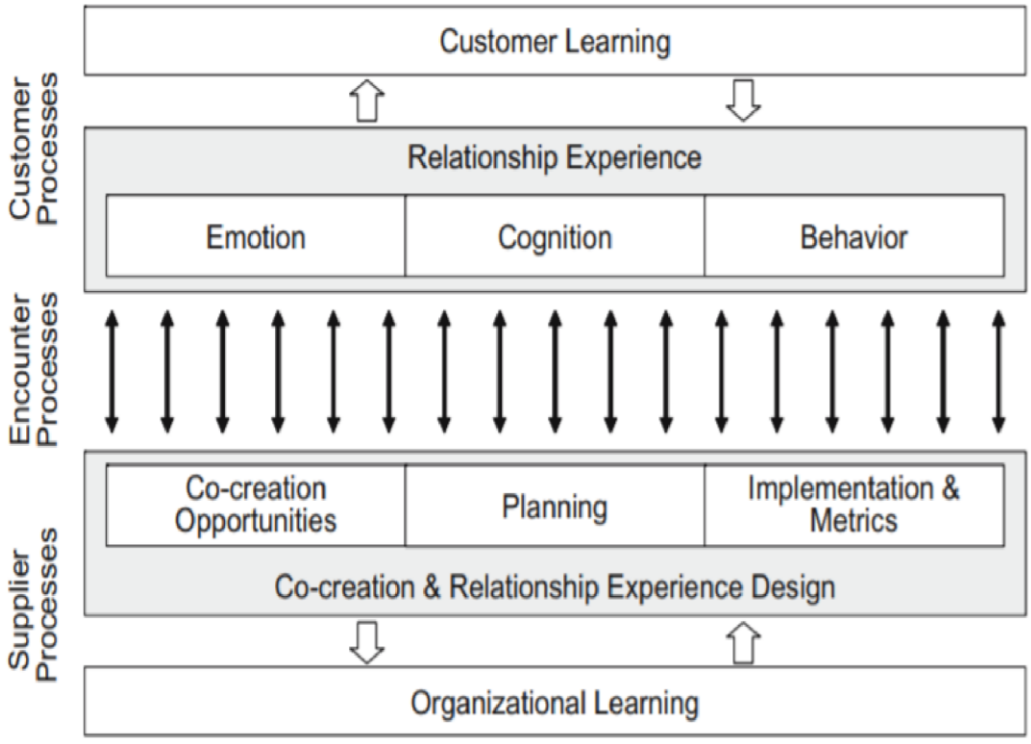


Figure 1: A conceptual framework for value co-creation

(Source: Payne, Storbacka and Frow (2008))

In our digital age, the primary driver that change the value creation to value co-creation between companies and customers is big data (Jagadish et al., 2014; Jobs et al., 2015). Customers behavior and location, for example, create big data and companies use these big data by analyzing and improving their strategic decisions (Tirunillai and Tellis, 2014; Clarke, 2016).

Based on S-D logic concept in value co-creation, this research focus on how customers and companies play their roles to create (by customers) and to use (by companies) the big data for value co-creation and outcomes (benefits) from this value for both players.

Companies can sense, capture, respond to market changes, and manage the big data that produced by customers by using big data-related technologies (Xie et al., 2016). The main 3-Vs characteristics of the big data that generated by customers are: volume, velocity, and variety. There are mainly two types of big data: structured (e.g. answering some questions) and unstructured (e.g. need process to be usable). Big data platforms are the key for the firms to co-create value with customers. Xie et at. (2016, p. 4) specified these platforms as "Big data platforms range from common online transactional platforms (transactional exchanges), virtual social networking platforms (consumer community communication), opendesign platforms (customer self-help design), and mobile interaction platforms (firm–customer communication)".

In summary, the main goal of this research is to build a conceptual measurement model of creating big data from customers and using these data by companies to construct the co-creation value based on service dominant (S-D) perspective. The proposed model will help to transfer the concept of value in marketing to value co-creation that depends on two players: customers as creators for the big data and companies as users for the big data that collected from customers.

Proposed Model

In value co-creation process, Xie et al. (2016) discuss how big data was transformed from resource to cooperative assets and identify four types of big data information resources (transactional, participative, communicational, and transboundary) that created by customers and how companies use and utilize these big data information resources. In summary, companies provide customers with the required technologies, software, and platforms to collect huge data from customers who use

these resources in value co-creation context. For example, customers' orders online are considered transactional information resources that saved and analyzed by companies, customers' communications with their companies about the orders are considered communicational information resources to be saved and analyzed for future use by companies, and customized design for products by customers considered as participation resources that analyzed by companies for identifying customers' needs and wants (Xie et al., 2016). Based on that, customer's role is big data creator and company's role is big data user and platforms provider in value co-creation to generate positive outcomes (benefits) for both customers (satisfactions) and companies (performance) (see Figure 2).

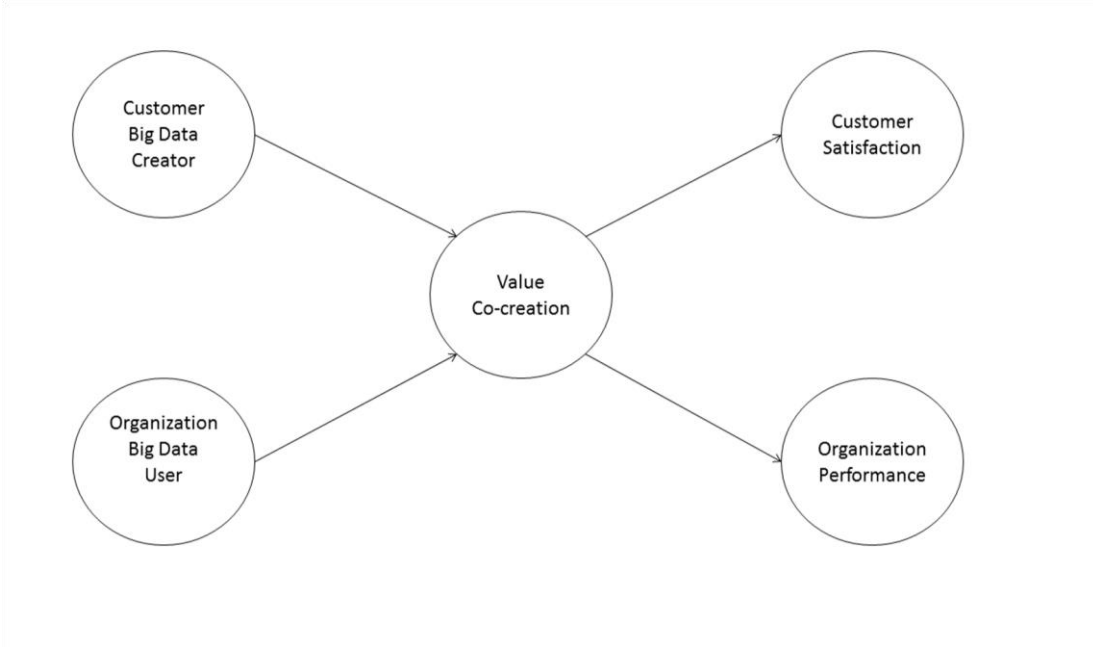


Figure 2: The Proposed Model

Taking the above discussions in consideration, the question is: how we can measure value co-creation?. Many studies are trying to answer this question and they proposed theoretical and practical models to measure cocreation value. For example, Neghina et al. (2015) proposed six dimensions or types of action (individualizing actions, relating actions, empowering, ethical, developmental, and concerted actions) shared by customers and companies to find out the antecedents of value co-creation. Tregua et al. (2015) proposed four dimensions of analysis of value cocreation: sharing, engagement, brand meaning, and awareness. McCollKennedy et al. (2012) identify eight activities in healthcare context

to measure value co-creation: co-learning, changing ways of doing things, cooperating, combining complementary therapies, connecting, collating information, co-production and engaging in cerebral activities. Finally, Tommasetti et al. (2017) proposed eight complex activities as a conceptual model for the measurement of customer value co-creation behavior: the combination of complementary activities, changes to habits, cerebral activities, cooperation, information research and collation, co-production, co-learning, and connection.

Methodology

Qualitative and quantitative methods will be used to collect the required data to test the proposed model. Interviews will be held with managers, marketing managers, and big data specialists to validate the four types of big data information resources (transactional, participative, communicational, and transboundary) that proposed by Xie., (2016) and to choose the best proposed model (or mixed models) to measure the cocreator value from the models described in the previous section. After finalizing the measurement items, a survey will target the major service companies in UK (Financial sector, Hotels, Communications, Insurance, …etc). The target sample will be managers, marketing managers, and big data specialists to collect the required data to test the proposed model.

In this research, qualitative data is used for the following reasons. Firstly, to know the types of big data information resources that used by companies. Secondly, the outcomes of this phase will be used to construct the modified and final proposed model. Thirdly, group interviews are used to identify key items (subcategories or themes) that will be used to develop items for potential inclusion in a survey questionnaire.

Categories were identified to classify the data within different categories. Based on the indicators used in the literature, the main subcategories were given to the coders to explain the categories in much more detail. Accordingly, the coders identified the subcategories by grouping similar themes associated with the main categories. Moreover, these data are tested and validated in the second phase by using confirmatory factor analysis techniques (CFA) on larger samples.

Conclusion

This research discusses how customer as big data creator and companies as big data user develop value co-creation and the outcomes (benefits) for both players. A conceptual model was developed based on servicedominant

(S – D) logic and Xie et al. (2016) study about big data creator and user concepts. In addition, the paper discusses the main theoretical and practical models that used to measure value co-creation to be based to empirically test the conceptual model.

References

Clarke, R., 2016. Big data, big risks. *Information Systems Journal*, *26*(1), pp.77-90.

Grönroos, C., 2008. Service logic revisited: who creates value? And who co-creates?. *European business review*, *20*(4), pp.298-314.

Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R. and Shahabi, C., 2014. Big data and its technical challenges. *Communications of the ACM*, *57*(7), pp.86-94.

Jobs, C.G., Aukers, S.M. and Gilfoil, D.M., 2015. The impact of big data on your firms marketing communications: a framework for understanding the emerging marketing analytics industry. *Academy of Marketing Studies Journal*, *19*(2), p.81.

Lusch, R.F. and Vargo, S.L., 2014. *Service-dominant logic: Premises, perspectives, possibilities*. Cambridge University Press.

McColl-Kennedy, J.R., Vargo, S.L., Dagger, T.S., Sweeney, J.C. and Kasteren, Y.V., 2012. Health care customer value cocreation practice styles. *Journal of Service Research*, *15*(4), pp.370-389.

Neghina, C., Caniëls, M.C., Bloemer, J.M. and van Birgelen, M.J., 2015. Value cocreation in service interactions: Dimensions and antecedents. *Marketing Theory*, *15*(2), pp.221-242.

Payne, A.F., Storbacka, K. and Frow, P., 2008. Managing the co-creation of value. Journal of the academy of marketing science, 36(1), pp.83-96.

Prahalad, C.K. and Ramaswamy, V., 2004. Co-Creating Unique Value with Customers. *Strategy & Leadership*, 32(3), pp. 4-9.

Tirunillai, S. and Tellis, G.J., 2014. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), pp.463-479.

Tommasetti, A., Troisi, O. and Vesci, M., 2017. Measuring customer value co-creation behavior: Developing a conceptual model based on

servicedominant logic. *Journal of Service Theory and Practice*, *27*(5), pp.930950.

Tregua, M., Russo-Spena, T. and Casbarra, C., 2015. Being social for social: a co-creation perspective. *Journal of Service Theory and Practice*, *25*(2), pp.198-219.

Vargo, S.L. and Lusch, R.F., 2004. Evolving to a new dominant logic for marketing. *Journal of marketing*, *68*(1), pp.1-17.

Xie, K., Wu, Y., Xiao, J. and Hu, Q., 2016. Value co-creation between firms and customers: The role of big data-based cooperative assets. *Information & Management*, *53*(8), pp.1034-1048.

BIG DATA AND THE REAL ESTATE MARKET: BENEFITS FOR PUBLIC ADMINISTRATIONS AND CORPORATIONS

Caterina Marini, Vittorio Nicolardi, Elisabetta Venezia
Department of Economics and Finance - University of Bari Aldo Moro
caterina.marini@uniba.it, vittorio.nicolardi@uniba.it, elisabetta.venezia@uniba.it

ABSTRACT

The problem of dealing with enormous databases is undoubtedly the new challenge that the scientific world needs to face to analyze a reality in continuous and fast development. In this paper, we focus the analysis on the Italian real estate phenomenon and how the administrative data are powerful in adding new information to the official statistics. The importance of the analysis we yield in this work is unique and original in its attempt to underline the opportunity that public management and real estate business can afford through a complete and harmonized data warehouse that needs Big Data analytic practice to be yielded. Our work is the first attempt in this sense.

