Volume 1, 2016

Proceedings on Big Data Analytics & Innovation

(Peer-Reviewed)

4TH Big Data Analytics and Innovation Conference

(DATE IN 2017 & VENUE IN NIGERIA TO BE CONFIRMED IN DUE COURSE VIA WWW.CONFERENCE.BIGDATANIGERIA.ORG)

Thematic Sections/Tracks:

Big Data, E-Governance, Public Service Delivery & ICT

Big Data, Telecommunications & National Security

Big Data, Energy, Oil & Gas & Environment

Big Data, Agriculture & Food Security

Big Data Analytics, Innovation & Enterprise

Big Data Analytics & Higher Education

*There will be **opportunities to publish papers** in the Conference volume—**Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 2 2017**—and selected papers might be published in the associated journal of the conference—see www.ijkie.org*

Proceedings

on Big Data Analytics & Innovation

(Peer-Reviewed)

2016 KIE Publications:
Proceedings on Big Data Analytics & Innovation (Peer-Reviewed)

Produced and Published in London by the KIE Publications
Printed and Bound in Great Britain by CDS Corporate Document Services, Leeds, England, United Kingdom

$15.95

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed)

Peer Review Process

All articles in the *Proceedings on Big Data Analytics & Innovation (Peer-Reviewed)* have undergone peer review, based on initial editor screening by at least two anonymous referees.

CONTENTS

CONTENTS

EDITORIAL NOTE

'Ever-expanding Application of Big Data Analytics'

The Organising Team of the Big Data Analytics and Innovation Conference is very delighted to publish this volume—*Proceedings on Big Data Analytics and Innovation* —as part of the 2016 edition of the conference. It is also a real privilege for us to have a wide range of subject specialists—academics and practitioners—to contribute to this volume.

The wide range of topics covered in the volume are a testament to the ever-expanding application of Big Data Analytics—be it in education, as Etuk et al. paper shows, or in telecommunications, as demonstrated by a collection of papers from colleagues at the FUT Minna, or in national security and crime prevention, as Ujah and Adejoro's paper show—it is clear that Big Data is a low-hanging fruit that is just waiting to be plucked, which is applicable to just about any aspect of our national live and national economy.

Of course much has been said and written about Big Data and I do not intend to deal with the subject in any comprehensive way here except to say that the term 'big data' describes the innovations that surround the possible uses of digital information and that big data is now a significant feature of enterprise computing.

We are witnessing an explosive growth in digital and physical data and in the number of different types of datasets in the public domain. These datasets—especially unstructured or non-metric data—are huge and complex in volume, velocity, variety and veracity they threaten the capability of modern data processing and analytic tools.

Nigeria has an estimated US $200 billion Big Data market (Akwaja, 2014). The size of the Big Data market was approximately 58% bigger than what the country earned from oil in 2013 and estimated to rise to 70% in 2016. But despite Nigeria's comparatively huge Big Data market, data assets in key sectors of the economy such as the telecommunications, financial services, energy, oil and gas are still lying comparatively dormant as are in much of corporate Nigeria. Within the Nigeria's SMEs community, Big Data is unheard off, arguably. So, with dwindling foreign exchange reserves and collapsing oil prices, harnessing big data opportunities seems a sensible way to go.

Finally, I'm grateful to all the authors in creating time from their very busy schedules to contribute to this maiden volume. Thank you.

James Ogunleye, PhD, FRSA
Conference Chairman

Reference

Akwaja, C. (2014) 'Big Data: Nigerian Operators Jostle For $200bn Opportunities', *Leadership Newspaper (*Business section*)*, May 19, available at: http://leadership.ng/business/371431/big-data-nigerian-operators-jostle-200bn-opportunities; accessed: 25/10/14.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp. 8-20

Big Data in Big Giant in Big Continent

UGOCHUKWU ONWUDEBELU
Federal University Ndufu-Alike Ikwo, Nigeria

SANJO FASOLA OJENIYI
University of Ibadan, Nigeria

ADEBAYO JOSEPH
Federal University of Technology, Minna, Nigeria

ABSTRACT Welcome to the age of Big Data, where the quantity of data in the world is soaring very high on a second to second basis and in both scope and power. Indeed, Big Data has come to stay and there is no turning back since data is key to change and it is at the centre of knowledge-based economic and society. In our time, an enormous amount of data is recurrently being generated and flowing from various sources, through different channels in this digital age. Our world revolves around data and hence come the revolution. With the world experiencing a data revolution or *"data deluge",* we must be ready for changes as big data technologies are evolving rapidly. This data revolution has not just restricted itself to the industrialised world rather it is also being experienced in developing countries. The change and development we yearn for in Nigeria is data-dependent rather than oil-dependent as has been the case in the past decades. However, many organizations are rich in data but poor in insight. These data hold the potential, as yet largely untapped, to allow decision makers to track development progress. Consequently, we need to turn-on our own tap of innovation and socio-economic development through Big Data. With the promise come questions about the analytical value and thus, policy relevance of this data, especially, concerns over the relevance of the data in developing country contexts. This paper does not offer a grand theory of technology-driven social change in the Big Data era. Rather it aims to describe the big impact of Big Data and to suggest ways to address at least a few aspects of concerns raised here. This paper is also a call-to-action for stakeholders for intensive action to ensure that big data helps the individuals and communities who create it as well as the government at large. The paper concludes that Big Data will help deliver better services, improve efficiency of operations and enhance communications between the governing and the governed. Furthermore, the economy of a nation can greatly be raised by the Big Data paradigm shift, which will result in not only seeing Big Data as the oil rather as the air, that is, you cannot breathe without it.

Keywords: Big Data, prediction, open government, cloud computing, data deluge, Nigeria, Business Intelligence

Introduction

Big Data is the extremely large and varied datasets that may be analyzed to reveal patterns, trends, and associations; they are often too large to store, process and analyze with traditional storage and computing methods (Davenport, 2012). Gartner refers to data sets that have "high variety, volume and velocity as Big Data" *(*Franks, 2012). Consequently, Big Data is an umbrella term for the explosion in the quantity and diversity of high frequency digital data. Nevertheless, some have called it, *the next frontier* (Harrison and Hrdinova, 2013) for advances in organizational innovation and productivity. For others, the term was spoken with a bit of awe and pushiness, and always a breathlessness about what incredible potential big data held for making our world more understandable and more fixable (Kitner and Thea, 2015). Big Data bears the promise of enabling policy makers and organizational leaders to extract knowledge from the torrents of data now generated through business transactions, network-based human interaction, or by increasingly ubiquitous sensors monitoring events in the physical, technological, and social world (Harrison et al, 2013). It should be recalled that the fundamental reason for using big data is to find a way to present any data in a usable form to solve problem.

In 2010, there were over five billion mobile phones in use all over the world, and of those, over 80% in developing countries (unglobalpulse.org) and the number continues to rise quickly. The spread of mobile-phone technology to the hands of billions of individuals generates a lot of data when such data are analysed would help in knowing the individuals more. This mobile phone technology has been used as a substitute for usually weak telecommunication and transport infrastructure in addition to underdeveloped financial and banking systems in Africa. The promises and potential of Big Data in transforming digital government services, governments, and the interaction between governments, citizens, and the business sector, are significant (Bertot and Choi, 2013). As the digital data universe grows bigger by the day local and national governments are embracing innovative methods to harness Big Data technology to better serve citizens (Ben-Zvi , 2016).

One clear example of Big Data is the Square Kilometer Array (SKA) (www.skatelescope.org) planned to be constructed and completed in South Africa and Australia (Harris, 2012) by 2024. If completed, the SKA will be able to produce in excess of one Exabyte of raw data per day, which is more than the entire daily internet traffic at present. The SKA is a 1.5 billion Euro project that will have more than 3000 receiving dishes to produce a combined information collecting area of one square kilometer, and will use enough optical fiber to wrap twice around the Earth. Another example of Big Data is the Large Hadron Collider, at the European Organization for Nuclear Research (CERN), which has 150 million sensors and is creating 22 Petabytes of data in 2012. Many projects fail due to neglect of invisible data work (Kitner et al, 2015).

Key Sources of Big Data

Human-Generated: These are made possible through the following activities:

a. Online behaviours such as orders, transactions, payment history, usage history etc.
b. Online attitude such as opinions, preferences, interests, needs and desires, social media posts etc.
c. Online user-generated content such as blog posts and Tweets, online searches, satellite images, online videos etc.
d. Interaction data such as click digital pictures, streams, email and chat etc.
e. Mobile phone interactions: call logs, mobile-banking transactions, cell phone GPS signals etc.

Sensors-Generated: These are made possible through the following activities:

a. Wearable devices such as calculator watch, wristband, Google glass, digital eye wear, smart watches, domestic robot, etc.
b. Home Appliances such as refrigerator, toaster, dishwasher, clothes dryer etc.
c. Climate Information such weather, traffic, etc.

A Deluge of Data: The Promise of Big Data in Nigeria

We are faced with a deluge of data that, when combined with new technologies and analysis techniques, has the potential to inform decision and policy making in unprecedented ways (Bertot et al, 2013). This data deluge has significantly be promoted by the advent of social network websites, where users replicate an electronic version of their lives by generating records of their lives by daily posting details of activities they performed, events they attended, places they visited, pictures they took and things they enjoyed and fancied. This data deluge is often referred to as Big Data (Bell, Hey, and Szalay, 2009; Franks, 2012). The striking thing about this data deluge is the speed and frequency with which data is emitted and transmitted as well as the rise in the number and variety of sources from which it emanates. Increasingly, the research and scientific communities, governments, and the private sector are generating large-scale data sets on a range of topics, including traffic patterns, disease data, purchasing behaviour, and social behaviour through social media interactions (Bertot et al, 2013). Furthermore, other areas where a deluge of data is being generated include: Health data, environment such as climate change, emergency response and disaster resiliency, manufacturing, robotics and smart systems, secure cyberspace, transportation and energy, education, and workforce development. Thus, we are cumbersome with a lot of data which if not appropriately harness will lead to information overload. In short, our ability to harness Big Data have the potential of reducing information overload, leading to new scientific

and research insights, create economic development, and generate new policies that benefit the public served by governments.

Figure 1: Data Deluge (*Source*: Ben-Zvi, 2016)

Nigeria, the giant of Africa, both in terms of economy and population (with about 180 million people) needs to harness the benefits of Big Data as well as cloud computing to remain at the top. The challenges of data collection and mining are surely a challenge particularly with various organs of government that are involved in the management and usage of data for different purposes. Some of the agencies saddled with the responsibility to collect and manage data in Nigeria include (Hilbert and Lopez, 2011):

i. Federal Road Safety Commission (FRSC) responsible for drivers' license and vehicle number plates;
ii. Independent National Electoral Commission (INEC) responsible for voters registration exercise.
iii. National Bureau of Statistic (NBS) responsible for production of national official Statistics;
iv. National Identify Management Commission (NIMC) accountable for national identity database
v. National Population Commission (NPC) in charge of national demographic data.
vi. Other organizations including the banks in the financial sector and Telecommunication companies in the telecommunication sector such as MTN Nigeria, Globacom, Airtel, Etisalat etc.

Over 80% of World's Data are currently Unstructured (Iyilade, 2015). Most of these data collected by these organizations (FRSC, INEC, NBS, NIMC, NPC, etc.) are structured in nature and are very essential for e-governance services. Nevertheless, the replication of the same data by different government agencies poses great threats to security and privacy concerned. The collected data is used merely for statistical purpose and not solving any mission critical challenge that can improve the quality of the government service or any private sector challenges. The Nigeria government needs to have an open access to data across Government ministries just as is seen in most of the developed world – US, Canada, Britain, Australia, France, Japan - that has a national agenda and initiatives on Big Data. This massive amount of data should be recognized as a national asset that can improve the economy of Nigeria, which in time past has sorely depended on oil.

Data is the New Oil

Data is the new '*Oil'* of the 21st Century with potential to spur innovation and socioeconomic development in many sectors (World Economic Forum, unglobalpulse.org). Like oil, data must be refined before it can be used. This oil can be found at the Federal, State and Local government levels, in fact, it is located in each state, thus the need for Nigeria as a country to digitize all government departments, scheme and services in the three tiers of government. The amount of data is expected to grow as new technologies are adopted resulting in an equivalent increase in the amount of both structured and unstructured data available outside government circumference. The application of Big Data analytics to this growing resource can increase the value of this asset to government and the people seeing that data is an asset. Therefore, there is the need to have national ICT strategy that will identify the need to have a framework for Big Data analytics so as to further develop government capability in Big Data exploitation.