KEYWORDS

Big data; Real estate values; Public management; Harmonized database; GIS process; Real estate business.

1. INTRODUCTION

In this paper, we focus the analysis on the Italian real estate phenomenon and how the administrative data are powerful in adding new information on the phenomenon in terms of both volume and value comparing with the limited evidence that normally arises from the official statistics yielded by the National Institutes of Statistics (NIS, hereafter). The importance of the analysis we yield in this work is unique and original in its attempt to describe an economic phenomenon that, not only in Italy but in many European countries, still suffers the consequences not only on the real estate business, but also on the public management, of the dearth of a complete and harmonized data warehouse. The analysis is restricted to the territory of the city of Bari, in the South Italy, because that is part of a national research project but the outcomes we generated can be perfectly replicated in any dimensional territorial area. As result of our work, we first built a unique administrative database starting from 4 independent administrative databases, normally managed by two independent Public Administration offices, the Italian Real Estate Registry and the Italian Revenue Agency (IRA, hereafter), which provide autonomous information. The basic Big Data analytic practice has been necessary to guarantee the record linkage and, although we have circumscribed the analysis to one city, the amount of data has also required the application of GIS processes to guarantee the exact matching of data and depict the real estate framework in detail.

2. LITERATURE FRAMEWORK.

Big data are receiving high attention from different industries and functional areas, and not only from the scientific/academic world. Presently, some attempts of big data utilization in the real estate enterprise have achieved some success in terms of decision-making capability and economic benefits, while the systematic research and use in the Public Administration (PA, hereafter) is still far from being a real practice. As suggested by Du et al. (2014), the applications of big data in realty benefits would diversify development and innovative investment and contribute to realty marketing. Scholars on almost all research fields are trying to deal with massive data that are available in many private and public environment. Independently of the scenario in which the so-called Big Data are positioned, one of the main issues that regards the

Big Data topic is the way by which they need to be handled and managed considering that the conventional and traditional statistical and informatic tools are significantly inappropriate. In accordance with Pyne ed al. (2016), it is expected a gradual and robust convergence of data architectures and statistical practice that will be designed for big data and high-performance computing in the next future. The utilization of big data will undoubtedly help in many areas ranging from efficiency of operations to decision making processes, in the private as well as in the public sector. In fact, big data have the potentiality to contribute significantly to the analysis of phenomena with unprecedented breadth and depth and this is what we are proposing in this paper with regard to the real estate market values for which enormous datasets are normally available but very often intricate to manage. The idea, therefore, is to correct imprecise measures of the real estate values which are obviously problematic in a nowadays otherwise hyper-efficient capital market.

3. METHODOLOGY AND ECONOMIC EVIDENCE.

The use of Big Data related to the real estate market is very valuable and interesting for the Public Real Estate Management (PREM, hereafter) research, which is based not only on the Corporate Real Estate Management theory but also on the public management approach. Literature review (Marona and van den Beemt-Tjeerdsma, 2018) clearly indicates there is little research that tries to merge the real estate management perspective with public management concepts, more specifically the New Public Management and the Good Governance approaches as a part of the Public Governance concept. In our view, the main scientific obstacle encountered since now in developing new approaches of analysis in this sense relies on the real lack of integrated and complete datasets that include all information on the phenomenon. In fact, all data, which since now were used in literature, are complete in relation to the database used but partial on the overall view of the phenomenon. The main problems that are normally encountered when it is necessary to work with administrative databases are the typology of data and the corresponding quality that they contain. In our work, we decided to independently work on each database to pre-process data and select the key features of each database to finally proceed with the merging action. The robustness of the result and the potential for use are truly appreciable. Three of the 4 datasets belong to the Real Estate Registry (RER, hereafter) of the city of Bari and they contain all technical information related to the real estates, such as cadastral category and the corresponding economic cadastral income. Although the PA office is the same, the 3 databases are independent and autonomous in providing the corresponding information.

Therefore, the Italian RER has the complete information on real estates though utilizes that in a roundabout way that complicates its same use. The 3 databases experience different sizes because of technical reasons due to the administrative information they report. Therefore, we dealt with missing data, duplication and erroneous information that complicated their merging. Once the numeric and structural homogeneity of the database size has been obtained, the successive step is to merge them. We have identified a merging field, the Cadastral Office Real Estate Identification Code (COREIC, hereafter), in common between the 3 datasets and proceeded to create the Unique Real Estate Registry (URER, hereafter) database by means of COREIC. The fourth database we used belongs to the Real Estate Italian Observatory (REIO, hereafter) of IRA. In Italy, this source of data is the main and one of the most reliable to analyze the real estate monetary value dynamics. The REIO real estate value data are calculated based on the trade price per square meter of the properties and reflect the economic and socio-environmental characteristics of the REIO homogeneous area in which the properties are located. In this work, we use the REIO dataset of the city of Bari for the years 2015 and 2018, referred to 14 typologies of real estates. In Table 1, all information related to the dataset contents is reported.

Dataset	Original Size		Final Size	
	Fields	Records	Fields	Records
Real Estate Units	29	283,217	9	262,977
Cadastral Identifiers	13	421,324	6	262,977
Cadastral Addresses	5	668,302	5	262,977
Real Estate Italian Observatory	24	7,814	6	7,814
NIS Census Sections	4	82,576	4	82,576

Tab. 1: Dataset contents: original size and final size after cleaning action.

To compare the real estate cadastral income with the corresponding market value, which is the most valuable purpose of our work in terms of public management return and business opportunities on the territory, it is necessary to align the URER and REIO databases because they are effectively unlinked and not directly connectable through any field although they are referred to same object. In literature, there is not any attempt in this sense and our work is the first challenge. We solved the geo-technical obstacles on the way of the final objective referring to a

GIS procedure we generated to link each real estate unit to the Italian NIS Census Section database and the REIO homogeneous areas. In fact, the GIS procedure yielded two distinctive maps that we overlapped to obtain our own original database (BRIDGEDB, hereafter) in which all the cadastral units are linked with each REIO area. In the end, the creation of a Transformation Matrix allowed the conclusion of the alignment process of URER and REIO in terms of real estate typology classifications. Therefore, finally, the use of BRIDGEDB and the Transformation Matrix allows to assign the REIO real estate value to each real estate unit for each cadastral category and compute the market value through the cadastral size of each real estate unit. The final database includes, therefore, all the harmonized cadastral and market data for each real estate unit. The extraordinary potentialities of this outcome are very large and can involve many aspects of the PA activities on one hand, and the household/private business on the other. Figure 1 shows, for instance, the percentage average differentials of the Economic Dwellings in the Murat neighborhood and compares 2015 with 2018 data.

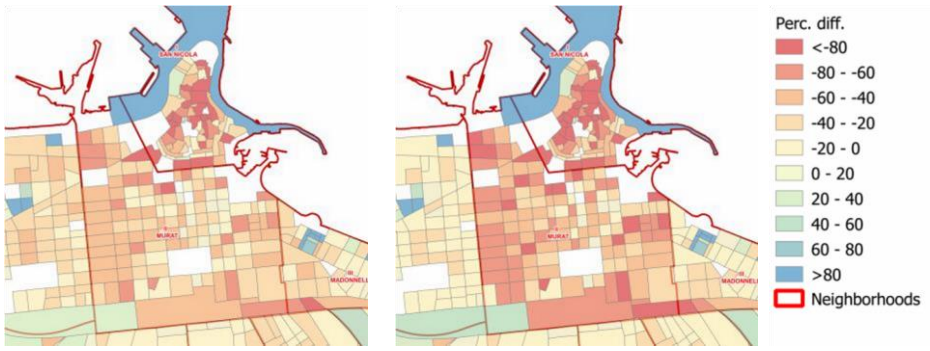


Fig. 1: Murat neighborhood differentials. Years 2015 (left) and 2018 (right).

As we can see on the maps, in this area of Bari the percentage average differentials highlight that the real estate cadastral income is always much lower than the market value, up to 80% in some cases. Furthermore, time comparison shows an increase of the differentials between 2015 and 2018 underlining the effects of requalification actions that involved the Murat historical area of the city. In terms of public management, this outcome and all the others, not described here because of writing limitations, would suggest to municipalities to reconsider not only the economic attention on some specific areas of the city to evaluate hypothetical effects of plausible

requalification actions, but also the fiscal policy for those portions of the city whose market values are significantly greater than the cadastral incomes. At the same time, for households and private enterprises, the opportunity to evaluate at 360 degrees the economic profitability of an investment in the real estate market would have incomparable return in terms of business.

4. CONCLUSIONS

The importance of the analysis we yield in this work is unique and original in its attempt to describe an economic phenomenon that still suffers the consequences not only on the real estate business, but also on the public management, of the dearth of a complete and harmonized data warehouse. The outcomes of the study are economically and statistically noteworthy in depicting a value discrepancy that is de facto considered as known but never numerically quantified. Furthermore, in the consideration that there is little research that tries to merge the real estate management perspective with public management concepts and none of PREM research takes into consideration the ideas here discussed, we can conclude that the outcomes of this study can be valuable at different administrative levels and in other geographical areas.

REFERENCES

Du D., Li A., Zhang L., Li H. (2014) Review on The Applications and The Handling Techniques Of Big Data. In: Chinese Realty Enterprises, Annals of Data Science, 1(3–4), pp. 339–357.

Kok N., Koponen E., Martínez-Barbosa C. (2017) Big Data in Real Estate? From Manual Appraisal to Automated Valuation, The Journal of Portfolio Management, Special Real Estate Issue.

Marona B., Van Den Beemt-Tjeerdsma A. (2018) Impact of Public Management Approaches on Municipal Real Estate Management In Poland And The Netherlands, Sustainability, pp. 1-15.

Pyne S., Prakasa Rao B.L.S., Rao S.B. (2016) Big Data Analytics: Views from Statistical and Computational Perspectives. In: Pyne S., Rao B., Rao S. (eds) Big Data Analytics. Springer.

Quran Analytics and Islamic Banking

Prof. Issam Tlemsani
(Corresponding author)
Prince Mohammad Bin Fahd University
College of Business Administration
Al Khobar, Saudi Arabia
Email: itlemsani@pmu.edu.sa

Prof. Farhi Marir
College of Technological Innovation
Zayed University Dubai,
UAE
farhi.marir@zu.ac.ae

Prof. Munir Majdalawieh
College of Technological Innovation
Zayed University Dubai,
UAE
munir.majdalawieh@zu.ac.ae

Abstract

The aim of this research is to provide a thorough and comprehensive data base that will be used to examine existing practices in Islamic banks' and improve compliancy with Islamic financial law.

An exploratory research approach is used to developed a new text mining Algorithm based on recursive co-occurrence of synonym terms to reflect the "blockchain" like structure of the Quran and Hadith, where stories on aspects of human life (social relationships, moral, trust, economic, ethical… etc.) are spread like pieces of puzzles amongst the 6,236 verses from 114 chapters of the holy Quran and in around 1,300 Hadiths. First, we collect the words reflecting each task or action that compose a business process (e.g. Murabaha) used in current financial institutions like Islamic banks and then we use each term and its synonyms to run the recursive co-occurrence Algorithm. This textual analysis of the holy Quran and Hadiths will be used to build a tree structure that refers to each retrieved verse and hadith that will later be converted into a sequence of financial Sharia compliant actions or tasks.

The main findings of this investigation have led to the classification of all activities across all functional areas in processing any Islamic banks' product. This research seeks to bring research on Islamic banking into the mainstream financial services landscape. It serves as a potential foundation to scientifically engineer the Islamic banking business model. This study contributes to the understanding of Islamic banking mechanism and framework principles and its value as a solution to the current and future Islamic financial complaint products. The built business process resulting from mining the Quran and Hadiths will be presented to experts in the religious domain to compare and validate the findings.