Table 1: Units of Information

Value	Metric	Overview of Data Scale from Kilobytes to Yottabytes	
1000	KB	Kilobyte	1,000 bytes
1000^2	MB	Megabyte	1,000,000 bytes
1000^3	GB	Gigabyte	1,000,000,000 bytes
1000^4	TB	Terabyte	1,000,000,000,000 bytes
1000^5	PB	Petabyte	1,000,000,000,000,000 bytes
1000^6	EB	Exabyte	1,000,000,000,000,000,000 bytes
1000^7	ZB	Zettabyte	1,000,000,000,000,000,000,000 bytes
1000^8	YP	Yottabyte	1,000,000,000,000,000,000,000,000 bytes

The amount of available digital data at the global level grew from 150 exabytes (see Table 1) in 2005 to 1200 Exabytes in 2010 (Dirk and Balietti, 2011). It is projected to increase by 40% annually in the next few years (Manyika et al, 2011), which is about 40 times the much-debated growth of the world's population (UNDP, 2010). This rate of growth means that the stock of digital data is expected to increase 44 times between 2007 and 2020, doubling every 20 months (unglobalpulse.org). Accord to a pictorial estimate made by Hilbert and Lopez (2011), if all the data used in the world today were written to CD-ROMs and the CD-ROMs piled up in a single stack, the stack would stretch all the way from the Earth to the Moon and a quarter of the way back again. In terms of figures, the International Data Corporation (IDC, 2010) in their report in 2010 estimated that by the year 2020 there will be 35 Zettabytes of digital data created per annum.

An organisation willing to use analytics technology frequently acquires expensive software licences; employs large computing infrastructure; and pays for consulting hours of analysts who work with the organisation to better understand its business, organise its data, and integrate it for analytics (Sun et al, 2011). This joint effort of organisation and analysts often aims to help the organisation understand its customers' needs, behaviours, and future demands for new products or marketing strategies. Such effort, however, is generally costly and often lacks flexibility. Nevertheless, research and application of Big Data are being extensively explored by governments, as evidenced by initiatives from USA (science.house.gov) and UK (www.cs.ox.ac.uk); by academics, such as the bigdata@csail initiative from MIT; and by companies such as Intel (istc-bigdata.org).

The Nigeria government needs to support the idea of creating of a Nigerian Big Data Analytics Service Network (NBDAS) to increase exchanges between data scientists and businesses in order to improve skills capacities within the software industry which will lead to the next generation of data-analysis products and applications which will be similar to that of European Big Data Analytics Service Network (EBDAS) (NESSI, 2012).

Discourse on the Extent to which Big Data Analytics is used in Nigeria

Nigeria is gradually making a foothold with respect to the issues of Big Data analytics although it is still struggling with it because of many limiting factors as mentioned Table 2. Since 2014, Nigeria operators are jostling for $200 billion market values associated with big data (Akwaja, 2014). Consequently, Globalcom, MTN Nigeria, Airtel Nigeria as well as Etisalat Nigeria are already deploying the next generation networks and Business Intelligence software tools to tap into the big data opportunities. For instance, Globalcom is spending $ 1.25 billion to build an Internet Protocol network, MTN Nigeria is spending over $ 1 billion yearly on its network, Airtel Nigeria and Etisalat Nigeria are both spending over $ 3 billion collectively to retool their network to harness into the big data prospects. These reactions to emerging technologies and leverage tools like big data are all geared at gaining a competitive advantage. Furthermore, with the fantastic yearly income that averages about N 11.8 million, Big Data is expected to catalyse the growth of the economy and thus boost its gross domestic product (GDP). While it has been

projected to create over one million jobs in the United State experts say in Nigeria Big Data promises to create twice that (Ajanaku, 2015). Soft Solutions Limited is company that is involved in Big Data in Nigeria while a good example of companies compiling data for their use in the country includes Renaissance Capital among others.

Turning Data into Products

We have entered the era of high rate of production of information of physical, biological, environmental, social and economic systems. For data to be helpful to any government there is need to turn the data into finish product, as was noted by O'Reilly Media when it declared: The future belongs to the companies and people that turn data into products. Big Datasets should be used to inform funding and science policy decisions. Just as the bibliometricians have successfully made use of the data deluge to make enormous advances in understanding how to manage scientific documents (Lane, 2012), Nigeria needs to use the data deluge to make enormous advances in its economy. There is a link between data and information, and the way in which that information is created and transmitted, and hence used to generate scientific, social, economic and workforce products.

Big Data brings the following benefits data view, changing thinking methods and tools, expanding the application scene, providing better service to the society, enhancing the value of the opportunity. The Big Data is changing thinking methods and tools, bring data intensive science and data exploration science (Miao, 2014; Shukla, 2015).

SWOT Analysis

SWOT is a study undertaken by an organization to identify its internal strengths and weaknesses, as well as its external opportunities and threats. SWOT stands for **S**trengths, **W**eaknesses, **O**pportunities, **T**hreats, and it is one of the simplest tools for mapping the signals relevant to change in the business or society is the SWOT analysis ((Tidd, Bessant and Pavitt, 2005). It is a simple, structured way of exploring the key challenges facing an organization and consequently a nation. The SWOT analysis in Table 2 reveals several opportunities but limited strengths. It is beneficial for Nigeria to make the most of the strengths and opportunities when directing efforts and crafting policies. A number of the identified threats can be limited through appropriate legislation.

Table 2: SWOT Analysis for Big Data Applications in Nigeria

Strengths	***Weaknesses***
a. Advantage of later adopter. b. Changing thinking methods and tools. c. Expanding the application Scene. d. Growing interest from companies using cloud computing and software services.	a. Insufficient knowledge and understanding on how to gain new insights by using data analytics on citizens, customers, products and services. b. Big Data as a Service in Nigeria is currently non-existing c. Very few research organizations are known for their activities and initiatives in the field of Big Data. d. Close access to data, in other word, no wide spread knowledge about freely available data. e. Ineffective research and development networks between universities, research centers and industries are being established.
Opportunities	***Threats***
a. Providing betters services to the society. b. Enhancing the value of the opportunity. c. Rising citizen demands in Nigeria for smarter products. d. Better use of freely available data. e. Developing new products and services enhanced with Big Data analytics and privacy by design.	a. There are serious concerns of expectations not being met and security and privacy threats rising. b. Insufficient capabilities for companies in Nigeria to scale to the world market. c. The fear of using big data analytics while neglecting privacy standards. d. There is a globally fast growing knowledge about using data as an asset class for leveraging the industrial competitiveness.

The is a need for Nigeria to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data just like other governments (developed World) such as US, Canada, Britain, Australia, France, and Japan have done. However, the potential value of Big Data will not be achieved without first overcoming barriers such as monetary cost of managing, collecting and maintaining such data (Harrison et al, 2013).

Challenges for Big Data Uptake in Nigeria

Big Data presents Nigeria many opportunities to influence every sector and sphere of life. Nevertheless, it also presents her with some challenges at the same time. These challenges include:

Data Inaccessibility

Today a huge amount of data which is easily accessible in electronic form is produced by both research and human activity as shown in section 1.2. Data accessibility is a problem in Nigeria, there are various hurdles of different forms in accessing data form FRSC, INEC, NBS, NIMC, NPC etc. Consequently, there is a need to take significant steps in promoting open data initiative in government ministries and parastatals so as to encourage innovation by providing access to public data. Examples of countries with their open access sites include: Canada - http://open.canada.ca/en, UK - https://data.gov.uk; US - https://www.data.gov etc. These steps when taken will make public datasets to be available and accessible for analysis and use. Some countries across the globe have begun adopting norms to standardize and publish datasets, they include: Kenya, Norway, Brazil, South Korea, and international institutions like the World Bank.

Lack of a National Strategy and Initiative on Big Data

The amount of data in the world is exploding - large portion of this comes from the interactions with the Internet and electronic devices. This huge volume of data is pregnant with meaning and hidden treasures if exploited. The lack of a national strategy and initiative on big data is actually affecting the economy of Nigeria indirectly. Most of the developed world such as Australia, Britain, Canada, France, Japan and US has a national agenda and initiatives on Big Data. These have helped them greatly in generating and spur innovation in many sectors of their economy. Nigeria likewise should developed a national strategy and initiative on Big data that will help in advancing various sectors in our nation's economy – agriculture, business, climate, ecosystems, education, energy, finance, health, local government Admin, manufacturing, telecommunication and transportation.

Privacy and Security of the Data

The Nigeria government in undertaking projects in Big Data must always have these two words at heart: the *privacy* and *security* of the data. The value of data (both in terms of personal value and the data's circulatory value) should be calculated; and the simple cost of data streams and access to the same must also be adequately accounted for (Kitner et al, 2015). One of the most challenging scenario, however, is the one represented by *analytics over big data* where analytics procedures run on large-scale amounts of distributed big data, hence leading to critical privacy and security breaches in cloud-assisted outsourced databases (Cuzzocrea, 2014).

Scarcity of Skilled Personnel

Gartner revealed that software vendors and service providers have cited scarcity of skills as the most important impediment in developing Big Data initiatives (Kart,

Heudecker, and Buytendijk, 2013). Nigeria, the giant of Africa, especially in terms of population, produces large and varied dataset which is useless without skilled personnel to harness and generate value from it. As a result, skills for data science and associated skills which are at present very low are in high demand. For instance, currently, there is a large shortage of skilled data scientists globally, estimated at 140,000 in next five years in US. Skills in Big Data Analytics and Visual analytics (visual design skills) are very much needed. With regards to visual analytics, skills in it help to promote the need for understanding the relevance and relatedness of information. Moreover, there is a need to know more about appropriate visual representation of different types of data at different spatial and temporal scales (NESSI, 2012). One of the ways of reducing skill shortages is by establishing innovation networks and spreading knowledge with online trainings as well as generating hands-on expertise based on commoditized analytical services.

Recommendations

i. Nigeria needs data privacy laws and a privacy commission to provide oversight on privacy implications of how personal data is collected and used by government and businesses so as to reduce privacy concerns and security threats raised above.
ii. The Nigeria government needs to support the idea of creating of a Nigerian Big Data Analytics Service Network (NBDAS).
iii. Nigeria needs to encourage the development of infrastructure and tactical approaches to Big Data analytics for universities' depositories.
iv. We need to encourage early education on data science and training in appropriate computer programming, machine learning, and Mathematics/ Statistics.
v. Our Universities need to be offering specialized Masters Degree in Big Data Analytics.

Conclusion

Researchers and policymakers are beginning to realize the potential for channelling these torrents of data into actionable information that can be used to identify needs and provide services (Bertot et al, 2013). As Big Data is growing and changing the mainstream business agenda, especially in the developed world, governments of the developing countries must realize its great potential as a key decision maker and tap into it. The cost of data is high but still desirable. Face with the situation of *more is less*, where the more data we collect, the less is our ability to derive actionable information from the data. It is necessary for us to move to *more is more*. The more data we collect, the more value we derive from the data. There is a need to exploit the huge hidden potentials embedded in Big Data analytics, so as to harness its benefits for the economy development of our nation. Big Data will help deliver better services, improve efficiency of operations and enhance communications between the governing and the governed. Furthermore, the economy of a nation

can greatly be raised by the Big Data paradigm shift. As this paper argues, the relevance of *Big Data in Big Giant in Big Continent* for a nation's development hinges primarily on the type of analysis it is subjected to and the use that is made of the resulting information. If properly harness, big data will become plucking the diamond from the waste.

Limitation of paper

The limitation of this paper of course is not being able to explore more Big Data companies so as to ascertain their distributions in the country at large. Generally, most data in Africa is largely offline bringing it online requires astronomical undertakings with enormous funding that private companies simply do not have. Therefore, we would have also loved to carry out a survey to learn more on the hindrances that are peculiar to the private sector as some companies have data they are not prepared to share. This data, if shared can be something useful to our nation.

Correspondence
Ugochukwu Onwudebelu
Department of Computer Science
Federal University Ndufu-Alike Ikwo (FUNAI)
P.M.B. 1010, Abakaliki, Ebonyi State, Nigeria

Sanjo Fasola, Ojeniyi
Department of Computer Science
University of Ibadan, Ibadan, Oyo State, Nigeria

Adebayo Joseph
Cyber Security Science Department
Federal University of Technology
P.M.B. 65, Minna, Niger State, Nigeria

References

Ajanaku, L. (2015) Big Data, Big Jobs. The Nation Nigeria. http://thenationonlineng.net/big-data-big-jobs/ accessed on Thursday, 15th September 2016

Akwaja C. (2014) Big Data: Nigeria Operators Jostle for $200 bn Opportunities, Leadership Newspaper, http://leadership.ng/business/371431/big-data-nigerian-operators-jostle-200bn-opportunities, accessed on Thursday, 15th September 2016

Bell, G., Hey, T. and Szalay, A. (2009) Beyond the Data Deluge, Science 323 (5919) 1297–1298.