Keywords: Islamic Finance, Islamic financial business processes, Qur'anic corpus, Murabaha

I. Introduction

Text data mining also known as text analytics is a task of extracting high-quality information from text**,** one of its major benefits is that it provides institutions with an effective method for analyzing massive and complex volumes of information known as big data in the form of structured data or free texts. Since the olden times, knowledge in medicine for instance was established through recording and analyzing human experiences. This paper presents a new text mining technique based on recursive co-occurrence of synonym words to build up meaningful stories related to Islamic finance from the Quran, which includes around 78 thousand words, these words are grouped into 6,236 verses, a set of verses are grouped into114 chapters and around 1,300 contextual Hadiths (actions and words of the prophet Muhammad). This text mining technique is proposed to fit the recursive structure of the Quran and Hadith in which verses and hadiths are partially inter-connected but spread like puzzles all over the holy Quran and Hadith texts. These constructed stories will be further analyzed and validated by the Islamic experts to contribute to the development of standard Sharia compliant business processes for Islamic banks.

Islamic finance has grown rapidly over the past few decades and this is expected to continue for the foreseeable future. The Islamic financial market has become increasingly competitive with 1,143 Islamic financial institutions operating globally, thus Islamic banks need to innovate and develop its products beyond Sharia compliance to attract more business and opportunities in the competitive financial environment (Zaraq, 1983; Khan, 1986; El-Gamal, 2001). According to Thomson Reuter's projections, Islamic finance is expected to grow to $3.2 trillion by 2020, with Islamic banking constituting $2.6 trillion of this figure (IFDI, 2015). It has evolved from a niche offering, limited to banks in the Middle East to part of the mainstream financial services landscape globally.

In principle, Islamic financial institute are governed by Sharia law, which includes a combination of primary sources such as the Quran and Hadiths also secondary sources such as Ijma (consensus), Qiyas (analogy), Ikhtiyar (choice), Darurah (necessity), Ijtihad (interpretation) and fatwas (the rulings of Islamic scholars). Islamic financial services institutions' instruments must be Sharia compliant. However, Sharia law lacks uniformity and the interpretation of Sharia differs depending on the branch of Islam and the school of thought within them. Islamic financial institute's clients are becoming more educated about Islamic products and are seeking convincing evidence of the sharia compliance of products that the institute is offering (Kuppusamy et al. 2009). In addition, a large percentage of capital providers, shareholders and investors in Islamic banks are concerned that to what extent the business model and their invested funds are Sharia-compliant (Chapra and Ahmed., 2002). Various bodies in several countries are formed to address the creation of a coherent set of Islamic banking standards. This puts more pressure on the managers of Islamic banks to not only maximize the value of their investments, but also to achieve these objectives in a Sharia compliant way (Archer et al., 1998).

Sharia law compliance has not been researched or explored from a business process management perspective and the current literature has shown a lack of a well-defined methodology for integrating Sharia compliance controls into Islamic Sharia business processes (Vayanos et al.,

2008). Kuppusamy et al., (2009) indicated that the lack of Islamic financial standards would affect the ability of the banks to implement Islamic products and services.

This research on textual mining analysis of the holy Quran and Hadith texts can fulfil these strategic needs in Islamic finance and initiate a new era of research on utilising data analytics on Islamic holy book. The research will look in particular at Sharia compliant business processes like, Murabaha (cost-plus financing modes).

This research paper is organised as follows; section (I) is a background to the subject and provides the purpose of this research. Section (II) provides an overview of previous research work on text data analytics within the Quran and Hadith. Section (III) focusses on the selection of one of the existing Islamic business process to be compared to the business process we developed through text mining the Quran and Hadith texts. Sections (IV) describes the different text mining approaches. Section (V) outlines the research methodology which relies extensively on recursive co-occurrence Algorithm and how it is implemented to mine the Quran and Hadiths to develop a sharia compliant Islamic finance product. Section (VI) highlights what was achieved and how to progress in the future.

II. Previous Research Work on Analysing the Quran

As stated in (Mohammad A. et al., 2015) the holy Quran is the word of God and hence needs careful handling when processed by automated methods of machine learning, natural language processing and artificial intelligence. The language of the holy Quran used in this research is Arabic, which is known to be one of the challenging natural languages in the field of natural language processing and machine learning. This is due to some of its special characteristics such as diacritic, multiple derivations of words, complicated Diglossia and others (Habash, 2010). These make dealing with Arabic language a challenging task when applying machine learning and artificial intelligence techniques.

Limited research has considered the Arabic text of the Quran. Most research has focused on finding frequent patterns (Ali, 2013), developing semantic lexicons (Al-Yahya, 2010), syntactic annotation (Dukes, 2013), statistical extraction and visualization of Quran topics (Panju, 2014) and semantic indexing and retrieval of Quran words and topics (Hikmat, 2013) and (Aliyu, 2013; Kais et al., 2014) have created an open source Quranic corpus (Dukes, 2014) using both Arabic words as well as translations of these words.

EL-Kourdi, et al., (2004) built an Arabic document classification system to classify non-vocalized Arabic web documents based on Naïve Bayes algorithm, while (AL-Kabi, et al., 2005) represent an automatic classifier to classify the verses of "*Fatiha*". In (El-Halees, 2006) a system called ArabCat based on maximum entropy model to classify Arabic documents, and (Saleem et al., 2004) present an approach that combines shallow parsing and information extraction techniques with conventional information retrieval, while (Khreisat, 2006) conducts a comprehensive study for the behavior of the N-Gram frequency statistical natural language technique processing for classifying Arabic text documents. The work of (AL-Kabi, et al., 2007) on comparative study represents the efficiency of different measures to classify Arabic documents. Their experiments[i] show that NB

method slightly outperforms other methods, while (AL-Mesleh, 2007) proposes a classification system based on Support Vector Machines (SVMs), where the classifier uses CHI square as a feature selection method in the pre-processing step of text classification system procedure. In the Quran the verse of "*Yaseen*" is based on predefined themes, where the system is based on linear classification function (score function) and (Hammo, et al., 2008) discussed the enhancement of Arabic passage retrieval for both diacritisized and non-diacritisized text, they propose a passage retrieval approach to search for diacritic and diacritic-less text through query expansion to match user's queries**.**

Akour, M. et al. (2014) investigated Quranic verses similarity and Surah Classification using NGram. There is search work reported by (Mohammad A. et al., 2015) in which he initiated a series of statistical analysis and information retrieval techniques such as TF and TF_IDF to the holy Quran to show a variety of characteristics of the holy Quran such as its most important words, its word cloud and chapters with high term frequencies. The most recent reported by (Rahima B et al., 2017) measures the strength of the semantic relations of the AND conjunction between two semantically same terms in different position around the AND conjunction in the holy Quran.

To our knowledge no research work on using machine learning to analyse the holy Quran and Hadiths to support research question on Islamic finance has been undertaken before this research study which is aiming to support the effort of making current Islamic financial business process compliant to Islamic Sharia' law.

III. The Selected Business Process

As part of the exploratory research, the team has investigated and collected information on the problems facing the Islamic banks who opted for Islamic financial models and Islamic business processes. The findings of the data collection and investigation has led to the classification of all activities across all functional areas in processing any Islamic banks products. For this research, we will focus on Murabaha (shown in figure 1 below) which is the most popular mode of Islamic financing. This choice will lead to the following research question for this study: Is the Murabaha business process compatible with the Sharia law?

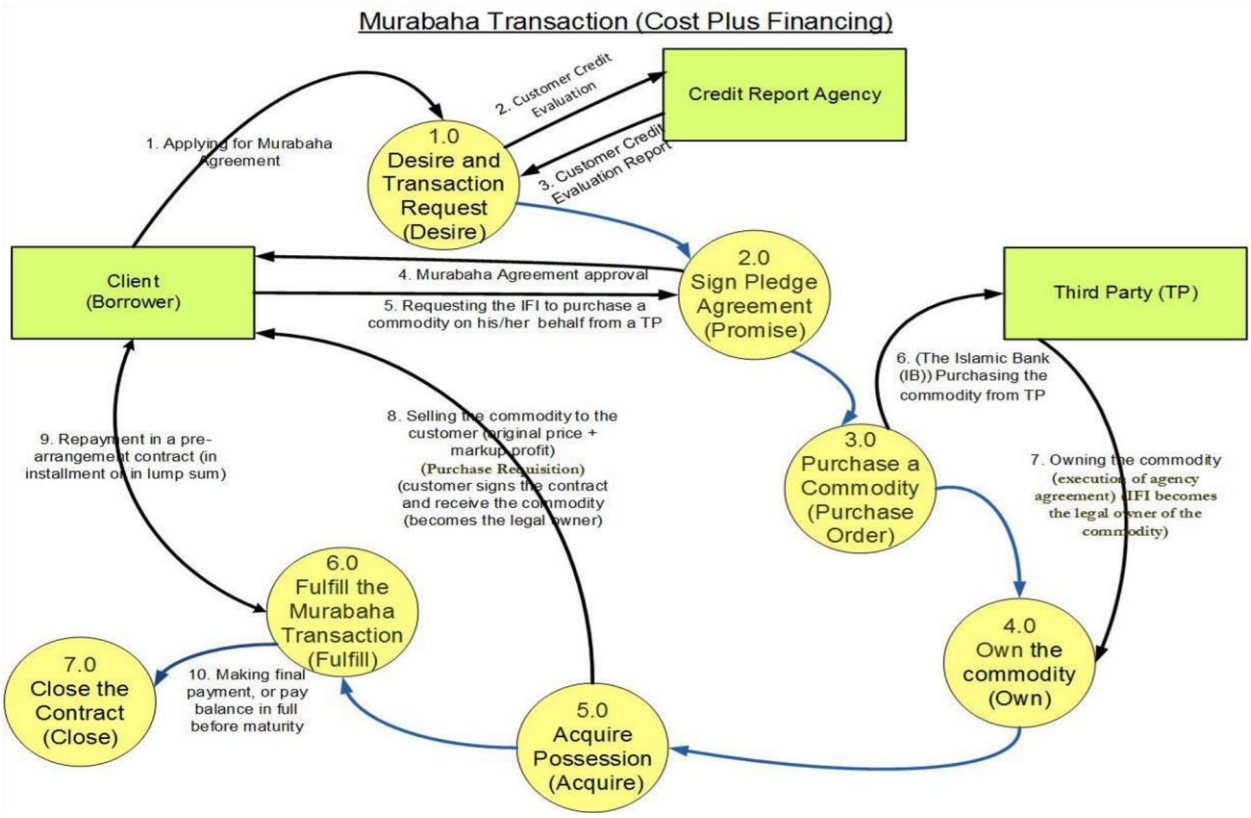
To answer this research question, we will identify the factors needed to consider in our text mining techniques of the holy Quran and Hadith texts for Murabaha business process. These text mining analytics will follow these steps:

1. Compile the holy Quran and Hadith texts adding annotations at multiple levels of Arabic linguistic analysis to produce an indexed corpus easy to link verses and Hadith texts semantically to facilitate the recursive co-occurrence retrieval process.
2. Information extraction: identification of explicitly stated facts by entity recognition, information and event extraction (Quran and Hadiths).
3. Summarization: process by which the salient aspects of one or more Quran's verses and Hadith texts is identified, presented succinctly and coherently.
4. Visualization for human observation, analysis and evaluation in all the steps discussed so far.

These steps will be repeated incrementally, validated and results will be added to the evolving knowledge base of the given financial business (Murabaha) processes. The goal of this phase is to design Murabaha business process based on text mining, which comply with the Sharia law, then critically analyse and compare it with the Murabaha business process that we have developed based on the data collection.

The final phase is to identify any variations and correlation between the Murabaha business processes produced using text mining process of the Quran and Hadith texts against the Murabaha business process implemented by current Islamic banks.

Figure 1: The seven Murabaha business process



Murabaha business process consists of seven main processes to manage and control the flow of Murabaha components. These processes are:

1. Desire process (desire and transaction request) is responsible for presenting the client (borrower) with an application form and provide assistance to complete the form. The Islamic banks through this process requests a credit evaluation report from a credit report company.

2. Promise process (singing a pledge agreement). The Islamic banks is responsible for informing the client of the progress/approval of their application. In response, the client requests the Islamic banks to buy the commodity.

3. Purchase order process (The Islamic banks to purchase the commodity from the third party): It is responsible for buying the commodity after confirmed that everything is in order, and immediately sell it to the client at an agreed upon price (cost + mark-up). The contract's terms will be processed in the contract. This process will ensure that a pre-approval from the client is obtained before the Islamic banks can request to buy the commodity.

4. Owning process (Islamic banks will own the commodity): This Islamic banks will purchase the commodity in this process, register the commodity in the client name or in the name of a trust, a corporate (offshore & onshore), or a partnership.

5. Acquire process (acquire possession): It will be invoked to record the sale between the client and the Islamic banks in the Murabaha contract. Client first payment to the Islamic banks is made on the day of completion and is the initial contribution (down payment). Payments are fixed for the entire payment term.

6. Fulfil the Murabaha transaction: This process will be activated monthly almost one month after completion of the previous process. The monthly payments will be claimed by direct debit automatically from the client's current account.

7. Close the Murabaha contract: This is the final process in the Murabaha transaction and it takes effect when the Islamic banks collects all the full Murabaha debts from the client.

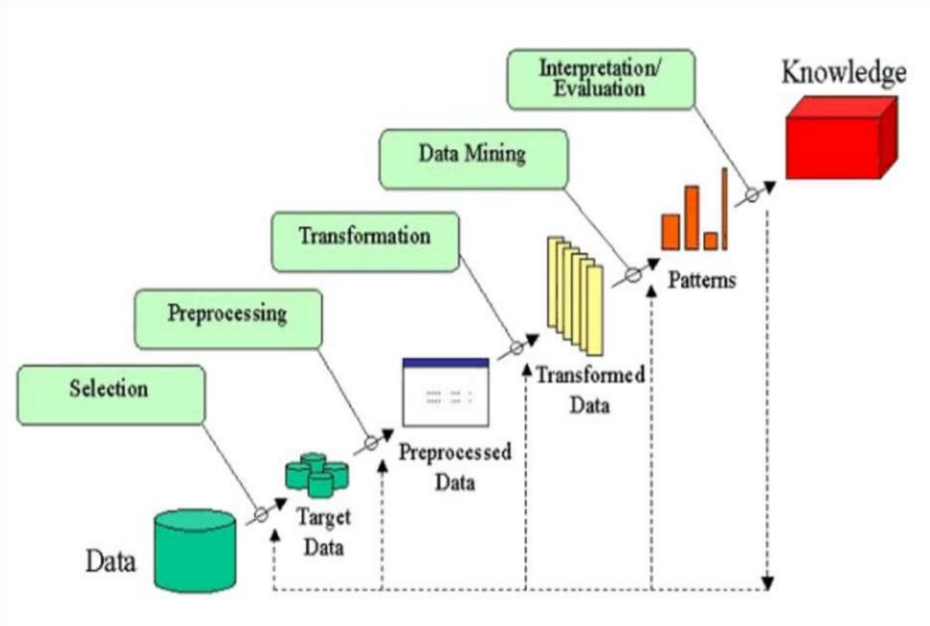
These processes interact amongst themselves and with external entities (represented by the rectangles) at the same time. External entities are clients (borrowers), credit rating/report company and third party who provide and receive data inputs and outputs from and to the Islamic information system. Such interactions are briefly described in figure 1 above.

IV. The Different Text Mining Approaches

Three basic types of approaches in text mining are used for new knowledge discovery in Databases (shown in figure 2 below); knowledge based, statistical or machine-learning based, and cooccurrence approach (Bretonnel K. et al., 2008). In general, knowledge based systems take a significant amount of time to develop and requires deep understanding of the domain to analyse (Cohen, 2004). Statistical or machine learning based systems are more appropriate for classifications and require large amounts of expensive labelled training data (Craven, 1999). However, the co-occurrence based methods look for concepts that occur in the same unit of text typically a sentence or verse in the Quran, but sometimes as large as an abstract and suggest a relationship between them as reported in (Swanson, 1986; Hui, 2009 and Jelier, 2005; Swanson, 1986) used basic text mining approach based on terms co-occurrence and transitive relationship (A -> B -> C) between terms to analyse medical research abstracts. He suggests a connection between dietary fish oil and Raynaud's disease, a circulatory disorder, which three years later was validated

through clinical trials. In addition, the recent research reported in (Hui, 2009.); which aims to automatically identify the status of obesity and 15 related co-morbidities in patients based on their clinical discharge summaries using terms co-occurrence approach to identify both explicit and implicit references to the diseases. In (Jelier, 2005) co-occurrence, based meta-analysis of scientific texts is used for retrieving biological relationships between genes and their results revealed more functional biological relations and could even achieve results with less literature available per gene. The performance achieved in this co-occurrence text mining work is in line with the agreement between human annotators, indicating the potential of text mining for accurate and efficient prediction of disease statuses from clinical discharge summaries.

Figure 2: Knowledge Discovery in Databases (KDD) process.



In line with these successful texts mining, we are proposing in this paper an enhanced approach of co-occurrence of text mining approach to reflect the blockchain like structure of the holy Quran and Hadith texts. The proposed approach is not limited to transitive co-occurrence as reported previously but goes beyond that into recursive co-occurrence of finance related terms to extract as much as possible of inter related terms to build a tree of key word terms that reflect Murabaha business process shown in figure 1.

V. The Recursive Co-Occurrence and its Implementation

5.1 The Recursive Co-occurrence Algorithm

The next step is to mine Quran, Hadiths and exegesis of Quran to find out how the Murabaha business process shown in figure 1 above is seen in these holy texts. Text mining techniques analyse and identify patterns through given free text is well known in different domains, for instance medical science especially mining clinical records, biomedical texts, and medical research paper.

In this research, we will use recursive words co-occurring text mining to extract recursively linked verses through synonym words that makes the link of verses in the Quran and Hadith texts, which are linked like the items of a blockchain. This retrieval of semantically linked Sharia' law business processes are compared with existing Islamic finance business used in current Islamic banks like Islamic banks. The proposed recursive co-occurrence of keyword terms algorithm is as follows:

Recursive-Co-Occurrence Algorithm**:**

1. *Gather all the terms related to a particular business process into a set* $T_{initial} = \{t_1, t_2, t_3, \ldots, t_m\}$
2. *Create a B-tree structure* T_{bp} *for a given business process and initiate its root,*
3. *Iteration =0; # initiate the counter of recursive iterations*
4. $T_{\text{current}} <- T_{initial}$
5. While *(Iteration < n) # given n the number of recursive iteration;*

   For *each term,* t_i *of* $T_{current}$ *set* Do
   - *Run the co-occurrence frequency process (word sketch) in Sketch Engine (lexical computing) to extract from Quran & Hadiths corpus the other terms in* $T_{current}$ *co-occurring with terms* t_i*,*
   - *Append the extracted terms at the right node of the B-tree* $T_{bp;}$
   - *Select manually new discovered relevant terms into a new set named* $T_{observed}$*,*
   - *Append all sentences related to the extracted co-occurring terms into a new subcorpus named* C_t*,*

   End
   $T_{\text{current}} <- T_{observed} \cap T_{initial;}$
   $T_{initial} <- T_{observed} \cup T_{initial;}$
   $T_{observed} <- \emptyset;$
   Iteration <- iteration +1; If ($T_{\text{current}} == \emptyset$) *then go to step 6;*

   Else *Go step 5*
   End while

6. *Mine all the sentences contained in the* C_t *sub-corpus based on the hierarchical relationships of the B-tree* T_{bp}
7. *Use the result of text mining process, produce an ordered list of terms that could be used to compose the alternative actions in the original business process*

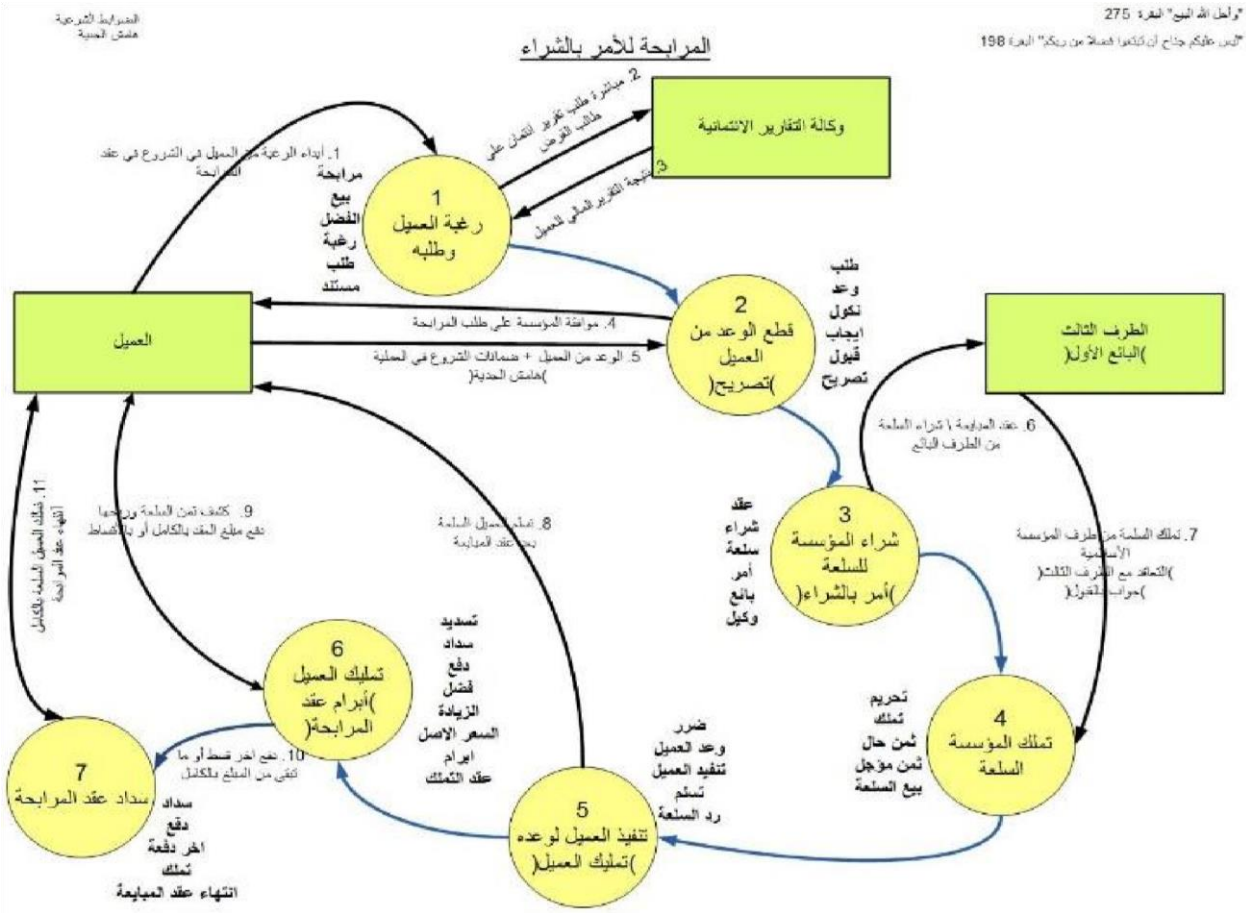
5.2 Preparing the Holy Quran and Hadith Texts

The Arabic version of the holy Quran has been downloaded from Tanzil project website (Tanzil, 2014) which represents an authentic verified source of the holy Quran text. The holy Quran is composed of 114 chapters, 6,236 verses and the Hadith texts is composed of around 1,300 collected from several sources and were compiled into a corpus using Sketch Engine text mining tool.

5.3 Running the Co-occurrence Algorithm

In the first step, we browse through Murabaha business process of figure 3 below to collect the initial terms that defines tasks or action as: *T initial* = {مضاربة (Mudaraba); مرابحة (Murabaha); بيع (elaS); شراء (esahcruP); عقد (tcartnoC)…cte.}.

Figure 3: Murabaha business process with keywords (Arabic Version)



5.3.1 Extract from KUCC corpus the other terms co-occurring with terms in $T_{initial}$

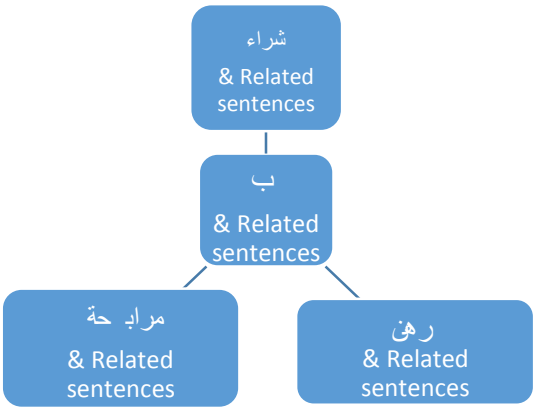
Then we used the process co-occurring terms (word sketch operation in the Sketch engine) to extract terms within initial co-occurring together (shown inside red rectangles) and manually collect the new observed terms (shown in blue rectangle) e.g. observed = {نمر (mortgage); ض رق (loan)…. etc.} relevant to the business process in hand as shown in figure 4 below. These new observed terms would be used to perform another recursive iteration to discover new relevant terms to the Murabaha business process.