Ben-Zvi G. (2016) Can Big Data Better Serve the Public? http://sqream.com/can-big-data-better-serve-the-public/

Bertot, J. C. & Choi, H. (2013) Big Data and e-Government: Issues, Policies, and Recommendations, *the Proceedings of the 14th Annual International Conference on Digital Government Research,* ACM, pp. 1-10

Cuzzocrea, A. (2014) PSBD 2014: Overview of the 1st International Workshop on Privacy and Security of Big Data, *CIKM '14*, ACM, 2100-2101.

Davenport, T. (2012) Enterprise Analytics Optimize Performance, Process and Decisions through Big Data. *FT Press*: 30-45.

Dirk, H. and Balietti, S. (2011) From Social Data Mining to Forecasting Socio-Economic Crises. *Arxiv* (2011) 1-66. 26 Jul 2011 http://arxiv.org/pdf/1012.0178v5.pdf

Franks, B. (2012) Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, 1st ed., in: Wiley and SAS Business Series, Wiley.

Harris, R. (2012) ICSU and the Challenges of Big Data in Science, *Research Trends Issue 30*, 11-12

Harrison, T. M. and Hrdinova, J. (2013) The Complexities of "Big Data": The Opportunities and Challenges for e-Government, *The Proceedings of the 14th Annual International Conference on Digital Government Research*, ACM, 263-264.

Heudecker, N. (2013) Hype Cycle for Big Data. Gartner G00252431

Hilbert, M. and Lopez, P. (2011) The world's Technological Capacity to Store, *Communicate and Compute Information,* Science 332, 1 April 2011, 60-65.

http://bigdata.csail.mit.edu/ bigdata@csail

http://istc-bigdata.org The Intel Science and Technology Centre for Big Data

http://science.house.gov/press-release (2013) Big Data has Big Potential to Improve Americans' Lives, Increase Economic Opportunities, Committee on Science, Space and Technology.

http://unglobalpulse.org/ (2012) Big Data for Development: Challenges & Opportunities.

http://www.cs.ox.ac.uk/news/639-full.html (May 2013) Prime Minister joins Sir Ka-shing Li for launch of £90m initiative in big data and drug discovery at Oxford.

IDC (2010) IDC Digital Universe Study, sponsored by EMC, May 2010, available at http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

Iyilade, J. S. (2015) Big Data & Analytics: Opportunities and Challenges, *12th NCS International Conference Information Technology for Inclusive Development*.

Kart L., Heudecker N. and Buytendijk F. (2013) Survey Analysis: Big Data Adoption in 2013 Shows Substance Behind the Hype, Gartner G00255160

Kitner, K. R. and Thea de Wet, (2015) Big City, Big Data, *Forum Evaluation and Usability*, Siegel, D. and Dray, S. (Editors), *Interactions*, 70 – 73.

Lane, J. (2012) Big Data: Science Metrics and the black box of Science Policy, *Research Trends Issue 30*, 7-8

Manyika, J., Chui, M., Brad B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H. (2011) Big data: The next frontier for innovation, competition, and productivity, *McKinsey Global Institute* (2011): 1-137

Miao, X. (2014) Big Data and Smart Grid, *Big Data Science '14*, Beijing, China, ACM

NESSI (2012) Big Data White Paper: Big Data A New World of Opportunities.

Salisu K. (2015) E-government Adoption and Framework for Big Data Analytics in Nigeria National Information Technology Development Agency (NITDA).

Shukla, S. K. (2015) Editorial: Big Data, Internet of Things, Cybersecurity - A New Trinity of Embedded Systems Research, *ACM Transactions on Embedded Computing Systems*, Article 61, 14 (4).

Sun, X., Gao, B., Zhang, Y., An, W., Cao, H., Guo, C. and Sun, W. (2011) Towards delivering analytical solutions in cloud: Business models and technical challenges, in: *Proceedings of the IEEE 8th International Conference on e-Business Engineering (ICEBE 2011)*, IEEE Computer Society, Washington, USA, 347–351.

Tidd J., Bessant, J. and Pavitt, K. (2005) Managing Innovation Tools: SWOT Analysis. www.wileyeurope.com /college/tidd, 2005.

UNDP (2010) World Population Prospects, the 2010 Revision, *United Nations Development Programme.* http://esa.un.org/unpd/wpp/unpp/panel_population.htm

Proceedings on Big Data Analytics & Innovation, (Peer-Reviewed), Volume 1, 2016, pp.22-34

Innovating with Data: the Aquila Technology Case in Petroleum Equalization Fund (Management) Board

ABUBAKAR, K, JUDE, J., YUSUFF, A.S., & EMMANUEL, O.G
Federal Ministry of Science and Technology, Abuja, Nigeria

ABSTRACT This study appraised the impact of data pool on innovation capability of Aquila Technology in the Operations of Petroleum Equalization Fund (Management) Board [PEF (M)B], Nigeria. Factors bothering on positive impacts made by data gathered using the Aquila Technology Platform in relation to the efficiency of tasks performed by the Operations Department of PEF(M)B were studied. The Narrative-Textual Case Study (NTCS) in amalgam with a tailor-made questionnaire and interview, administered on 140 and 30 interviewee respectively on randomly sampled population of about 198. However, one hundred and fifteen (115) of the tool representing about 82.5% retrieval ratio was recorded. The survey revealed 50.69% and 49.31% intensely agreed and agreed respectively that Aquila was designed using the Work Breakdown Structures (WBS) of Operations Department (OPs). Another 53.92% and about 80% strongly agreed respectively that the technology captured every data on all transactions which has transformed the payment process significantly. An overwhelming 70.05% and 36.61% also subscribed that data management has effectively helped the OPs Department in planning. The result of gap analysis indicated Aquila has met stakeholders expectations gaining 80% support and recording about 70% success. Another 52% observed that the pre-Aquila era was manned by fraudulent transactions. The paper submitted the need for immediate commissioning of the proposed project Aquila II.

Keywords: Innovation, Data, Aquila, Technology, Petroleum, Oil and Gas, Marketers, PEF (M)B

1. Introduction

Oil exploration in Nigeria began in 1908 when a German Company, Nigeria Bitumen Corporation was licensed to explore at Araromi area of Western Nigeria. Operations was truncated few years after the discovery due to First World War in 1914. Apparently, due to the interruption in operations, Anne, et al., (2015) claimed the discovery was now made in 1956 at Oloibiri, Niger Delta. Over time, however, the sector has emerged as extremely competitive. About 15 years after the expansion of exploration privileges, PEF(M)B (2013) observed most petrol service stations nationwide were characterized by long queues due to frequent severe shortages of petroleum products. In response to the yawning of Nigerians (The Equaliser, 2011b), the government set up an inter-ministerial committee to examine the situation and make appropriate recommendations. The impregnated recommendations gave birth to Uniform Pricing System (UPS) which steered the establishment of PEF(M)B through Decree No. 9 of 1975 (as amended by Decree No. 32 of 1989). The Board was charged with the responsibility of reimbursing

petroleum-marketing companies for losses suffered as a result of sale of petroleum products at uniform prices throughout Nigeria [PEF(M)B, 2013]. Towards the actualization of aforementioned mandates, as at December, 2011, PEF(M)B recognized the need to shift the paradigm of claims payment from manual data collection to generating a pool of data on petroleum marketers using the Aquila Technology Platform for the purpose of innovating the payment process. This made the Board unveil the Aquila Technology in December, 2011 (The Equaliser, 2011b).

Aquila, a Latin name for Eagle was chosen in view of the bird's strength, speed and accuracy. To successfully couple a set of the hardware, equipment required include; Radio Frequency Identification (RFID) Tags for Trucks, RFID Reader and CN3 Mobile Device. Aquila is not a tracking device to monitor movement of trucks or goods in transit, but it had scrutinized payments that were due for about 30,000 different petroleum marketers within its database as part of bridging claims for distribution of petroleum products across depots in Nigeria within the period from January to December 2013 (PEFMB, 2013). Baaziz and Quoniam (2014) recommended the adoption of big data technologies for optimization of operations in upstream sector of the petroleum industry. Hence, thousands of data-collecting sensors in facilities were used to provide continuous and real-time monitoring of assets for taking evidence-based and informed decisions. This paper argued the cost implications of taking wrong decisions far outweigh that of adopting a data extraction technology such as Aquila. Because a system is never perfect, it is expected that technologies be open to positive contributions from beneficiaries of the services rendered by the Aquila technology. Issues raised by petroleum tanker drivers, the need to stop illegal payments to unscrupulous petroleum marketers as bridging claims and relevant other issues bordering on the functionality of the Aquila further makes studies of this nature ever relevant. It is expected that the output of the research will serve as a blue print for policy makers and stakeholders to chart the right course of action for the development of Aquila II.

The object of this paper was fastened on appraising the impact of data collection and analysis through the Aquila technology platform on the innovation capability in the Operations of PEF(M)B, particularly in the last 5 years of the use of Aquila Technology. The paper sought answers to issues such as (i) the relevance of data gathered through the Aquila platform on the tasks of the Operations Department in PEF(M)B; (ii) the role of Aquila technology data in innovating the settlement of marketers' claims and; (iii) the observable lapses and imperfections of Aquila during implementation, which would assist in proposing an improvement to be labelled Aquila II Technology. This survey embraced data from both primary and secondary sources such as published and unpublished articles from the internet and intranet, as proposed by the Narrative-Textual Case Study (NTCS) method of research, see Abouzeedan and Leijon (2007). The justification for using of NTCS is due to the uniqueness of the Aquila technology and the insubstantial nature of academic publications on the study at hand. Section 2 of this paper dwelled on relevant literature materials reviewed to gain insight of the topic at hand. Method of data collection and analysis adopted were presented in Section 3. In section 4, the results obtained were presented and discussed, as the paper wrapped up with submissions for making informed decisions as regards the role of data on innovation.

2. Conceptual Outline

2.1 Innovating with Data and e-payment in Upstream Petroleum Industry

The upstream sector of the petroleum industry is no stranger to using data for innovations as observed by Baaziz and Quoniam (2014). Organizations use thousands of sensors installed in surface wells and facilities to provide continuous data-collection, real-time monitoring of assets and environmental conditions. Settlement of customers' claims were also facilitated by data mining technology (PEFMB, 2013). Other sources of big data include sensors, spatial and GPS coordinates, weather services, seismic data, and various measuring devices. "Structured" data is handled with specific applications used to manage surveying, processing and imaging, exploration planning, reservoir modeling, production, and other upstream activities. Most of the data in use before Aquila were "unstructured" or "semi-structured" such as emails, word processing documents, spreadsheets, images, voice recordings, multimedia, and data market feeds, which made it difficult or costly to either store in traditional data warehouses or routinely query and analyze. Therefore, appropriate tools for Big Data was recommended for use (Baaziz and Quoniam, 2014) such as the Aquila Platform which was applied in loading and offloading of petroleum products managed (PEFMB, 2013).

The regime of e-payment system in Nigeria oil and gas sector could be traced to the early 2010 when Easy Fuel Limited, a Nigerian company engaged in the provision of e-payment solutions. The firm introduced a range of products. One of such integrated e-payment solutions had the potential of revolutionizing payment system in the upstream sector. The intent of the technology was to eliminate the use of cash in buying petroleum products at filling stations across the country over time (Tayo, 2011). The author also proposed that with these solutions in place, a customer only needs a bank card or any of Easy Fuel's proprietary e-payment products which may be installed in the vehicle or issued as contactless Free on Boards (FOBs) to purchase fuel. With the speed pass, customers can set predetermined limits on how much and how often a vehicle can refuel. The product offered the highest level of control and accountability for fleet managers. The Sprint is a smart RFID device, which when installed in a vehicle and worked by identifying the vehicle to the Easy Fuel system (Tayo, 2011).

However, Gbodume (2014) observed that for an e-payment system to be efficient there must be a way for merchants to verify the validity of the purchase and payment must be easily convertible to cash. He also suggested that since most merchants in Nigeria were in business on subsistence basis, cash will be better appreciated. This is where an e-payment or e-transactions solutions came handy. These payment solutions target most of the concerns of merchants and more. Despite its advantages, e-payment solutions in the upstream sector have not enjoyed much acceptance as anticipated in Nigeria. To this end, the Equalizer (2012) queued behind Tayo (2011) on stakeholders expectations from Aquila.

2.2 Technology and Data, the Aquila Technology Case

As Shehu (2014) observed, the number of marketers in the downstream sector had availed data and communication technologists significant investment opportunities waiting to be explored by data experts. The author observed there were over 6,000 independent petroleum products marketers and 6 major marketers distributing and marketing petroleum products across the country. His argument was hinged on the provision of ICT and data management related services to oil companies and marketers alike.

Since PEF(M)B recognized that for any organization to metamorphose from being reactive to being proactive, it must adopt the application of big data technology for informed decision to be made. PEF(M)B has demonstrated this as the first organisation to successfully launch and implement an electronic loading and delivery system for petroleum products in Sub-Sahara Africa as observed by Abubakar, et al., (2015). The platform has automated data capture mechanism, which enabled faster processing and payment of claims with an online, real-time information management system facilitated by the Radio Frequency Identification (RFID). In addition, the technology provides data for strategic business decision planning and making. Prompt settlement of marketers' claims accrued during pre-electronic management and payment system era was made with ease. This technology also promoted a healthy working environment by reducing the traditional paperwork, thereby reducing cost in printing and by extension transportation [PEF(M)B, 2013].

Project Aquila is a software written to work on RFID technology. The launching of the business solution also made Nigeria to join other world leaders like Canada, India and South Africa where RFID technology had been in use successfully. PEF(M)B (2013) also observed that this technology is a first of its kind in Sub-Sahara Africa. Although the process work flows was designed by an external consultant, the software codes for Aquila were written by a staff of PEF(M)B, thereby saving government an estimated two million dollars ($2million) [PEF(M)B, 2013] and promoting local content as enshrined in the Local Content policy of the government.

The adoption of data technologies in many countries by different sectors of the economy has made direct positive impact on the organizations' efficiency. Baaziz and Quoniam (2014) sighted instances where Chevron used proof-of- concept for seismic data processing. Shell also uses piloting Hadoop in Amazon Virtual Private Cloud for seismic data processing too. To this end, much of the innovation that is key to the digitization of big data is happening at oil service contracting companies. Pool of data has made tremendous improvements possible on decision making processes in PEF(M)B (The Equaliser, 2011a). David (2004) had proposed the application of computers to process, analyze and display information about many processes involving moving "trucks." Entering the information about the status of moving things requires repeated data entry which is cumbersome, expensive and error-prone. Consequently, many automated systems have being developed to perform this data entry task. Together these systems are referred to as automatic iden-

tification ("Auto ID") systems. Other important areas of application of Auto ID include product identification in consumer goods industries, swipe card access systems and GPS otherwise referred to as on board computer systems used in trucking. The availability of such data entry systems has provided the impetus for development of more sophisticated decision support and control systems (Duncan and Yossi, 2004).

In line with the developmental process of using technological approach for rapidly and accurately delivering product data into supply chain operations, core is the decoupling of the physical item from the information representing it (as is the case with bar codes). In particular, digital identity is the only piece of product data that must be directly located on the product itself. All the other data can be stored elsewhere, with the identity providing a unique code to access it.

Project Aquila's benefits to PEF(M)B in particular and to the oil and gas sector in general went beyond the capacity to generate data pool on national consumption pattern of petroleum products. The benefit also included the planning and determining of the volume of petroleum products bridged across the country. Again, the manual method of determining accruals to the marketers and transporters before the adoption of this technology had become a thing of the past (Equalizer, 2013). This technology uses the RFID in identification of tags placed on trucks which serves as a chip that monitors the loading of the truck. The RFID reader and the CN3 mobile device combine perfectly to extract required data of trucks. The same Aquila platform is used at the discharging point of trucks to capture the data of the discharging trucks which is used in processing the payment of marketers' claims.

While introducing the business solution to stakeholders, the Executive Secretary (ES) of PEF(M)B emphasized the transparent and sustainable nature of "Project Aquila". She emphasized it will eliminate the guesswork and perceptions on which the then consumption pattern of petroleum products were based. The information assessed on the database would show products' loading and off-loading at depots as well as areas of the country to which they were assigned.

In addition, the success of the project relies on the principles of winning key compliant. To achieve this, there is need on the part of the staff to possess the ability to initiate a winning mentality in spite the challenges in the industry. Pertinent also are the roles and responsibilities of staff towards the creation of a winning team and highly innovative capability which are paramount to the success of the goals of the project as outlined by Abubakar, et al., (2015). The Equalizer (2012b) also proposed staffs' initiative and innovative capability, self-confidence, enthusiasm, data analytical skills, flexibility and technical know-how must be brought to bear.

2.3 Auto ID System Architecture and Radio Frequency Identification (RFID) Technology

Automated Identification (Auto ID) involves the automated extraction of the identity of an object. The Auto ID system applied in the project Aquila as described in this paper draws heavily on past and current developments in the area of Radio Frequency Identification (RFID) (Abubakar, et al., 2015). The RFID technology

provides a simple means of automatically obtaining the unique identity of an item. In the instance of Aquila, the item in question is a petroleum product loading truck. This operation is done at increasingly low cost. The systems can then be coupled with networked databases which enable access to additional product data. Much of this functionality can readily be provided by bar code systems (See Duncan and Yossi, 2004 for details on the basic principles of operations of a typical RFID). An RFID-based Auto ID system generally comprises the following elements:

i. A unique identification number which is assigned to a particular item
ii. An identity tag which is attached to the item with a chip capable of storing - *at a minimum* - a unique identification number. The tag is capable of communicating this number electronically.
iii. Networked RFID readers and data processing systems which are capable of collecting signals from multiple tags at high speed (100s per second) and of preprocessing this data in order to eliminate duplications and mis-reads.
iv. One or more networked data bases that store the product information (Duncan and Yossi, 2004).

Similarly, (Duncan and Yossi, 2004) juxtaposed RFID Based Auto ID Systems and Bar Code Based Systems and observed that the RFID is a logical extension to to-day's barcode-based on systems that have been so successfully applied throughout supply chains. Barcodes identify products at transition points such as shipping, receiving, and checkout. General features of these technologies are that both are inexpensive, ubiquitous and, in principle, very accurate. There are two advantages of the RFID technology over conventional bar code systems:

i. Bar codes have to be scanned deliberately by a person in a process that is difficult to
automate. On the other hand, RFID tags can be readily scanned automatically without human involvement.
ii. Bar codes require line-of-sight to read, while RFID tags can be read in any orientations as long as they are within the reader's range (Duncan and Yossi, 2004).

3. Research Methodology

3.1 Research Design

This paper draws facts and relevant literature from historical data. It then adopted the Narrative-Textual Case Study (NTCS) galvanized with the survey research using questionnaire to pull information from respondents on the significance of data gathered through the Aquila technology platform on the innovations in the operations of the Board. NTCS is a social science research method that employs intensively, the information, data and academic materials made available and easily accessible by information and communication technology facilities such as intranet, internet, World Wide Web, online databases, e-libraries et cetera

(Abouzeedan and Leijon, 2007). The method combines the use of quantitative and qualitative, observation, text content analysis and available official statistics in different proportions for problem-solving or problem-identification as designed by the research.

3.2 Research Instrument

A self-designed questionnaire was used for this survey. The justification for the choice of questionnaire was hinged on the uniqueness of the tool in addressing research questions as designed by the researchers. The instrument was divided into three (3) sections. Section A contains general information on respondent such as Organization, Schedule of duty and Educational qualification, etc. Section B has twenty (20) items which sought the views of respondents on the significance of data collected through the Aquila technology platform for innovations of the tasks performed by the operations department of the understudied organization. This section used a five (5) point liken rating options comprising Strongly Agree (SA), Agree (A), Uncertain (U) Disagree (D), and Strongly Disagree (SD). In section C, attempt was made to conduct a gap analysis of the pre and post implementation of Aquila technology.

3.3 Method of Sampling and Data Collection

The target population of this essay cut across the transporters of petroleum products under the auspices of the National Association of Road Transport Owners (NARTO), Petroleum Tanker Drivers Union and National Union of Road Transport Workers (NURTW), staff of the Petroleum Equalization Fund Management Board (PEFMB) and the Independent Petroleum Marketers Association of Nigeria (IPMAN). Random sampling technique was adopted in selecting respondents from a population of about one hundred and ninety-eight (198) for the survey. A total of one hundred and forty (140) questionnaires; representing about 70% of the population, were administered on the target population. Another thirty (30) respondents representing about 12.63% of the population were interviewed. However, one hundred and fifteen (115) of the questionnaire representing about 82.5% retrieval ratio was recorded.

The inclusion criteria followed the focus and technicalities of the research questions. Questions were targeted at the Operations and the Information and Communication Technology (ICT) Departments of PEF(M)B. Other stakeholders like IPMAN, NARTO, NURTW and Petroleum Tanker Drivers Union were included because they were beneficiaries of the output of Aquila technology. Data analysis was performed using Microsoft Excel.

The tool was given to seven (7) Subject Matter Experts (SME) drawn from PEF(M)B and academics to answer the question and validate the content. The result of validation revealed that five (5) of the experts in the panel answered “yes, relevant” with minor recommendations. Their views were respected and implemented before the production of the final research tool.

4. Results

4.1 Result on the Role of Aquila Data in Transforming the Settlement of Marketer's Claim

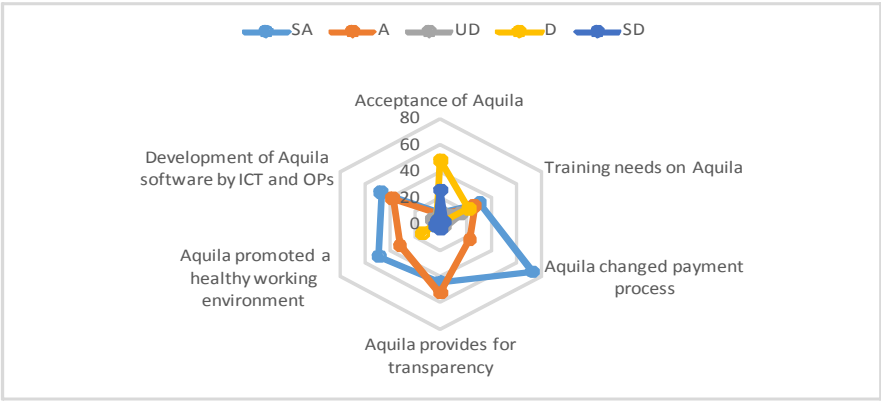


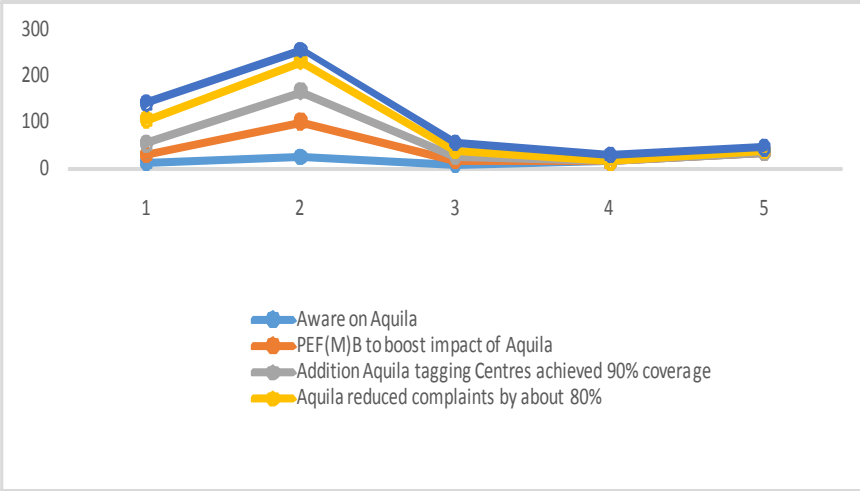
Fig. 1: Role of Aquila Data in the Innovation of the Settlement of Marketer's Claim

4.2 Result on the Relevance of Data Gathered through Aquila on Tasks in PEF (M) B

S/N	STATEMENT	SA	A	U	D	SD
		%F	%F	%F	%F	%F
1	The data captured by Aquila has helped OPS in planning	36.61	35.27	3.13	6.25	18.75
2	Aquila was developed using the WBS of the OPS	50.69	49.31	0.00	0.00	0.00
3	Data gathered on Aquila has positive impact on job efficacy	59.69	26.02	7.14	7.14	0.00
4	Data on every transaction is captured on Aquila database	53.92	26.73	3.23	12.90	3.23
5	Aquila bid data capturing capacity	57.64	21.67	17.24	0.00	3.45
6	Aquila made effective data management possible	70.05	29.95	0.00	0.00	0.00
7	Most staff of OPS lacked skills to explore Aquila	9.05	7.62	13.33	56.67	13.33

Table 1: Relevance of Data Gathered through Aquila on Tasks in PEF (M) B

Fig. 2: Gap Analysis of Aquila Business Solution



5. Discussion

5.1 Impacts of Aquila Technology on Innovating the Operations of PEF(M)B

Aquila technology has obviously made some positive impact on the operations of the Operations Department and by extension the job of PEF(M)B. This is evidenced in Figure 1, as an overwhelming 70.05% and about 29.95% of respondents settled for an opinion that the business solution has made data management more effective hence it has changed the payment process in PEF(M)B. To this end, it is suggested that training and re-training of officers be done regularly in the areas of using the Aquila technology to proffer swifter responses to issues of marketers' claims payment process. It was observed that there was weak synergy between the developers of the Aquila data extraction platform and the Operations Department of PEF(M)B. Aligning with above observation are about 47.78% and 38.42% like-minded to this proposition. This may be hinged of the local content development policy of the government, hence bulk of the design of the software was developed by a staff of PEF(M)B. This argument was supported by Abubakar, et al, (2015) and is in tandem with the report of the Equalizer (2012).