Figure 4: Iteration for extracting co-occurrence of Murabaha business process related terms

5.3.2 Create a B-tree structure for a given business process and initiate its root

Based on the inter relationship occurrence, a B-tree will be built to explicitly reflect these relationships to be used as an indexing scheme for accessing related sentences or verses of the Quran or Hadiths for further analysis. An example of this B-Tree in figure 5 below could be deduced from figure 4 above.

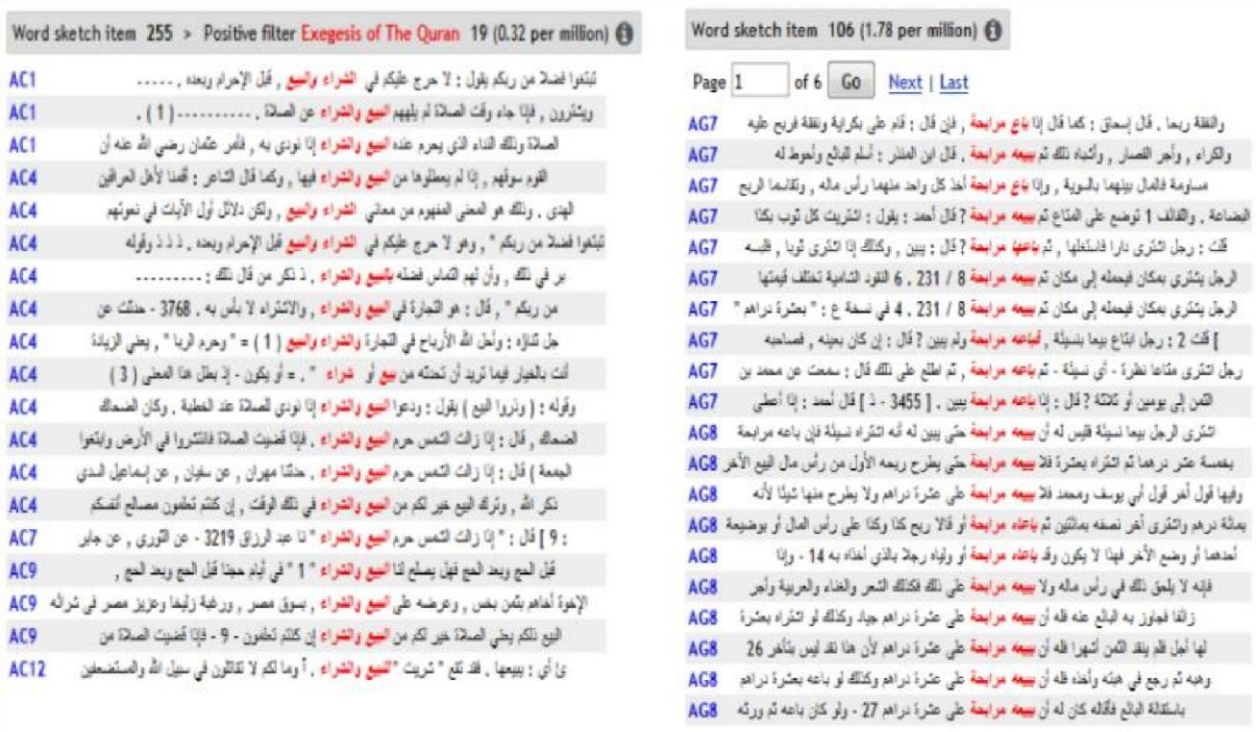
Figure 5: An instance of a B-Tree showing the relationship between terms



5.3.3 Append all sentences related to the extracted co-occurring terms into a new sub-corpus

Using the hyper link of co-occurring terms shown in figure 6 below, we drill down to the indexes of related sentences containing the co-occurring terms like {بيع (Sale); شراء (Purchase) and مرابحة (Murabaha)} as shown in figure 5. These indexes will help further drill down to download new sentences to be saved into a new sub corpus for further text analysis to find more patterns relevant to the Murabaha business process and more importantly in producing ordered list of words to be composed into an alternative Islamic business process that is built using the holy Quran and Hadiths.

Figure 6: Sample of co-occurring (بيع (Sale); شراء (Purchase) and مرابحة (Murabaha))



Word sketch item 255 > Positive filter Exegesis of The Quran 19 (0.32 per million)

AC1 لتبتغوا فضلا من ربكم يقول : لا حرج عليكم في الشراء والبيع , قبل الإحرام وبعده , ...
AC1 ويشترون , فإذا جاء وقت الصلاة لم يلهيهم البيع والشراء عن الصلاة , ...(1) .
AC1 الصلاة وذلك النداء الذي يحرم عنده البيع والشراء إذا نودي به , فأمر عثمان رضي الله عنه أن
AC4 القوم سوقهم , إذا لم يعطوها من البيع والشراء فيها , وكما قال الشاعر : فقدنا الأهل العراقين
AC4 الهدى . وذلك هو المعنى المفهوم من معاني الشراء والبيع , ولكن دلائل أول الآيات في نعوتهم
AC4 لتبتغوا فضلا من ربكم " , وهو لا حرج عليكم في الشراء والبيع قبل الإحرام وبعده . ذ ذ ذ وقوله
AC4 بر في ذلك , وأن لهم التماس فضله بالبيع والشراء , ذ ذكر من قال ذلك : .....
AC4 من ربكم " , قال : هو التجارة في البيع والشراء , والاشتراء لا بأس به . 3768 - حدثت عن
AC4 جل ثناؤه : وأحل الله الأرباح في التجارة والشراء والبيع (1) = " وحرم الربا " , يعني الزيادة
AC4 أنت بالخيار فيما تريد أن تحدثه من بيع أو شراء " . = أو يكون - إذ بطل هذا المعنى (3)
AC4 وقوله : (وذروا البيع) يقول : وذعوا البيع والشراء إذا نودي للصلاة عند الخطبة , وكان الضحاك
AC4 الضحاك , قال : إذا زالت الشمس حرم البيع والشراء , فإذا قضيت الصلاة فاشتروا في الأرض وابتغوا
AC4 الجمعة) قال : إذا زالت الشمس حرم البيع والشراء , حدثنا مهران , عن سفيان , عن إسماعيل السدي
AC4 ذكر الله , وترك البيع خير لكم من البيع والشراء في تلك الوقت , إن كنتم تعلمون مصالح أنفسكم
AC7 : 9] قال : " إذا زالت الشمس حرم البيع والشراء " ثنا عبد الرزاق 3219 - عن الثوري , عن جابر
AC9 قبل الحج وبعد الحج فهل يصلح لنا البيع والشراء " 1 " في أيام حجنا قبل الحج وبعد الحج ,
AC9 الإخوة أخذهم بثمن بخس , وعرضه على البيع والشراء , بسوق مصر , ورغبة زليخا وعزيز مصر في شرائه
AC9 البيع تلكم يعني الصلاة خير لكم من البيع والشراء إن كنتم تعلمون - 9 - فإذا قضيت الصلاة من
AC12 يا أي : بيعها , فقد تقع " البيع والشراء , آ وما لكم لا تقاتلون في سبيل الله والمستضعفين

Word sketch item 106 (1.78 per million)

Page 1 of 6 Go Next | Last

AG7 والنفقة ربحا . قال إسحاق : كما قال إذا باع مرابحة , فإن قال : قام علي بكرابة ونفقة فربح عليه
AG7 والكراء , وأجر القصار , وكأنه ذلك ثم بيعه مرابحة . قال ابن القاسم : أسلم للبائع وأحوط له
AG7 مسؤمة فالمال بينهما بالسوية , وإذا باع مرابحة أخذ كل واحد منهما رأس ماله , وتقاسما الربح
AG7 البضاعة . والقالف 1 توضع على المتاع ثم بيعه مرابحة ? قال أحمد : يقول : اشتريت كل ثوب بكذا
AG7 قلت : رجل اشترى دارا فاستغلها , ثم باعها مرابحة ? قال : يبين , وكذلك إذا اشترى ثوبا , فلبسه
AG7 الرجل يشتري بمكان فيحمله إلى مكان ثم بيعه مرابحة 8 / 231 . 6 النقود الشامية تختلف قيمتها
AG7 الرجل يشتري بمكان فيحمله إلى مكان ثم بيعه مرابحة 8 / 231 . 4 في نسخة ع : " بعشرة دراهم "
AG7] قلت 2 : رجل ابتاع بيعا بنسيئة , فباعه مرابحة ولم يبين ? قال : إن كان بعينه , فصاحبه
AG7 رجل اشترى متاعا نظرة - أي نسيئة - ثم باعه مرابحة , ثم اطلع على ذلك قال : سمعت عن محمد بن
AG7 الثمن إلى يومين أو ثلاثة ? قال : إذا باعه مرابحة يبين . [3455 - ذ] قال أحمد : إذا أعطى
AG8 اشترى الرجل بيعا نسيئة فليس له أن يبيعه مرابحة حتى يبين له أنه اشتراه نسيئة فإن باعه مرابحة
AG8 بخمسة عشر درهما ثم اشتراه بعشرة فلا يبيعه مرابحة حتى يطرح ربحه الأول من رأس مال البيع الآخر
AG8 وفيها قول آخر قول أبي يوسف ومحمد فلا يبيعه مرابحة على عشرة دراهم ولا يطرح منها شيئا لأنه
AG8 بمائة درهم واشترى آخر نصفه بمائتين ثم باعاه مرابحة أو قالا ربح كذا وكذا على رأس المال أو بوضيعة
AG8 أحدهما أو وضع الآخر فهذا لا يكون وقد باعاه مرابحة أو ولاه رجلا بالذي أخذاه به 14 - وإذا
AG8 فإنه لا يلحق ذلك في رأس ماله ولا يبيعه مرابحة على ذلك فكذلك الشحر والغناء والعربية وأجر
AG8 زائفا فجاوز به البائع عنه فله أن يبيعه مرابحة على عشرة دراهم جياد وكذلك لو اشتراه بعشرة
AG8 لها أجل فلم ينقد الثمن أشهرا فله أن يبيعه مرابحة على عشرة دراهم لأن هذا نقد ليس بتأخير 26
AG8 وهبه ثم رجع في هبته وأخذه فله أن يبيعه مرابحة على عشرة دراهم وكذلك لو باعه بعشرة دراهم
AG8 بالمقالة البائع فقاله كان له أن يبيعه مرابحة على عشرة دراهم 27 - ولو كان باعه ثم ورثه

VI. Conclusion

Muslims believe that the Qur'an is a sole authoritative source for knowledge, wisdom, guidance and legislations for mankind. It was challenging research work to be one of the first to initiate the use of machine learning to mine the holy Quran and Hadith to extract knowledge to consolidate the Islamic financial business processes and make them more compliant with Islamic Sharia' law. Another challenge is the Arabic version of the holy Quran, which is known to be one of the most challenging natural languages in the field of natural language processing and machine learning.

To our knowledge, there is no similar research work on mining the holy Quran and Hadiths to extract knowledge to support other aspects of human life. In this research work we achieved the first results by assembling some key words that could provide an alternative or an improvement to the existing Islamic Murabaha (cost-plus financing) business process and be more compliant with Islamic Sharia' law on the condition that it is accepted and validated by experts in Islamic Sharia' law. These results are milestone for further research to cover other Islamic financial business processes.

In the late 1970s early 1980s, it was shown, mostly by (Minsky, 1982), that conventional banking systems were inherently prone to instability because there would always be maturity mismatch between liabilities (short-term deposits) and assets (investment long-term). Because the nominal values of liabilities were guaranteed, but not the nominal value of assets, when the maturity mismatch become a problem, the banks would go into a liability management mode by offering higher interest rates to attract more deposits. There was always the possibility that this process could not be sustained resulting in erosion in confidence and bank runs. Such a system, therefore,

needed a lender of last resort and bankruptcy procedures, restructuring processes, and debt workout procedures to mitigate contagion. Perhaps new ways of being in the world will emerge: recognition of interdependence is a step towards realising the need, at a practical as well as a spiritual level for compassion. We could cite a number of Quranic messages:

> "*that God does not change the condition of a people until they change their own inner selves*" (Quran: 13:11) and "*Mankind was created as one nation, but they became divided because of differences among them*" (Quran: 10:19).

References

- Ali, I., (2012), "Application of a mining algorithm to finding frequent patterns in a text corpus: A case study of the Arabic" *International Journal of Software Engineering and Its Applications,* Vol. 6, pp. 127–134.
- Aliyu Rufai Yauri A. A. et al. (2013), "Quranic verse extraction base on concepts using owldl ontology" *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 6, No. 23, pp. 4492 - 4498.
- Alrabiah, M. S (2014), "King Saud University Corpus of Classical Arabic (KSUCCA)" Department of Computer Science, *King Saud University,* April.
- Alrabiah, M., Al-Salman, A., Atwell, E., (2013), "The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic" *In: Second Workshop on Arabic Corpus Linguistics (WACL-2),* Lancaster, UK.
- Al-Yahya M., Al-Khalifa H., Bahanshal A., Al-Odah I., and Al- Helwah, N. (2010), "An antological mdoel for representing semantic lexicons: An application on times nouns in the holy quran" *The Arbaian Journal for Science and Engineering,* Vol. 35(2c), pp. 21–37.
- Akour, Mohammed; Alsmadi, Izzat; and Alazzam, Iyad. (2014), "MQVC: Measuring Quranic Verses Similarity and Sura Classification Using N-Gram" *WSEAS Transactions on Computers,* Vol. 13, pp. 485-491.
- Archer S., Ahmed, R. and Al-Deehani, T. (1998), "Financial contracting, governance structures, and the accounting regulation of Islamic banks: An analysis in terms of agency theory and transaction cost economics" *Journal of Management and Governance,* Vol. 2, pp. 149 - 70.
- Bentrcia, R., Zidat S., Marir F. (2017), "'Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns" *Journal of King Saud University – Computer and Information Sciences, http://dx.doi.org/10.1016/j.jksuci.2017.09.004*
- Bretonnel K. C. and Lawrence Hunter (2008), "Getting Started in Text Mining", *PLoS Comput Biol.* Vol. 4(1).
- Chapra, M. U. and Ahmed, H. (2002), "Corporate Governance in Islamic Financial Institutions, Islamic Development Bank", *Islamic Research and Training Institute, Periodical Document* No. 6.
- Daniel McDonald (2014) , "Text Mining Analysis of Religious Texts", *The Journal of Business Inquiry,* Vol. 13, Issue 1 (Special Issue), PP. 27-47 (www.uvu.edu/woodbury/jbi/articles/)

- Dukes K., Atwell E., and Habash N., (2013), "Supervised collaboration for syntactic annotation of quranic arabic", *Language Resources and Evaluation,* Vol. 47, No. 1, pp. 33–62. *[Online]*. *Available:* http://dx.doi.org/10.1007/s10579-011-9167-7
- Habash, N. Y. (2010), "Introduction to Arabic Natural Language Processing" *G. Hirst, Ed. Morgan and Claypool Publishers.* .
- Hikmat Ullah Khan M. S. et al. (2013), "Ontology-based semantic search in holy quran", Vol. 2, No. 6, pp. 562–566.
- Hui Y., Irena S., John A. K., Goran N. (2009), "A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries" *Journal of the American Medical Informatics Association,* Vol. 16, No. 4
- International Monetary Fund, (2015), "Islamic Finance and the Role of the IMF" *Retrieved from http://www.imf.org/external/themes/islamicfinance/index.htm o*n April 30 2016.
- Institute of Management Accountants, (2000), "Implementing Process Management for Improving Products and Services", *ISBN 0-86641-288-3, 2000.*
- Kais D., (2014), "Quranic Arabic corpus". [Online]. Available:http://corpus.quran.com/
- Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D., 2004. 'The Sketch Engine' *Proceedings of EURALEX, Lorient,* pp. 105–116.
- Kuppusamy, M.; Raman. M.; Shanmugam, B.; Solucis, S. (2009), "A Perspective On The Critical Success Factors For Information Systems Deployment" *In Islamic Financial Institutions."* EJISDC, Vol.37, No. 8, pp. 1-12.
- Alhawarat, M.; Mohamed Hegazi. M and Hilal. A, (2015), "Processing the Text of the Holy Quran:a Text Mining Study" *International Journal of Advanced Computer Science and Applications,* Vol. 6, No. 2.
- Panju, M. H. (2014), "Statistical extraction and visualization of topics in the qur'an corpus" *Master's thesis, University of Waterloo.*
- Swanson D.R. (1986), "Fish Oil, Raynaud's syndrome, and undiscovered public knowledge" *Perspectives in Biology and Medicine,* Vol. 30, pp. 7-18.
- Vayanos, P., Wackerbeck, P., Golder, P. T., Haimari, G., (2008), "Competing Successfully In Islamic Banking", *Booz & Co, Retriev*
- Cohen KB, Hunter L. (2004), *"Natural language processing and systems biology" In: Dubitzky W, Pereira F, editors. Artificial intelligence and systems biology. Berlin: Springer Verlag.*
- Craven M, Kumlein J.(1999) *"Constructing biological knowledge bases by extracting information from text sources" Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology;* pp. 5–10, Heidelberg, Germany. Menlo Park (California): AAAI Press; 1999. pp. 77–86.
- Hui Y., Irena S., John A. K., Goran N. (2009), "*A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries" J*ournal of the American Medical Informatics Association Vol.16, No. 4.
- Jelier R.,G. Jenster, L. C. J. Dorssers, C. C. van der Eijk1, E. M van Mulligen1, B. Mons1 and J. A. Kors1 (2005), "Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes" *Bioinformatics,* Vol. 21, No. 9, pp. 2049 - 2058
- Tanzil.net. (2014), "Tanzil" *Quran text download from http://tanzil.net/download/* (latest access is on 26th Feb 2018.

Abstract submission for 2019 International Conference on Big Data in Business, 29-31 October, 2019, London UK

Learning from failure.
Big data analysis for detecting the patterns of failure in innovative startups

by

Ulpiana Kocollari[47] and Maddalena Cavicchioli[48]

Understanding business failure represents still a challenge for researchers since most studies deal with prediction of failure and only few focuses on its understanding. Most of these prediction models depend on accurate quantitative data over several years prior to failure that are problematic especially for startups and small ventures that are known to be weak in financial data records. Furthermore, the multidimensional nature of failure implies that numerous variables of a nonfinancial kind should be analyzed for marking the causes of failure.

In this realm, *the purpose of the paper is to individuate appropriate models for analyzing large dataset in order to detect the patterns of failure in the case of innovative startups and understand the interaction of their economic, context and governance variables and their influence over the different patterns.*

The study is based on financial, governance and context data of all 180 Italian innovative startups failed from 2012 to 2015. The considered sample collects data on the entire population of Italian unsuccessful startups, so it is representative of the population as a whole.

Failure patterns have been uncovered integrating the use of factor and cluster analyses, where the factor scores for each firm are used to identify a set of homogenous groups based on cluster analysis. The integrated use of those large dimensional data techniques permits to classify the data in rigorous ways and to unfold structures of the data, which are not apparent in the beginning. Considering the startups' age, the number of different patterns of failure reduces from three to two. Important evidence was found about the presence of different patterns of failure in different sectors for most firm age groups whereas the patterns of failure are not strongly associated with the geographical area in which the startups are born. The numerosity of shareholders features different patterns of failure only in mature startups.

The research shows that there are few common features that transcend the various patterns of failure, as startup phase of business will inevitably carry with them remnants of the business idea and the financial structure, and some features will be relatively persistent, despite the changing needs as startups move from one stage to another. But at the same time, some features are unique to each pattern, as the main needs in which the startup decision-making process takes place will change significantly over the different stages of business idea development.

[47] University of Modena and Reggio Emilia, Italy, and Softech-ICT (ulpiana.kocollari@unimore.it)

[48] University of Modena and Reggio Emilia, Italy, and ReCent (maddalena.cavicchioli@unimore.it)

The study highlights the usefulness of large data techniques for monitoring startups' performance together with their governance and context characteristics in order to detect and manage their different potential failure patterns. The analysis suggests that each pattern of failure is a multidimensional construct and as a consequence can generate different managerial implications. Therefore, an effective handling of failure requires management to use appropriate intervention targeted at the challenges faced at that particular pattern of failure in different firm's age.

Application of Rough Set Theory to Predict Telecom Customer Churn

Tu Van Binh

University of Economics Ho Chi Minh City and CFVG

59C Nguyen Dinh Chieu Strict, District 3, Ho Chi Minh City, Vietnam

tuvanbinh@gmail.com or tvbinh@cfvg.org

Abstract

This paper is an application of algorithms in machine learning program, e.g. WEKA to predict customer churn. To do this, 211,777 instances in the telecommunication sector are used in the method with the candidate of six attributes. Based on an approach of RST with supported machine learning, the paper found an accurate method, in which the genetic algorithm is a strong candidate to outperform rules generation algorithms in terms of precision, accuracy, recall, and F-measure, which Decision Table is the best candidate to predict churn. The study also reveals the performance of machine learning algorithm will be a good reference for telecommunication provider to think of decision making to retention of customers who are churn risks.

Keywords: Churn, telecom, rough set theory, decision table

I. Introduction

Customer churn absolutely threatens a share of telecommunication operators, so a predictive churn model is indispensable. Once churn customers are estimated, the business cost can be saved. As argued by (Sharma & Kumar Panigrahi, 2011), the cost of keeping the existing customers is less expensive than that of acquiring new customers. However, only a one way measure of the churn rate is not powerful good to accurately identify its business consequence, because it can be affected by multiple ways as from many others side. As a result, to estimate a proper churn, besides the churn rate, more attributes are needed to be added to study what mutual impacts between them are. There are many methods to predict customer churn, e.g. Neural Net, Native Bayes, Decision Tree, and so on. However, these methods sometimes face a problem of uncertainty data, while rough set theory (RST) can be seen as a mathematical approach to intelligent data mining and intelligent data analysis.

According to proposed theory of (Pawlak, 1998), using RST the researchers can discover hidden rules and solve the unknown data distribution. However, an application of RST is not popularly used to predict customer churn in telecommunication industry (Amin et al., 2019). This is partly a reason this paper is going to use RST to address the challenging customer churn prediction problem.

According to (Vashist & Garg, 2011), reduct and score are two important concepts of RST, in which reduct is a reduced subset of initial database and to reduce unnecessary attributes towards decision making applications, while the score is the story after the reduct processed and can be found as the set of all singleton entries in the discernibility matrix. Doing that, the RST is used as a mathematical tool to obtain precise and perfect information which (Nafis, Makhtar, Awang, Rahman, & Deris, 2016) employed after a process decanting database.

To predict customer churn in the telecommunication sector properly, besides a calculation of the attribute of churn information is concerned, others attributes are also enrolled, e.g. data usage of subscribers, length of stay, social calling, top-up, types of mobile or handset. The paper organization is structured in five sections, the next section is literature review. The method proposition is built in section 3 through related l works. section 4 offers the results with findings and discussion. The final section as conclusion is presented section 5.

II. Related work

As argued in previous scholars, RST was found early in the years 1900's. To deal with imprecise concepts, (Pawlak, 1998) introduced this theory. Because of competition and quick technology development, some studies concerned RST to analyze database to predict customers 'churn are coming more. RST is one of the most powerful mathematical frameworks to decant certain data and can be used to find subsets of relevant features. Churn prediction is not easy and a challenging activity for telecom operators. To predict customer churn, (Amin et al., 2019) grouped database into two categories: (i) data with high certainty, (ii) data with low certainty. While (Trabelsi, Elouedi, & Lingras, 2011) used rough sets with a support of two classification approaches. "Belief rough set classifier" as the first technique and "belief rough set classifier based on generalization distribution table" are employed in the study, in which an application of the heuristic method of attribute selection are concerned. As arguments of (Trabelsi et al., 2011), a lot attributes always

are appeared in a database. It, of course, can be problems of redundant for rule discovery. This can waste researchers' time to obtain a certainty attribute selection. Accordingly, Dong, et al. (2001) employed the rough set theory with heuristics method for feature selection. (Zhong& Skowron 2001) had an introduction to rough set, in which "generalization distribution table and rough set system", "rough sets with heuristics", and "rough sets and Boolean reasoning" are argued. Continuously, with application of algorithm toward rule generation, some authors employed learning machine, in which WEKA is the main candidate to predict customer churn by using RST, such as (Shaaban, Helmy, & Khedr, 2012) and (Amin et al., 2017).