The implication of this outcome also is that with adequate training of officers in the Operations Department of the understudied Board, schedules of officers can be innovatively discharged. It is also crystal clear from Figure 1 that the solution has improved the transparency of the payment system. It is believed that this is a pointer to sustainability of the project, as it gains better acceptability from relevant stakeholders. Apparently, Aquila has eliminated sharp practices by reducing number of human interference in operations of PEF(M)B, hence it has succeeded in managing the payment process better, especially when juxtaposed with the hitherto manual process of settling claims by the Board.

On the volume of data retrievable by the Aquila, about 48% of respondents agreed that all transactions passing through designated tagging depots across Nigeria were fully captured. Such a huge data is extracted at regular intervals for informed and evidence based decisions to be made. However, Table 1 revealed that about 17.24% finds it difficult to decide if Aquila has the capacity to capture voluminous data as may be required in the operations of PEF(M)B. The consequence of this outcome on the impact made by the business solution is that the acclaimed success must be consolidated to ensure the required improvement on the ease of performing tasks in the Operations Department is strengthened.

Apparently, knowledge management and human capacity development promote a healthy working environment. The duo were promoted by the introduction of Aquila in the operations of PEF(M)B. This assertion is sequel to the sustainable prospect when about 36.61% of respondents strongly agreed and another 35.27% agreed respectively that planning of logistics has been better since data is fast assessable by the Operations Department when needed to process payment of marketers' claims. The interview revealed there is need for improvement on the level of privileges administered to officers of the Operations Department by the ICT Department to allow the Operations Department access information as and when needed.

Another challenge of the technology is the level of privileges availed to ICT personnel, which stakeholders dread may be compromised if not checkmated. We argued that if secured privileges are extended to the Operations Department, then the bar of security is raised and porous pots closed against unnecessary intruders from the ICT Department, the technology will be more secured and this will improve on the transparency and the confidence accorded the operations of PEF(M)B.

Much as the technology has enjoyed support from stakeholders, the issue of product diversion remains a challenge. In Figure 2, about 37% of respondents strongly agree and another 25.12% agree respectively that product diversion has not being addressed in Aquila. We also gathered from the interview conducted that since the technology can only access information at the loading and discharge depots, the monitoring of the content of the truck, location or visibility of the product and truck; challenges earlier identified by Supply Chain Management (SCM) following Abubakar, et al, (2015) are germane to the success of the technology. On this basis, the gap to be covered by project Aquila II widens with time, since Duncan and Yossi (2004) orated the possibility of deploying the technology to tracking of truck and products. This survey further strengthened the claim of Olamade, et al, (2014) that visibility is the topmost among other challenges facing supply chain managers.

5.2 Pre and Post Implementation of Aquila Technology

We defined gap analysis in this context to involve the comparison of actual performance with potential, designed or desired performance. It provides a footing for measuring investment of resources against expected outcome (e.g. to turn the claims payment process from paper-based to paperless with the use of Aquila business solution).

The result revealed over 50.69% strongly agrees that the business solution has reduced complaints on the operations of the Board by about 80%. This is a pointer to the level of acceptability and popularity the solution has enjoyed since inception. The result also supported the findings of Abubakar, et al., (2015). From Figure 2 90% agree that the addition of Aquila tagging Centres in Kaduna, Lagos, Ibadan and Enugu has made the solution achieve about 90% coverage of marketers. Deductible here is the possibility of increase in the volume of data generated through the Aquila Platform over time. While it is argued that this will give room for qualitative decision making, handling such data poses even more challenges as proposed in Baaziz and Quoniam (2014).

While conducting a gap analysis, it was revealed that the pre implementation anticipated performance of Aquila technology was realized. This position was also supported by about 78% of respondents who claimed the solution has achieved about 70% success. About 19% ranked low that the system has achieved less than 60% of the expected outcome. Although stakeholders are yawning for improvement as the issue of product diversion remains an impediment to Just-In-Time delivery of petroleum products across the country.

It was also discovered that the pre Aquila era, for instance, between year 2000 to 2006 was characterized by cumbersome, falsified and shady practices, hence the process of preparation of marketer's claims was potholed and inaccurate. This claim was supported by 55% ranking above average on the one hand. On the other hand, 24% ranked the process and operations of PEF(M)B was faring well even amidst irregularities and difficulties officers were subjected to during assessment and preparation of marketers' claims. This paper claimed that Aquila has changed the face of the Board internally and externally. Deducible from this survey also is that the wealth of experience of the respondents that has been brought to bear, since about 33% have about 12-17 years experience. To consolidate this argument, on the gap analysis between expected outcome and obtained deliverable, about 78% of respondents agree that the business solution has played a leading role during the periods under review. The benefit of good teamwork was further established as has being enjoyed by the Aquila implementation team. We therefore argued the sustainability of every system is hinged on the quantum of human capital development built around it.

6. Conclusion

Aquila technology is a data extraction platform designed exclusively for the purpose of pooling data at both loading and discharging depots in Nigeria for the purpose of facilitating the process of making payments to marketers in compliance with the PEF(M)B mandate of equalizing the price of white petroleum products across board. The intent of this paper was to appraise the impact of the data pool through the Aquila technology platform on innovation capability of the operations in the understudied Board. Responses were fetched on the relevance of data gathered through the Aquila platform on swiftness of tasks delivery in the operations department of PEF(M)B. The survey which adopted a blend of the random sampling techniques and the NTCS gathered that the place of data in transforming the

process of settling of marketers' claims can never be over emphasized, since such data empower the officers in the Operations Department to facilitate payments just-in-time. It was also observed during the gap analysis that the development of Aquila technology has created a clean and transparent system to the operations of the Board. The system may not be perfect, but it has largely exceeded stakeholders' expectations. To gain stakeholders support, the need for teamwork and robust human capital development was underscored as enjoined by the local content policy, hence the contributions enjoyed by the business solution from staff of the Operations and the ICT Departments can be sustainable.

The paper put forward the need for a better cyber security policy in order to regulate levels of privileges given to the ICT personnel. The data capturing at both loading and discharging points should also be galvanized with product movement chips or trackers to eliminate the challenges of visibility and diversion of products. In addition, the paper proposes immediate commencement of Aquila II as a solution to the aforementioned challenges which should be anchored on a strategic implementation legal framework for the sustainability of success recorded. An innovative tagging should be deployed to avoid marketers abusing the process, penalties for non-adherence to the modus operandi of the Aquila may be an option to deter saboteurs. The technology should be made more robust and interactive enough to allow easy transfer of tangible facts and data capturing. The areas of how to curb product diversion using the Aquila technology and integrating the technology to an e-payment system are open to further studies.

Correspondence
Abubakar Kazeem
National Centre for Technology Management
Federal Ministry of Science and Technology
North Central Office, Abuj, Nigeria
Email: kz4tawa@gmail.com
Tel.: +2348028751764

References

Abubakar, K., Jude I. J., Zainab A. G (2015) Impact of Aquila Technology on the Supply Chain of Nigeria's Oil and Gas Industry. International Journal of Management and Commerce Innovations. Vol. 3, Issue 1, pp.597-606. Available at: www.researchpublish.com.

Abouzeedan, A. and Leijon, S. (2007) Critical review of the usage of narrative-textual case studies in social sciences and the connect to traditional research methods. Paper presented at the 10th Uddevalla Symposium, 14 -16 June, Uddevalla, Sweden. Research Report, University West.

Baaziz, A. and Quoniam, L. (2014) How to use Big Data technologies to optimize operations in Upstream Petroleum Industry. 21st World Petroleum Congress, Moscow, Russia. 15-19 June, 2014.

Duncan, M. and Yossi, S.(2004) The Impact of Automatic Identification on Supply Chain Operations. Available at: http://web.mit.edu/sheffi/www/documents/genMedia.sheffi-McFarlane.pdf. Assessed on 06/10/2016.

David, O. K.(2004) The Impact of Information Technology on the Nigerian Economy: A Study of Manufacturing and Services Sectors in the South Western and South Eastern Zones of Nigeria. ATPS Working Paper Series No. 39, Nairobi, Kenya. 2004.

Gbodume, A. (2014) Impact of E-Payment in Oil and Gas Industry. ED, Finance and Admin, MRS Oil Nigeria Plc. Available at http://cardatmandmobilexpo.com/wp-content/uploads/2014/06/IMPACT-OF-E-PAYMENT-IN-OIL-AND-GAS-INDUSTRY.pdf. Assessed on 06/10/2016.

Olamade, O.O, Abubakar, A. and Yusuff, S.A. (2014) Impact of Total Innovation Management on Supply Chain in Nigeria's Automobile Industry. Journal of Entrepreneurship and Business Innovation. ISSN 2332-8851 2014, Vol. 1, No. 1.www.macrothink.org/journal/index.php/jebi 3. Assessed on 06/10/2016.

Shehu S. A. (2014) Information Technology in the Nigeria Oil and Gas Industry and Nigerian Content Developmen. Presentation at the Stakeholders meeting on IT Local Content in the Oil and Gas sector held on 5th February, 2014 at NITDA, Abuja. Available on line at http://nigeriacomputers.com/tech-news/information-technology-in-the-nigeria-oil-and-gas-industry-and-nigerian-content-development. Assessed on 06/10/2016.

Tayo, O.(2011) EasyFuel out with e-payment Solutions for Oil and Gas Industry Posted in Technology. Available at http://www.nigeria70.com/nigerian_news_paper/easyfuel_out_with_e_payment_solutions_ for_oil_ gas_i/353721. Assessed on 06/10/2016.

The Equaliser (2011a) Capacity Building for the Project Aquila. Available at http://www.pefmb.gov.ng/?page_id=17. Assessed on 02/10/2016.

The Equaliser (2011b) PEF(M)B takes project Aquila to Majors Marketers. Available at http://www.pefmb.gov.ng/?page_id=17 . Assessed on 02/10/2016.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.36-51

Big Data Mining and Analytics for National Security in Nigeria

UJAH BRIDGET CHINALU
National Defence College Abuja, Nigeria

ADEJORO CONELIUS ONIMISI
University of Nigeria, Usukka, Nigeria

ABSTRACT In the recent times in Nigeria, the talk of the day has been from one national security issue to the other. The attacks by the Boko haram terrorist group in the northern region, herder-farmer clashes, kidnapping, oil pipeline vandalism and so on. The successes recorded by these criminal elements in Nigeria have been attributed to lack of actionable intelligence that would enable early detection, preventive and prediction of such terror attacks and other forms of security threats to the nation. In this study, we have carefully examined the measures adopted by the Nigerian security agencies to ensure internal security, and have found them inadequate as they are mostly based on the traditional approach of sense and response to security threats. However, owing to the technological advancement globally, technology can be put to good use in combating terror in Nigeria. Hence, to successfully defeat insecurity in Nigeria, there is a need for actionable intelligence gathering for effective repositioning of the country's security system. This study seeks to highlight how big data analytics can be leveraged to generate investigative lead and electronically gather intelligence for combating terrorism and other forms of security threats in Nigeria. The study recommends Big Data collaborative frame work for proper intelligence gathering, data analysis and information sharing amongst the security operatives for the purpose of efficient situational awareness and preparedness towards ensuring national security in Nigeria. This will enhance transcending from reactive approach to insecurity to evidence-based proactive approach aimed at nipping the act of insecurity on the bud.

Keywords: National Security, Actionable Intelligence, Analytics, Big Data, Intelligence Gathering

1. Introduction

According to Olajide (2015), a secured and safe environment is the desire of any nation that strives to harness its full potentials towards attaining economic development and improving the lives of its citizens. To this end, nations adopt several measures such as use of technology to gather intelligence on potential threats towards safeguarding the lives and properties of its citizenry. The application of technology in intelligence gathering is vast across the world. Osiah (2015) reveals that in most countries of the world, effective application of technology in intelligence has enhanced national security. For example, 75 per cent of former Union of

Soviet Socialist Republics (USSR) satellites were dedicated to missions including intelligence collection, geodesy, communication relay, weather and radar calibration. An important part of technology application in intelligence gathering is to offer response that is proactive to enemy intent so as to safeguard the sovereignty of state, prevent individual and groups from critical pervasive threats that undermine national security (Olajide, pg 20, 2015).