While (Zhong & Skowron, 2001), (Dong, et al., 2001), (Trabelsi, et al., 2011) have arguments on who to use rough sets with heuristic to gain right selection of features, while (Amin et al., 2017), (Arun Kumar, Sooraj, & Ramakrishnan, 2017), and (Amin et al., 2019) approached RST based on different algorithms and testing them carefully, which WEKA is a candidate of program to recruit to test feature selection by algorithms. Based on achievements relevant to this study, the method of RST with employed rule generation algorithms is taken into account through WEKA program, this is a very important point to address a necessary reduction on a lot attributes sourced from big data, and remove redundant ones.

As a result, Using RST with algorithm combination based on learning machine to predict customer churn in the telecommunication industry will bring a advanced approach to the literature contribution for studies in telecommunication sector in Vietnam. This study is a new page opened to wake telecommunication operators up thinking of the mathematical approach to identify how much customers churn and what reasons should be concerned. This finding will be an interesting story for researchers as companies toward advanced methods in the future to improve business.

III. Methodology

Database information

Data used in the study is sourced from big data of a telecommunication operator in Vietnam. Because of a security of information and competition, the database of this telecommunication service provider is asked to be hidden its name. As known, big data is the source available, which present proper consumption history of users toward, while database sourced from the survey is opposite. As a result, this is important advantages to contribute to the significance of model development toward right prediction. Anyway, the data are provided by the department of information technology of the company, which a sample size collected is 211,777 prepaid subscribers (as well as instances) who are active customers by a cell phone. Sampling is randomly selected from entire customer base of more than 34 million subscribers from population of the telecommunications company. As stated, all of these customers were active. Months considered to collect to measure customer's churn behavior are followed for 6 months, this means from May 1, 2018 as the origin of time to October 31, 2017 as the observation terminated time. Subscribers who have a length of stay (LOS) less than three months is not included, due to be sure clear consumption behavior after three months.

Churn is conceptualized differently in various sector. It brings a message of customer loss, also evaluate customer retention efforts of service providers (Jahanzeb & Jabeen, 2007). In this paper, the telecom churn is defined as one subscriber who is not using service of telecommunication operators anymore after three months. To predict telecom churn , (Hassouna, Tarhini, Elyas, &

AbouTrab, 2015) develop churn modelling techniques of logistic regression and decision tree, in which they concern variables of data usages and voice usages to investigate how changes in those to the churn situation. While (Olle, 2014) employed tariff to investigate its impacts on the telecom churn, which tariff is defined as the billing type available by individual subscriber. As a result, top-up in this paper is a kind of the monthly expenditure of subscribers to recharge their service usage, displayed as the monthly billed amount as proxy of tariff. Likely, (Ahn, Han, & Lee, 2006) concerned service usage, in which the monthly expenditure as the monthly billed amount of subscribers is displayed as the important behavioral predictors to predict churn.

Unlikely with other scholars, this paper employs a variable of offline calls, it is defined as a measure of how much communications of subscribers by calls and messages with others who are using a telecom service of different operators. Consequently, this can be hypothesized that an increase in the subscriber's offline calls to or/and from others ones under different operators causes a rise in churn risk. Particularly, (Amin et al., 2017) selected features of number of messages and calls to predict customer churn behavior by rough set theory. There are three levels (i.e. LowMedium-High) to identify on subscribers who communicate others using different telecommunication operators. To identify determinants affecting subscriber churn and customer loyalty in the Korean telecommunication industry, (H. S. Kim & Yoon, 2004) recruited handset to investigate customer behavior toward measure of customer lifetime value toward market segmentation.

To predict customer lifetime value, (Rosset, Neumann, Eick, & Vatnik, 2003) concerned a length of stay (LOS) in the model, which the LOS supports the method to investigate the customer's churn probability during period time. Based on this, LOS is employed in the model to find out its relationship with the churn behavior of describers by the method of RST.

With references mentioned previously, six features at beginning are enclosed in this paper, such as LOS, top-up, offline call, handset (type of mobile) and churn. Except to top-up, LOS, and data usage are available on the system of the data warehouse, while churn, offline calls are generated through criteria designed and an algorithm method as follows

- The churn behavior (Jahanzeb & Jabeen, 2007) defined is based on the consumption history during twelve months. Accordingly, telecommunication operators have different measures to classify the churn and non-churn customers. This paper is based on the criteria of 3K3D, which 3K means 3,000 VND[49] of describers' expenditure per month, while 3D is a measure of 3 days per month keeping the mobile on (not off). An identification to know whether the mobile of describer turning on or off is based on the system of the Visitor Location

[49] Currency: $1 = 26.200VND during the end of the year 2018.

Register (VLR)[50]. Any customer does not meet the criteria during three consecutive months to meet 3K and 3D, it is defined as the customer churn. This means the describer's expenditure is less than 3,000 VND per month and its mobile phone is off during the three consecutive days per month, every time VLR scanned and recorded.

- Offline call is generated based on number of messages sending-out or sending-in and number of calls going-out or going-in that the subscribers received or call-out as send-out to or from others subscribers who are using a service from different telecommunication operators. Once the one's offline call is high, it properly cause a churn risk and he or she can switch.

The information of feature selection is depicted in table as below

Table 1: Information of attributes and its transformed

Attribute	Description	Based on reference of
DATA	Data usage	(Ahn et al., 2006) and
LOS	Length of stay	(Rosset et al., 2003)
Top-Up	Understood as tariff that subscribers recharge its service	(Ahn et al., 2006), (Olle, 2014)
EXCOM	External communication based on offline call, which calls and messages out an in to/from others subscribers using different service operator	(Amin et al., 2017) and own ideas
Handset	Type of mobile that subscribers are using	(S. Y. Kim, Jung, Suh, & Hwang, 2006)
Churn	Based on calculation to meet 3K3D	(Jahanzeb & Jabeen, 2007)

Rough set theory and its application

(Amin et al., 2017) developed a model with respect to RST to predict customer churn in the telecommunication industry through a test of four methods, EA, GA, CA, and LEM2. The finding confirmed GA as the best accurate method, which GA is order-based couple with a heuristic. It is quite supported to reduct the computational cost in a large feature and complex decision table (Bazan, Nguyen, Nguyen, Synak, & Wróblewski, 2000). As a result, RST with rule generation algorithm is concerned in this study. An attempt to apply, this study is based references of Amin et al. (2017), (Arun Kumar et al., 2017), (Ning Zhong& Andrzej Skowron, 2001), and (Dong, Zhong, & Ohsuga, 1999).

[50] VLR is a database in a mobile communications network associated to a Mobile Switching Center. Mobile subscribers currently present the exact location in the service area, and extract subscribers who is turning or off its cell phone.

RST can be briefly as follows. Accordingly, a decision table is denoted T = (U, A, C, D), where U is universe of discourse, A is a family of equivalence relations over U, C and D are two subsets of features and so-called condition and decision features, respectively, they are contained in A (Pawlak, 1998).

Approximations: $\underline{R}X$ is the set of all elements of U, which can be with certainty classified as elements of X, in the knowledge R, where $X \subseteq U$. It can be depicted as below.

$$\underline{R}X = \cup \{Y \in {}^{U}/_{R} : Y \subseteq X\}$$

The positive region: This region is used to denote the low approximations of a set. Let Y and Z be equivalence relation over U, $Y \subset U$ and $Z \subset U$. The positive region of Z, POSY(Z), is the set of all objects of universe U which can be properly grouped into classes of U/Z employing knowledge expressed by the classification U/P.

$$POS_Y(z) = \bigcup_{X \in U/Q} \underline{R}X$$

Dispensable and indispensable features are defined as follows

Let $c \in C$. A feature c is dispensable in T, if POS(C-c)(D) = POSC(D). Otherwise feature c is indispensable in T.

Deleting c from T will cause T to be inconsistent if c is an indispensable feature. Conversely, c may be deleted T. If all $c \in C$ are indispensable in T, T = (U, A, C, D) is independent.

Reduct: A set of feature $R \subseteq C$ will be called a reduct of C, if *T = (U, A, R, D)* is independent and $POS_R(D) = POS_C(D)$.

Core: All indispensable features set in (C, D) will be denoted by CORE(C,D)

$$CORE(C, D) = \cap RED(C, D)$$

where RED(C,D) is the set of all reducts of (C,D)

All indispensable features should be contained in an optimal feature subset, because removing any of them will cause inconsistent in a decision table. As stated previously, CORE presents the set of all indispensable features. Therefore, finding CORE is depended on the process seeking indispensable features. Also searching CORE is based on the discernibility matrix proposed by (Skowron & Rauszer, 1992). A brief presentation below is the basic idea of the discernibility matrix.

Let $T = (U, A, C, D)$ as a decision table, with $U = \{u_1, u_2, u_3, \ldots, u_n\}$. By a discernibility matrix of T, $M(T)$ denoted as the matrix, which n x n matrix is defined as follows

$m_{ij} = \{a \in C: a(u_i) \neq a(u_j) \wedge D(u_i) \neq D(u_j)\}$ for i, j = 1,2,3,…n where m_{ij}

is the all attributes set that discern objects u_i and u_j.

The CORE can be denoted now as the set of all single element entries of the discernibility matrix, which CORE is defined as below

$$CORE(C) = \{a \in C: m_{ij} = (a), for\ some\ i, j\}$$

Considering CORE is to seek such a set of features in which each feature is unit to discern some objects.

IV. Empirical analysis

Many features of the telecommunication company are collected from big data, a selection of best features is very important, this is based on the evaluation criterion in an induction algorithms to gain an appropriate reduct (Amin et al., 2019). Initially, there are six attributes employed to take cut and discretization, the process of induction is a common approach used in the RST. As depicted in table 2, with the six attributes are concerned, in which the LOS and TOP_UP are continuous values, while DATAUSAGE, EXCOMM and HANDSET are nominal ones. All of them presents as condition attributes. Accordingly, the continuous variables are partitioned in to a finite value of groups or intervals. The decision attribute is CHURN.

Table 2: Attribute values and its sets

Name of attribute	Measure	Attribute value sets
Condition attribute		
(1) LOS: Length of stay	Months Min = 4.0; Max = 217; Mean = 61.2	Continuous variable
(2) TOP_UP (tariff)	VND/month Min = 0; Max = 13 million; Mean = 92,578	Continuous variable
(3) DATAUSAGE	YES and NO using data for third or fourth generation	Nominal variable: Y & N
(4) EXCOM: offline calls	LOW and HIGH	Nominal variable: HIGH & LOW
(5) HANDSET	Feature phone, Smart, Unknown	Nominal variable: F, S, U
Decision attribute		
(6) CHURN	Churn and No-churn	Nominal variable: CHURN, NOCHURN

Note: $1 = 23,300 Vietnamese Dong (VND)

According to the finding of (Amin et al., 2017), the genetic algorithm (GA) is the best method when compared with exhaustive algorithm (EA), covering algorithm (CA), and LEM2. This is proved by the authors through testing the method on the dataset of 3,333 instances, where customers are no-churn, accenting for 85.5%, while that of customers are churn, occupying 14.5%.

In line with this term, the method of GA is calculated in this paper by WEKA program with 211,777 instances, which customers with churn account for 13%, customers are non-churn as 87%. As resulted, four condition attributes are selected, they are LOS, EXCOM; TOP_UP; HANDSET, while DATAUSAGE is reduced. Continuously, calculation of GA derives reducts with probability of 60% with an initial population of 20 chromosomes. This chromosome is resulted from six features in the dataset. The GA converges in the span of 20 generations and produce an exhaustive rule base including more than one thousand and three hundred rules. Based on this proposed method, important decision rules are based on the training set. The decision rules present the rules in the form of "If C (conditional attribute) then D (decision attribute)". Based on this statement, we can get a association rule of condition attributes and decision attribute.

Table 3: GA Parameter for feature reduction

Parameter of GA	Value
Population size	20
Number of generations	20
Probability of crossover	0.6
Probability of mutation	0.033
Report frequency	20

Application of rule generation with RST based on a feature selection of GA, classifications of Naïve Bayes, J48 (Decision tree) and Decision table are concerned and tested. This application is an extent concern from (Arora, Asstt, & Cse, 2012), (Shaaban et al., 2012), and (Arun Kumar et al., 2017)

Accordingly results derived in table 4 is predicted from the sample of training sets. A comparison amongst these three classification methods based on the rule generation algorithm of GA with RST, which the finding confirms that Decision table as the best accuracy of 88.8%, while that of Naïve Bayes is 84.1%. Besides that, if comparing other criteria, such as Precision, Recall, FMeasure, the method of Decision table is the best choice. This means an appropriate genetic algorithm employs a candidate of Decision table. This is partly consistent to (Amin et al., 2017), but different from (Shaaban et al., 2012).