In the USA, after the terrorist attacks of 11 September 2001, the USA employed extensively the use of satellite and drone technology for intelligence acquisition against al-Qaeda leaders in Afghanistan, Pakistan and Yemen among others. These efforts yielded fruit with the killing of Osama bin Laden and 5 others in Abbottabad, Pakistan on the 2 May 2011. To hunt bin Laden, the USA Central Intelligence Agency (CIA) in addition to the satellites, flew an advanced Stealth Drone, the RQ-150, over Pakistan to eavesdrop on electronic transmissions. The CIA also was able to penetrate guarded communications among al-Qaeda operatives by tracking calls from their mobile phones. Also, pinpointing the geographical location of one of the phones to the compound in Abbottabad, Pakistan, where other evidence suggested Bin Laden was hiding. These feats achieved with the application of technology in intelligence disrupt the operation of al-Qaeda worldwide thereby enhancing the national security of the USA as recorded by Olajide (2015). Such a remarkable success is achievable in Nigeria with the help of an actionable intelligence tool that relies on technology to improve intelligence by identifying and predicting potential threats and crimes before they occur as well as preventing them from occurring, finding critical security information faster, enabling information sharing and collaboration between investigative organizations. Fortunately, the evolution of smart technologies and ICT tools, such as unmanned aerial vehicles, surveillance drones and satellite reconnaissance for border security could generate intelligent data, which when analyzed would produce actionable intelligence upon which security operatives could take action. The data gathered via this smart weaponry, and electronic intelligence demonstrates the potential of technology to tip the scales of terrorism and other forms of security threats in Nigeria. It could also mitigate, if not entirely negate the manpower deficit that has so far crippled border policing in Nigeria (Chris, 2015).

There is a need for a tool that could maximize the exponential data growth in Nigeria by combining vast stores of structured and unstructured historical security data from various sources, aggregate and analyze the aggregated data to create a complete, holistic view of an entity or situation. This will provide investigators with highly accurate, real-time identity, location and relationship insights and intelligence about targets, areas of interest and patterns of life.

In Nigeria, the fast advancement of Information Communication and Technology (ICT) suggests that the efficient use of ICT means to gather intelligence used to detect and defeat or avoid threats could guarantee national security. The NCC monthly Subscriber Statictics (August 2014) suggests that the deregulation of the mobile phone market in Nigeria over a decade ago has led to the introduction of Global System for Mobile communications (GSM). Nigeria has the largest mobile market in the African continent with over 90 percent of individuals and corporate organizations relying completely on the mobile industry for their day-to-day trans-

actions. This has impacted positively on the country's GDP too. The statistics from National Communication Commission (NCC), the Nigerian telecommunication regulator put the teledensity in Nigeria at 94.4 percent in August, 2014 and active lines at 133 million subscribers in a country of about 160 million populations according to NCC monthly subscriber statistics (August 2014). A direct result of this growth is the generation of quintillion bytes of user and network-related data in the country. These huge data sets contain the footprints of users, which include the criminals; it can therefore be put to good use in ensuring national security. Additionally, we now live in a global world where almost every day activities have been digitalized creating room for fast generation of large volume of data every second.
In the last 2 years, we've created more data than in the history of mankind. And the variety of data sources is evolving—transaction data, mobile phone data, social media chatter, telemetry, and on and on. There are needles in that haystack, and finding them depends on the ability of public safety organizations to apply analytics to find patterns that can lead to an interdiction before disaster strikes. It is on this premise that the idea of leveraging on this large volume of data as a means of proactively detecting and preventing threats evolved (Grant, p34, 2014).

According to KDnuggets (2016), human beings now create 2.5 quintillion bytes of data per day. The rate of data creation has increased so much that 90% of the data in the world today has been created in the last two years alone. This acceleration in the production of information has created a need for new technologies to analyze massive data sets for the purpose of better decision making in terms of guaranteeing national security.

Presently, there is an urgent need for collaborative research on Big Data field as underscored by the U.S. federal government following the recent release of $200 million funding initiative to support Big Data research in the U.S. (KDnuggets, 2016). In many places across the globe also, information technology has been adopted to combat the problem of terrorism, insecurity and uproar Big data has great potential to predict crime, crime hot spots and criminals. (Akinode et al, 2013). KDnuggets (2016) defines Big Data as a large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. Big Data is differentiated from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety). Analysing so many different criteria over the entire population is done with Big Data technologies and algorithms. Miller et al (2012) buttressed further that the term Big Data has a variety of definitions and has been used in a variety of contexts. It is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store and analyze and is used to enhance decision making, provide insight, discovery, support, and optimize processes.

Data volumes are growing exponentially and there are many reasons for this growth, including the creation of most data in digital form, proliferation of data sensors and new data sources such as high-resolution image and video
Mark (2015) estimates that there are approximately 100 million security cameras worldwide at the moment and this continue growing as the years go by. These

cameras are used to control important economic areas/buildings/highways/events and smart cameras can be used to notify organizations in real-time when a security breach is noticed. Mark (2015) reveals that IP cameras that are directly connected to the internet account for approximately 60% of all sales in 2016 and the percentage of HD security cameras will grow to 50% in the nearest future. All these high-definition IP cameras will generate massive amounts of data that can be automatically analysed with the right Big Data tools. This will enhance detection and defeat or prevention of threats or attacks. In many places across the globe, information technology has been adopted to combat the problem of terrorism, insecurity and uproar. Big data has great potential to predict crime, crime hot spots and criminals. Like everyone else, terrorists, leave digital traces with much of what they do, whether using e-mail, cell phones or credit cards. This data can be mined and used in fighting them. (Akinode et al 2012). Big Data is relevant to Nigeria due to the vast amount, variety, complexity and variability of data produced in the country. (Hopkins and Evelson 2011) In 2012, the World Economic Forum identified Big Data as a very powerful tool for public safety and national security. A Beckstrom, President and Chief Executive Officer of the Internet Corporation for Assigned Names and Numbers (ICANN), during the World Economic Forum in 2012. In the coming years we will see a massive increase in the use of Big Data tools by governments to increase national security (Rajkumar, 2016). The usage of Big Data tools to secure organisations, prevent crimes and ensure national security

2.0 Literature Review

2.1 National Security

The concept of national security started in the United States after World War II. Initially focusing on military might, it now encompasses a broad range of facets, all of which impinge on the non-military or economic security of the nation and the values espoused by the national society (Romm, pg 56, 1993). Accordingly, in order to possess national security, a nation needs to possess economic security, energy security, environmental security, etc. Security threats involve not only conventional foes such as other nation-states but also non-state actors such as violent non-state actors, narcotic cartels, multinational corporations and non-governmental organizations; some authorities include natural disasters and events causing severe environmental damage in this category. Fahey (2010) highlights the major national security missions to include but not limited to Conventional Military Defense (CMD), Counter Nuclear Proliferation (CNP), Counter Chemical/Biological Weapon of Mass Destruction (WMD), Counter Terrorism (CT), Cyber Security (CS), Counter Intelligence and Counter Narcotics.

2.2 Big Data in Public Safety and Crowd Control

Mark (2015) outlines the main areas that Big Data can affect and improve security in the coming years to include but not limited to public safety and crowed control. Verizon (2010) opines that national security can be realized through public safety

which is a sole responsibility of the government, to ensure that civilians are secure and especially during large events. Different tools can be used to achieve public safety to ensure effective crowd management in an event which involves large crowd. One of the best of these tools is Big Data solution. It could be used to control crowds and keep events safe by using it to monitor the movement of the crowd during an event, to prevent too many people at one place and therefore prevent disasters such as the Love Parade disaster in Germany in 2010 (Hastie et al, 2009)

2.3 Big Data Analytics for Organizational Security

Mark (2015) affirms that organizations are swimming in security data which organizations can maximize for their advantage. In a panel of discussion at the 2012 RSA Conference in San Francisco, Ramin Safai, chief information security officer at Jefferies & Co., said his investment bank with 5,000 employees captures 25GB of security-related data every day. Hidden in the 25GB, they usually find 50 matters to examine more closely, 2 of which end up demanding demand real attention. According to a whitepaper by EMC, 47% of the enterprises collect, process and analyse more than 6 TB of security data on a monthly basis. Big Data impacts organisational security on three different themes (Mark, 2015):

- Big data analytics can be used to ensure employees' security as well as in detecting fraud or criminal activities and monitor risks among the employees of an organisation. Especially within large corporations, it is difficult to monitor all employees' actions, but with the right Big Data tools, organisations can watch employees without infringing on their privacy. Big data analytic tools can analyse full text emails or scrape communication channels and look for anomalies or patterns that indicate fraudulent actions. Only when the tool indicates an issue needs real attention, managers could dive into it to take necessary action. After all, organisations do want to protect their (intelligent) property and prevent an aggrieved employee to make sensitive data public for example.
- Big data mining could be used to prevent fraudulent actions by customers. Criminals always try to cheat and to make money or receive services without paying for it. Examples include insurance fraud, tax fraud or unemployment benefit fraud. With Big Data, organizations can prevent, predict, identify, investigate, report and monitor attempts at insurance fraud regardless of line of business and it can be done automatically. Using massive amounts of historical data they can determine what is normal and what is not and match that with actions happening in real-time. In combination with pattern analytics it can help identify outliers that require (immediate) action. This fraud-prevention industry is huge. The Insurance Information Institute estimated that insurance fraud accounts for $30 billion in annual losses in the USA alone (Mark, 2015).
- Big data Analytics could prevent organisations from being hacked by criminals who are after sensitive data such as credit card information, bank ac-

counts, passwords or who want to steal digital money. There has been a lot of news recently about organisations such as Facebook, LinkedIn and Evernote that were hacked and where massive amounts of passwords were stolen. With the right Big Data tools, organisations can much better detect abnormalities on the network or find intruders that are not allowed. Organisations should create an intelligence-driven security model that incorporates a 360-degrees view of the organisation and all risks that the organisation faces. Together with the right Security Information and Event Management (SIEM) solutions to provide real-time analysis of security alerts generated by network hardware and applications. Social media analysis is a great tool to do this and it is taking a ever-increasing role in crowd control. Governments worldwide use different Twitter analysing tools to scan and analyse tweets for security threats and will take action accordingly. A good example of a tool developed by the Dutch government in conjunction with control the crowd during a large event in December 2012 in The Netherlands (Mark, 2015).

2.4 Big Data and Deep Learning

Big Data and Deep Learning are two major trends that will impact and influence the future of data science. The exponential growth and wide availability of digital data offer great potentials and also bring new challenges in various disciplines. Harnessing the power of Big Data is not an ordinary task. Also, deep learning is a fast-growing field and one of the most promising methods for data analytics. It has been successfully applied to a wide range of application domains such as speech recognition, computer vision, natural language processing, and analytics for large-scale business data. To take advantage of the unprecedented scale of big data, developments in deep learning that can scale up are urgently needed (Mark, 2015).

2.5 Big Data Real Time Analytics for Counter Terrorism

(Akinode et al, 2013) In this age of big data, as this data is generated by people in real time, it can be analyzed in real time by high performance computing networks, thus creating a potential for improved decision making and insight. The idea of analyzing the data while it is generated is to allow the data to speak for itself, thus bringing out not just the obvious correlations and connections, but the unexpected ones as well. In other words, it is a systematic way of identifying and gathering footprints or traces of activities of an object of interest from a huge mound of data. The tracking of terrorist group can be achieved through analyzing the data generated from their activities, which left traces via phone calls, e-mails, videos, images, click-streams, logs from various said networks and telecommunication lines and facilities. Akinode et al (2013) affirms that identifying and understanding the full portfolio of issues facing the nation with the view of enhancing national security is possible through Big Data analytics. Not all data must be of the highest quality. The quality of the data will depend on (Akinode et al, 2013):

- The purpose of its use,

- The magnitude of related outcomes and potential resource investment to achievehese outcomes,
- The time to effect or time to impact of an issue, and
- The level of reliance and criticality to national strategy and operation.

On the road map for getting started with big data for Nigerian security agencies, Nwaga et al (2015) maintains that the world currently generates a huge volume of data. The capacities to store broadcast and compute this information continues to grow exponentially, with one estimate suggesting that the installed capacity to store information would reach 2.5 Zettabytes in 2012. International Data Corporation research suggests that the world's digital information is doubling every two years and will increase by fifty times between 2011 and 2020, according to. Terrorist across different jurisdictions heavily utilize modern transportation and communication Systems for relocation, propaganda, recruitment and communication purposes (Nwaga et al, 2015). The basic premise is that terrorist networks can be evaluated using transaction-based models. This type of model does not rely solely on the content of the information gathered, but more on the significant links between data (people, places and objects) that appear to be suspicious (Akinode etal, 2013). How to trace the dynamic evolution, communication and movement of terrorist groups across different jurisdiction in Nigeria and how to analyze and predict terrorists activities, associations, and threats becomes an urgent and challenging issue . Many terror-related groups use the web as a convenient, anonymous communication infrastructure. This infrastructure enables an exchange of information and propagation of ideas to active and potential terrorists. The part of the web used for such illegitimate and malicious purposes is referred to as Dark web.

2.6 Big Data Analytics for Network Security

According to the author in Girau and Wang (2012), Big Data analysis is a suitable approach to network security, to protect the network against advanced persistent threat (APT). APT attacks target organizations in sectors with high-value information, such as national defense, manufacturing and the financial industry. A challenge in detecting APTs is the massive amount of data to sift through in search of anomalies. The data comes from an ever-increasing number of diverse information sources that have to be audited. This massive volume of data makes the detection task look like searching for a needle in a haystack. In looking at the experimentation of Big data analytics in security, Girau and Wang (2012) noted that the Worldwide Intelligence Network Environment (WINE) provides a platform for conducting data analysis at scale, using field data collected at Symantec (e.g., anti-virus telemetry and file downloads), and promotes rigorous experimental methods.

2.7 Divers Applications of Big Data Mining

According to Rajkumar (2016), Big Data makes use of data mining technology to make intelligent decision: Data Mining is primarily used today by companies with a strong consumer focus — retail, financial, communication, and marketing organi-

zations, to "drill down" into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments. Rajkumar (2016) outlines 14 other important areas where data mining is widely used:

***Future Healthcare**:* Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs.

***Education**:* There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments.

Customer Relationship Management (CRM): CRM is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies.

***Fraud Detection**:* Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information.

***Intrusion Detection**:* Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection.

***Lie Detection**:* Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also.

Customer Segmentation: Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. Market is always about retaining the customers.

***Financial Banking**:* Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume of data is too large or is generated too quickly to screen by experts. The managers may find this information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer.

Corporate Surveillance: Corporate surveillance is the monitoring of a person or group's behavior by a corporation. The data collected is most often used for mar keting purposes or sold to other corporations, but is also regularly shared with government agencies.

Research Analysis: History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases.

Criminal Investigation: Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques.

Bio Informatics: Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience.

The need to analyse and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools. Advanced analytics solution requires several interacting design steps as stated by Murrow (2013). These underlying steps are reflected in the notional information flow.

2.8 Big Data Potentials for Defence and Public Sector

According to Neil and Bill (2013) Big Data potentially underpins substantial improvements in efficiency in the public sector, an analysis borne out by that of the Policy Exchange think tank. McKinsey identifies five categories of Big Data levers for the public sector:

- Creating transparency to make data more accessible, allow agencies to share data and minimize the repeat entry of data.
- Enabling experimentation through which to discover requirements, expose variability and improve performance. Big Data analytics can reveal wide variations in performance within agencies that are not visible in the highly aggregated analyses carried out by conventional means. Big Data also offers the opportunity for predictive analysis. By examining the relationships embedded in large datasets it is possible to build a new generation of models describing likely future evolution. These can be combined with scenario planning to develop predictions for how systems will respond to policy choices or other decisions.
- Replacing or supporting human decision-making with automated algorithms. Such techniques can often reveal anomalies contained in large datasets when compared to historical data and are applicable to counter-

terrorism and other police, security and defence intelligence scenarios. Algorithms 'crawl' through data sources identifying inconsistencies, errors and fraud.

- Segmenting population groups to allow targeted and tailored action plans.
- Innovating new business models, products and services.

These categories have been developed around tax and labour agencies, but McKinsey asserts that 'their use can be just as relevant to other parts of the public sector'. McKinsey's models are founded on exploiting the mass of data existing in a citizen -facing department. Their application to the defence sector will not necessarily be straight forward in all cases. However, those regarding transparency, experimentation and decision support do potentially fit the security model.

2.7 Processing Procedure of Big Data Analytics

The notional information flow (NIF) is also called information supply chain (ISC). This is a five-step underlying flow, representing the data processing design to bring the analytics visualization and specific insights of the given data (program). These five steps are:

1. Understanding source and data applications: this represents the first stage of deciding what data needs to be acquired and where it is going to be acquired from.
2. Data preparation: this is a stage of data filtering, cleansing and validation.
3. Data transformation and metadata repository: this is a critical step to preparing data for analysis by aggregating different data types and applying a structural format. This is where relevance is given to different data sets, even if they are seemingly unrelated data sets.
4. Business intelligence and decision support: this is the actual analytics stage, where statistics, algorithms, simulations and fuzzy are employed.
5. Analysts and visualization: at the end of this process, there is usually a human, the analyst, who needs to make sense of insights surfaced by the analytic engine that has run against the entirety of available data. Hopkins and Evelson (2011)

3.0 Methodology

Primary and secondary sources were used to collect data for the study. Secondary data entailed materials collected from published and unpublished books, journals, newspapers, the media etc. Primary data was equally obtained via unstructured Interviews with the representatives of the key security agencies in Nigeria, such as the Nigerian Army; the Director of Military Intelligence, the Nigerian Air force; the Director of Air Intelligence, the Nigerian Navy; the Director of Naval Intelligence, the Nigerian Police; the Deputy Inspector General of Police Force Criminal Investigation Department, Nigeria Security and Civil Defence Corpse (NSCDC); the Commandant General, Office of National Security Adviser (ONSA), National

Intelligence Agency (NIA); the Director General, Defence Intelligence Agency (DIA); the Chief of Defence Intelligence, Ministry of Defence (MOD); the Minister of Defence. Some other top officials from National Agency for Space Research Development Agency (NASRDA); DG NARSDA, National Communication Commission (NCC), Deputy Chairman, and National Identity Management Commission were also interviewed. The reason for choosing such individuals to be interviewed for this study is that they are the key officials in position of policy, operations, technology, communication, intelligence and security development in Nigeria.

3.1 Limitations of the Study

The major limitation of the study is the difficulty encountered in accessing classified information. Another major challenge is that the people interviewed handled the subject matter with high secrecy, owing to the secrecy involved in intelligence data handling, the people interviewed were not willing to open up on some certain security issues as they maintained that security matter is a sensitive matter.

3.2 Discussions

The data collected reviewed their mode of intelligence gathering and analysis, perception and capacity in Big Data collaborative framework for national security in Nigeria.

3.2.1 The Existing System

The problems identified in the existing system that would be solved in the proposed system include:

1. Sense and response approach to security threats; the existing measures adopted by the Nigerian security operatives are mainly based on sense and response approach. This approach does not ensure public safety and national security because it is not preemptive. Many things go wrong before the intervention of the security agents. Often times, actions are being taken when the criminal elements must have struck or in the verge of implementing their plans. At this point, it becomes too late to stop them or to remedy the situation.
2. Lack of a robust intelligence framework for efficient intelligence gathering, analysis and sharing of information; the security agencies in Nigeria still depend mostly on manual and physical method of intelligence gathering and analysis for their operations. The war on terrorism cannot be won just with arms and ammunition; we must evolve a robust frame work for intelligence. Also, owing to the technological advancement and in particular, ICT, technology can be put to good use in developing a robust tool in intelligence gathering and analysis for effectively curbing terrorism and other national threats in Nigeria.

3. Poor Collaboration between the security agencies; although efforts have been made to enhance collaboration between the security agencies in Nigeria, these efforts would yield much more fruits with stronger collaboration. Intelligence gathering would not make any difference if there is no collaboration among all security agencies and even the citizen. It is therefore imperative for all security agencies in the country to intensify collaboration which will eventually bring to rest the issue of terrorist in the country. There is a need for collaborative intelligence gathering, analysis and information sharing between the security agencies to successfully defeat terror in Nigeria.
4. Absence of actionable and real time intelligence; There is a need for actionable and real time intelligence that would guild the security decision makers in making decisions for effective repositioning of the country's security system. Actionable intelligence generated from big data analytics could guild decision makers in making right decisions towards ensuring National Security in Nigeria.
5. Absence of efficient data analytic tool; Presently, most security agencies rely on manual data analysis in carrying out analysis on the aggregated security data collected to draw insight on the holistic view of the security situation or what to expect. The manual analysis is not robust and is based on trial and error method. The problem with this method is that it is error prone and not based on scientific approach. Also, it cannot be used for large volume of data like in the big data analytics. There is therefore a need for an effective and robust analytic tool that could be used to analyze large volume of security data.
6. Lack of technical knowledge in performing automated analytics on aggregated data to generate the right result upon which decision could be taken. The personnels require adequate training in the area of modern data analytics for efficient situational management.

3.2.2 The Proposed System

We propose a Big Data Active Security Collaborative Framework as described in this manner:

This model advocates for a collaborative information gathering, analyzing and sharing system among the security agencies in Nigeria through Big Data Center. The framework consists of the Nigerian Defense industry, The Paramilitary Agencies, Big Data Center, National Communications Commission, Nigerian Communication Satellite Limited (NIGCOMSAT) and Network Operators from Ministry of Information and Communication Technology. The system links all the security agencies to ensure effective, real time information gathering, analysis and sharing among the security agencies for underpinning security threats in Nigeria. The GPS system is greatly employed for surveillance and monitoring to generate real time

information. Results generated from the Big Data Center can be shared across the security networks to forestall any security threat in the country.
In operational terms, the Ministry of Defense would ensure that information generated on any act of insecurity is effectively disseminated across the security networks.

In this regards, the military and paramilitary would utilize the data generated to track the criminals on the move. The Nigerian Communication Commission (NCC) tracks, monitors and regulates the network providers who provide call data, records etc. for analysis by the Big Data Center. The NIGCOMSAT hosts the GPS which ensures surveillances and efficient transmission of real time information generated through actionable intelligence of Big Data Center; the Ministry of Defence (MOD) alongside the Defence Intelligence Agency (DIA), the Directorate of Military Intelligence (DMI) and the Nigerian Police Force Criminal Investigation and Intelligence Department (FCIID) coordinate the the Big Data real time active collaborative analytics that generates the actionable intelligence, while the ministry of Information and Communication Technology ensures its smooth running.

4.0 Conclusion

The Security Agencies in Nigeria are playing a significant role in fighting terrorism and its likes, and to secure the life and property of the citizenry. A lot of efforts have been made by the government in this regard. The security agencies are equally doing their part to ensure that the country is a safe place to be at all times. The Big Data mining and analytics technology will greatly enhance this effort if well leveraged upon. The goal of applying Big Data analytics for national security in Nigeria is to obtain actionable intelligence in real time through an active collaboration. Big Data analytics have significant promise to national security. Hence, Nigeria must therefore maximize the full advantage of its huge data generation to realize the true security potential of Big Data in national security.

4.1 Limitations of Big Data Active Security Collaboration framework

- The cost of implementation is very high; so much money is required to purchase and install the networking equipment that will link all the security agencies, as well as modern intelligence and security facilities such as surveillance cameras, unmanned Aerial vehicles, GPS systems, very large data storage systems to house the data banks etc. required for analytics.
- A very high computing network is required to efficiently operate the system
- Great expertise is required in data capturing, storage, and to perform analytics on aggregated data
- The process of obtaining data for analysis involves very serious issues of privacy and ethics

4.1 Recommendations

The consequences of ignoring Big Data and associated third-platform technologies (cloud computing, mobile devices and social media) in the defence and security sector could be profound, including the loss of life and operational failure. In addition, the growing legal obligation regarding human rights of personnel and standards of care when on operations could apply in the future as much to the information provided to commanders and servicemen as it increasingly does to physical equipment and training. Beyond the clear moral imperative, therefore, the reputational and financial impact in an increasingly litigious society should not be ignored. The authors therefore recommend that the security agencies should:

- Define a collaborative Big Data work package as part of technology innovation studies. This should consider a broad range of candidate technologies and techniques from the commercial sector that may have application to the areas of defence outlined above.
- Consult widely on the responses to likely legal and ethical challenges that such an approach might require, particularly from a department of state security service.
- Select two functional areas (one from the Military and one from Paramilitary, such as the Police) that might benefit from pilot programmes or concept demonstrators, acting both to support the security operatives as a learning organization and as proofs of concept for Big Data techniques.
- In the pilot areas, the security agencies should: Assess training and educational needs for the functional areas expected to use Big Data, covering senior management and subject-matter experts (data analysts).

Initial assessment of the moral and legal issues to be addressed in any Big Data policy-development activity Clarify the role of industry in support of developing the capability, including potentially providing skilled data analysts to the reserve force element.