Table 4: Results by rules Generation of the GA through RST classification approach

Indicators	Naïve Bayes	J48 (Decision tree)	Decision table
Precision	0.837	0.871	0.871

Accuracy	0.845	0.887	0.888
Recall	0.845	0.887	0.888
F-Measure	0.841	0.871	0.870
Time to build (Seconds)	0.54	12.3	12.0

V. Discussion

With a review based o a combination of the finding of this paper with what (Amin et al., 2017) and others previous studies found, this study again confirms the GA is one of the best induction method to address a right feature selection, then to predict customer churn in the telecommunication sector. Evidently, the indicator values of precision, accuracy, recall, and Fmeasure are high, not so different from other scholars (table 5). In addition, the current study found that among classifier methods, such as Decision Table, Naïve Bayes, and Decision Tree (J48), Decision Tree is the most appropriate classifier to address rule generation reaching customer churn prediction.

Table 5: Performance predictive performance of this current finding and previous studies

Technique and reference	Precision	Accuracy	Recall	F-Measure
GA of Current finding	0.871	0.888	0.888	0.870
GA of (Amin et al., 2017)	0.860	0.981	1.000	0.925
(Shaaban et al., 2012)	0.842	0.837	0.834	0.838
(Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 2015)	0.917	0.969	0.875	0.846

Source: Based on (Amin et al., 2017) and current finding

VI. Conclusion

Up today, many scholars use many churn prediction models, but application of RST on practical dataset of the telecommunication sector is not widely used. This paper is an application of algorithms in machine learning program, e.g. WEKA to predict customer churn. To do this, 211,777 instances in the telecommunication sector are used in the method with the candidate of six attributes. With an approach of RST with supported machine learning, the paper found an accurate method and outperformed rules generation algorithms in terms of precision, accuracy, recall, and F-measure, which Decision Table is the best candidate to predict churn. The study also reveals the performance of machine learning algorithm will be a good reference for telecommunication provider to think of decision making to retention of customers who are churn risks

Reference

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, *30*(10–11), 552–568. https://doi.org/10.1016/j.telpol.2006.09.006

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*(October 2017), 290–301. https://doi.org/10.1016/j.jbusres.2018.03.003

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242–254. https://doi.org/10.1016/j.neucom.2016.12.009

Arora, R., Asstt, S., & Cse, D. (2012). Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. *International Journal of Computer Applications*, *54*(13), 975–8887.

Arun Kumar, C., Sooraj, M. P., & Ramakrishnan, S. (2017). A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets. *Procedia Computer Science*, *115*, 209–217. https://doi.org/10.1016/j.procs.2017.09.127

Bazan, J. G., Nguyen, H. S., Nguyen, S. H., Synak, P., & Wróblewski, J. (2000). Rough Set Algorithms in Classification Problem. In *Rough Set Methods and Applications* (pp. 49–98). https://doi.org/10.1007/978-3-7908-1840-6_3

Dong, J., Zhong, N., & Ohsuga, S. (1999). Using rough sets with heuristics for feature selection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1711*, 178–187. https://doi.org/10.1007/978-3-540-48061-7_22

Hassouna, M., Tarhini, A., Elyas, T., & Abou Trab, M. S. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research*, *8*(6), 224–237. https://doi.org/10.5539/ibr.v8n6p224

Jahanzeb, S., & Jabeen, S. (2007). Churn management in the telecom industry of Pakistan: A comparative study of Ufone and Telenor. *Journal of Database Marketing & Customer Strategy Management*, *14*(2), 120–129. https://doi.org/10.1057/palgrave.dbm.3250043

Kim, H. S., & Yoon, C. H. (2004). Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. *Telecommunications Policy*, *28*(9–10), 751–765. https://doi.org/10.1016/j.telpol.2004.05.013

Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, *31*(1), 101–107. https://doi.org/10.1016/j.eswa.2005.09.004

Nafis, N. S. M., Makhtar, M., Awang, M. K., Rahman, M. N. A., & Deris, M. M. (2016). Predictive modeling for telco customer churn using rough set theory. *ARPN Journal of Engineering and Applied Sciences*, *11*(5), 3203–3207.

Ning Zhong& Andrzej Skowron. (2001). *A ROUGH SET-BASED KNOWLEDGE 2 . Generalized Distribution Table and Rough Set System*. *11*(3), 603–619.

Olle, G. (2014). A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *International Journal of E-Education, e-Business, e-Management and e-Learning*, *4*(1), 55–62. https://doi.org/10.7763/ijeeee.2014.v4.302

Pawlak, Z. (1998). Rough set theory and its applications to data analysis. *Cybernetics and Systems*, *29*(7), 661–688. https://doi.org/10.1080/019697298125470

Rosset, S., Neumann, E., Eick, U., & Vatnik, N. (2003). Customer Lifetime Value Models for Decision Support. *Data Mining and Knowledge Discovery*, *7*(3), 321–339. https://doi.org/10.1023/A:1024036305874

Shaaban, E., Helmy, Y., & Khedr, A. (2012). A Proposed Churn Prediction Model. *Mona Nasr / International Journal of Engineering Research and Applications (IJERA)*, *2*(4), 693–697. Retrieved from www.ijera.com

Sharma, A., & Kumar Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. *International Journal of Computer Applications*, *27*(11), 26–31. https://doi.org/10.5120/3344-4605

Skowron, A., & Rauszer, C. (1992). The Discernibility Matrices and Functions in Information Systems. In *Intelligent Decision Support* (pp. 331–362). https://doi.org/10.1007/978-94-0157975-9_21

Trabelsi, S., Elouedi, Z., & Lingras, P. (2011). Classification systems based on rough sets under the belief function framework. *International Journal of Approximate Reasoning*, *52*(9), 1409–1432. https://doi.org/10.1016/j.ijar.2011.08.002

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

Vashist, R., & Garg, M. . (2011). Rule Generation based on Reduct and Core: A Rough Set Approach. *International Journal of Computer Applications*, *29*(9), 1–5. https://doi.org/10.5120/3595-4989

Accounting principles, disclosure and big data

Sabrina Pucci, Marco Venuti, Umberto Lupatelli (RomaTre University, Rome)

Extended Abstract

1. *Introduction*

The main scope of the financial statement is to provide stakeholders with qualitative or quantitative information useful in making economic decisions (Conceptual Framework for Financial Reporting, 2018). This principle has inspired IASB and other standard setters and accounting bodies to look for indications that permit shareholders to have more useful, complete, reliable, and unbiased information on a company's business. Principles like transparency, neutrality, comparability and fairness of the financial and non-financial information are becoming increasingly important under pressure from the market, investors, clients, workers, supervisory authorities, etc.

Disclosure is at the centre of a recent IASB Project: Discussion Paper on Principle of Disclosure (2017), which attempts to find an appropriate level of information for stakeholders that use financial statements to take their decisions.

In financial statements, some data is objectively measurable (for example cash or the amounts of bank accounts), while other data is estimated (inventory and related evaluation hypothesis such as FIFO and average cost) and other data strictly depends on the judgment applied by managers in the evaluation process (for example the impairment amount). It should be remembered that yearly return is an abstract item because the only objective return is the amount you can calculate after the liquidation of a company. The yearly return is affected by hypothesis, practical wisdom and judgement.

2. *Body of knowledge*

In some recent studies, accounting is combined with big data, with new technology instruments and with analytics (Amani F.F., Fadlalla A.M., 2017, Rikhardsson P., Yigitbasioglu O, 2018, CPA, AICPA, 2019). Some papers have highlighted that big data combined with sophisticated algorithms can improve management reporting (Cockcroft S., Russell M., 2018) permitting a better comparison between financial and non-financial variables. Others consider that big data should be evaluated as an asset, especially in the companies of the financial sector (Turner D., Schroeck M., Shockley R., 2013) or that big data should be an important way to increase the integrity and quality of accounting processes (CGMA, 2013).

Remembering that the Gartner model, formalized in 2012 and better defined later, assigns three main characteristics to big data: "*high volume, high velocity, high variety of information assets*", another important question is what will be the impact of new technology environments and instruments on financial statements, accounting principles and disclosure level presented to stakeholders. Will the use of big data and analytics be only an advantage or are there some threats to be managed?

We begin our study with a paper presented in September in Turin with a qualitative analysis of the replies to the Disclosure Initiative Paper of IASB. In this paper, we found that a lot of respondents ask for a specific replacement of technology issues referring both to the possibility to create a digital financial statement framework and "*technology neutral guidance*" and the possibility to insert non-financial data with financial information referring to the history of the company and the forecasts of long term sustainability of the business.

The aim of the current stage of this research is to evaluate whether some prescriptions of current accounting principles will be obsolete using big data and analytics or if they should only be updated, or maintained in their present form but improving the quality of the data they express. We also evaluate whether the principle-based approach used by IASB is neutral in comparison with the technology used for the preparation of financial statements and with the way in which data is collected and analysed. Analysed examples of accounting evaluation and principles that can be affected by big data to respond to the research questions are:

- Impairment test IAS 36 ○ Stakeholders are particularly concerned about the cost, complexity and high level of judgement used in the impairment test. In particular, critical aspects regard inputs and methods to determine the recoverable amount of a cash generating unit. Big data can be useful in the calculation of value in use especially in assessing the discount rate, in determining expected cash flows in multiple scenarios and terminal value;
- IFRS 17 ○ New IFRS 17 establishes a model to evaluate technical reserves that requires an important use of data, forecasts and judgement referring to both the amounts of liabilities (interest rates, future trend of contractual service margin, loss estimates) and the amount of insurance revenues (for all the estimates of onerous and non-onerous contracts with the use of annual cohorts and coverage service period);
- Disclosure ○ One of the main aspects could be the coexistence of paper financial statements with digital financial statements. These can create a binary system: one for traditional stakeholders and one for digital stakeholders, with important implications for transparency, supervision and level of information. Another aspect is the chance to present non-financial information in financial statements and the possibility to audit this information and the necessity to highlight auditable and non-auditable information.

3. Conclusion

The conclusion we expect to find is that most accounting principles could maintain their validity in substance in a regulatory framework but the level of information and the processes behind the numbers presented in financial statements could be improved with better estimates for uncertain variables using big data and new algorithms to create possibilities for more motivated judgement. However, at the same time, we are confident that judgement will remain an important characteristic of the financial statement owners' decision process.

Some other research questions that can be added to our analysis are connected with the way in which big data will be managed. Is it possible to use every kind of data independent from the devices that collected this data in the accountability processes? How will it be possible to harmonise this different data considering the different organizations, dimensions and sectors of the companies? Which possibilities of errors do these processes involve and how do they influence data quality? Some others regard the impact on models and

analytics. Will stochastics evaluation of some financial statement amounts be replaced by the use of analytics? What percentage of errors can affect analytics and big data? How could the behaviours of investors, clients, etc. affect the evaluation of financial statement items? How could non-financial information influence the value of financial information?

References

Amani F.F., Fadlalla A.M., *Data mining applications in accounting: a review of the literature and organizing framework*, International Journal of Accounting Information System, n.24, 2017, pagg.32-58

Cockroft S., Russel M., *Big data opportunities for accounting and finance practice and research: big data in accounting and finance*, Australian Accounting review, February, 2018

CGMA, *From insight to impact: unlocking opportunities in big data*, Report of the Chartered Global Management Accountant, AICPA, 2013

CPA, AICPA, *A CPA's Introduction to AI: from algorithms to deep learning, What you need to know,* 2019, www.cpacanada.ca

Gartner, *The importance of big data: a definition*, October 2013, www.gartner.com

IASB, *Conceptual Framework for Financial Reporting*, 2018, www.iasb.org

IASB, *Disclosure Initiative*, 2017, available on www.iasb.org

ICAEW, *Data Analytics for external auditors*, International auditing perspective, 2016, www.icaew.com

IMA, ACCA, *Big data: its power and perils*, November 2013, www…..

Rikhardsson P., Yigitbasioglu O., *Business intelligence & analytics in management accounting research: status and future focus*, International Journal of Accounting Information System, n.29, 2018, pagg.37-58

Turner D., Schroeck M., Shockley R., *Analytics: the real world use of big data in financial services*, IBM Global Business Services, 2013