Correspondence
Ujah Bridget Chinalu
Department of Science and Tchnology
National Defence College Abuja, Nigeria
Email: Bridgechi2000@yahoo.com

Adejoro Conelius Onimisi
University of Nigeria, Usukka, Nigeria
Email: Specialcialcornel@gmail.com

References

Akinode, J.I., Alawode A.J. and Ojuawo, O.O. (2013) "Improving national Security GPS Tracking System Technology" proceedings of the 1st International Technology, Education and Environment Conference, African Society for Scientific Research (ASSR) pp. 634-644.

Chris Ngwodo, Nigeria's Unending Counter-Terrorism War, Premium Times October 22, 2015.

Devet, C. and Goldberg, I. (2014), The best of Both Worlds: Combining Information Theoretic and Computational PIR for Communication Efficiency. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),.

Fahey, S. (2010), Big Data and Analytics for National Security. The John Hopkins University. Applied Physics laboratory. [Online] http://web.stanford.edu/group/mmds/slides2012/s- fahey.pdf. Retrieved 91-05-2015.

Giura, P & W. Wang. (2012) Using Large Scale Distributed Computing to Unveil Advanced Persistent Threats. New York, NY: AT&T Security Research Center.

Grant Woodward (2014). How Big Data can change the world of Public Safety. [Online] http://www.sas.com/en_ca/insights/articles/big-data/local/how-big-data-changes-public-safety.html. Retrieved 21-08-2016.

Hastie Trevor, Tibshirani Robert. Friedman Jerome (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 10-08-2016. [Online] from: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Hopkins, B. and Evelson, B. (2011). Promotional Webinar: Expand your Digital Horizon with Big Data. Forrester. [Online] http://solutions.forrester.com/Global/FileLib/webinars/ Big_Data_Webinar.pdf. Retrieved 10-10-2016.

KDnuggets (2016) Data Mining, Analytics, Big Data, Data Science and Machine learning. [Online] from http://www.kdnuggets.com/meetings/. Retrieved 5-05-2016.

Mark Van Rijmenam (2015), Improving Organizational and National Security with Big Data, [Online] https://datafloq.com/read/improving-organisational-national-security-big-dat/220. Retrieved on 20-07-2016.

Miller, S., Lucas, S., Irakliotis, L., Ruppa, M., Carlson, T. and Perlowitz, B. (2012). "Demystifying Big Data: A practical Guide to Transforming the Business of Government", Washington: Tech America Foundation.

Murrow, B. D. (2013). What Big Data means to you? IBM Corporation. [Online] http:// murrow.net/Portals/0/Publications/Brian%20Murrow%20-%20What% 20Big%20Data% 20Means%20To%20You.pdf. Retrieved on 15-7-2016.

NCC Monthly subscriber statistics Report, August 2014. http://www.ncc.gov.ng; accessed 23-05- 2016.

Neil Couch and Bill Robins, Big Data for Defence and Security. Occasional Paper, September 2013.

Nwanga, M. E., Onwuka, E. N., Aibinu, A. M. and Ubadike, O. C. (2015) "Impact of Big Data Analytics to Nigerian mobile Phone Industries". Paper accepted for publication in the proceeding of the 2015 Intl' conference on Industrial Engineering and Operations Management (IEOM).

Olajide K (2015), Intelligence and National Security (Research Project, National Defence College Abuja.

Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009). A Data Mining & Knowledge Discovery Process Model. In Data Mining and Knowledge Discovery in Real Life Applications. Book edited by: Julio Ponce and Adem Karahoca, ISBN 978-3-902613-53-0, pp. 438–453, February 2009, I-Tech, Vienna, Austria.

Osiah, MCM (2015), Information Management and National Security in Nigeria. (Research Project, National Defence College Abuja.

Rajkumar, P (2016). Applications of Big Data. [Online]

Romm, Joseph J. (1993). Defining national security: the nonmilitary aspects. Pew Project on America's Task in a Changed World (Pew Project Series). Council on Foreign Relations. p. 122. ISBN 978-0-87609-135-7.

Proceedings on Big Data Analytics
& Innovation (Peer-Reviewed),
Volume 1, 2016, pp.53-66

Data Mining Application in Crime Analysis and Classification

OBUANDIKE GEORGINA N.
Federal University Dutsinma, Katsina State, Nigeria

JOHN ALHASAN & M. B. ABDULLAHI
Federal University of Technology, Minna, Niger State, Nigeria

ABSTRACT The analysis of crime data helps to unravel hidden trends that will aid in better understanding of crime pattern and the nature of those who commits such crimes. It also enables appropriate strategies to be put in place to control such crimes. Literature revealed that data mining has been successfully applied in crime analysis and control. In this work the crime data used was collected from selected Nigerian prisons. Exploratory analysis was performed on the crime data for better insight into the dataset before application of data mining algorithms. The exploratory analyses revealed that majority of the offenders are within the ages of 18 to 34 years old. The mining algorithm was limited to two classification algorithms. C4.5 algorithm was used to classify the data into vulnerable and non vulnerable groups. To verify the reliability of the C4.5 algorithm, Naïve Bayes algorithm was also used to classify the dataset. The result showed that C4.5 classified the data better with higher accuracy of 97% against 93% from Naïve Bayes. The rule generated by the C4.5 classifier revealed that those without educational qualification and not gainfully employed are the vulnerable groups. The authors are of the opinion that education is the way out of crime in Nigeria since most of the offenders either have low educational qualification or none.

Keywords: Data Mining, Crime Analysis, Naïve Bayesian, Tree Classifier

1. Introduction

Crime is a societal ill and its cost is usually enormous hence, the need for analysis of crime data to learn the factors that enhances crime and the nature of offenders (Wilson, 1963; Brown, 2003). Many people desire to live and operate in a secure environment; they want to be sure that their lives and that of their loved ones are secured. The main duty of any Government is to secure the lives and properties of its citizens through relevant policies and strategies. Nigeria is currently having serious security challenges ranging from the Boko Haram attacks in the North-East to armed robbery and kidnapping in other parts of the country. Data mining has gained recognition in crime analysis (Jawei et al, 2012). It is a field that cut across many other fields. There are many definitions of data mining according related literatures (Julio and Adem, 2009). The emergence of computing and communication technology has produced a society that extremely depends on information

(Witten and Frank, 2000). Development in technology has helped in collection and storage of large amount of data in many organizations' database. These usually contain hidden information. Most of these organizations gather these data for operational purposes after which they are dumped in data repositories or even thrown away or deleted. This type of data when mined can help in discovering of relevant information which can help the organization to increase productivity and can serve also as essential information to the society at large. Data mining has the capability to unravel the information that is usually hidden in such databases. Naisbitt (1986) is of the opinion that we are being choked with data but lack relevant information because the data are not mined to get relevant information. Data mining tools and algorithms are used to find relevant trends and to make necessary predictions and associations in data. Data mining has been successfully applied in virtually all areas of human endeavour which include banking, marketing, manufacturing, telecommunication, e-commerce and education (ZaoHui and Jamie, 2005). Data mining is an intelligent and potent data extraction technique that uses different types of data extraction algorithms. Data analysts explore large data repositories by using these data mining algorithms (Chen et al, 2004; Fayyad and Uthurusamy, 2002)

The analysis of crime data will help in cost reduction and reduction of training time of officer involved in crime control. It will also help in distribution of scarce resources to the appropriate quarters (Megaputer, 2002). The rest of the sections discussed about data mining techniques, data mining process; classification of the crime data using two popular classifiers and discussion of results.

2. Classification Techniques for Crime Analysis

Classification

Classification is a technique used to predict an unknown class label using a function. Classification as a method comes in two steps, the number one step involves the construction of the classification model (model training) while the second step involve using the model to predict class labels. An instance R of an R m-dimensional attribute vector can be represented as $R = (r_1, r_2 \dots r_m)$ each i $R = (r_1, r_2 \dots r_m)$ nstance belongs to a class of determined attributes $T_1, T_2 \dots . T_m$. When $T_1, T_2 \dots . T_m$ an attribute class is discrete value or unordered, it is said to be a categorical or nominal attribute and it serves as the category or fields of the records. The records that are used for the construction of the classification model is usually trainers and is portioned out from the dataset being used for analysis. The training model can be represented as a function $Z = f(r)$ which $Z = f(r)$ represents the used fields Z of a g Z iven record R (Jawei R et al, 2012).

C4.5 Classifier

It is a statistical classifier that is used to create a decision tree. It is carved out from ID3 method to overcome its methodological challenges by pruning the decision tree after construction and handling discrete and continuous dataset. Let D be data D set $d_1, d_2, \dots d_n$ with m $d_1, d_2, \dots d_n$ dimens m ional attributes $t_1, t_2, \dots t_m$ and $k_1, k_2, \dots k_i$ r $t_1, t_2, \dots t_m$ epres $k_1, k_2, \dots k_i$ ents class groups. At each point in the tree C4.5 algorithm usually pick an attribute that gave the best split of the dataset. The attribute with the highest normalized value is chosen for the split and it is placed at the root of the tree. C4.5 is a supervised learning method that is simple and easy to implement. It divides dataset into portions with different characteristics. The last leave of the tree usually depicts predictions while the in between nodes depicts various test on the attributes (from the root node to the leaf node). The normalized value is calculated using equation 1 and equation 2.

$$\text{Gain}\,(q) = f(inf(T) - inf[(q,T)]$$

$$\text{Gain}\,(q) = f(inf(T) - inf[(q,T)] \quad (1)$$

where

$$\text{in}\,f(q,T) = \sum_{i=1}^{n} q_i \times \text{Ent}\,(q_i)$$

$$\text{in}\,f(q,T) = \sum_{i=1}^{n} q_i \times \text{Ent}\,(q_i) \quad (2)$$

Naïve Bayes Classifier

Naïve Bayes classifier is a probability based classifier and has proved its effectiveness in many areas where it has been applied. It is fast and easy to use which made it popular in data mining field. Though usually criticized for its attribute independent assumptions but it still competes favourably with other higher classifiers. Naïve Bayes calculates the probability value and selects the class with the highest probability (Taheri et al, 2014). It is represented mathematically as shown in equation 3

$$.P(K_i \cap Y) = \frac{P(Y \cap K_i)P(K_i)}{P(Y)}$$

$$.P(K_i \cap Y) = \frac{P(Y \cap K_i)P(K_i)}{P(Y)} \quad (3)$$

For a database with high dimension the computational cost is usually high thus the application of Naïve Bayes.

$$P(Y \cap K_i) = \prod_{k=1}^{n} P(Y_k \cap K_i) \quad (4)$$

$$P(Y \cap K_i) = \prod_{k=1}^{n} P(Y_k \cap K_i)$$

Naïve Bayes Algorithm
1) Input attributes and the class of the instances
2) Compute the posterior value for each attribute against the class
3) Compute the value before the existing class
4) Multiply the results from 2 and 3 for all the classes
5) Choose the highest value as the classification

Source: Taheri et al, (2014)

3. Data Mining Process Models

Data mining process required following some basic outlined steps when mining data. These steps outline all the necessary procedures for data mining. This process was originally proposed by Kurgan and Musilek (2006) and since then, many other mining processes have been developed. One common thing about all the process models is that they all outline steps which usually include loops and iterations (Kurgan and Musilek, 2006). CRISP-DM is a popular mining methodology that is generally accepted by data mining experts and it is a leading methodology used by data miners (Kurgan and Musilek, 2006). CRISP-DM is the process model that has been chosen for this work.

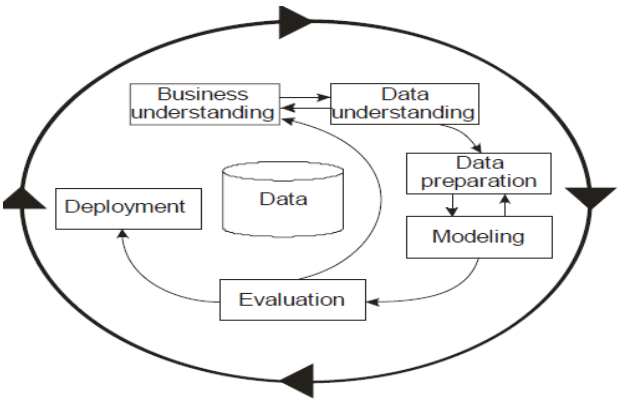


Figure 1: CRISP-DM Cycle

Source: (Gartner Group, 1995)

4. Methodology

The process followed in this research is as outlined in figure 2. The process started with business understanding which is all about understanding the problem domain and translating it to mining problem followed by the data collection stage where the required dataset was collected from selected Nigerian prisons. The exploratory analysis was done to get a better insight into the dataset before analysis. The classification stage was where the data mining proper was carried out using the two classifier namely C4.5 and Naïve Bayes classifiers. The evaluation stage is where the classification results from the two classifiers are discussed.

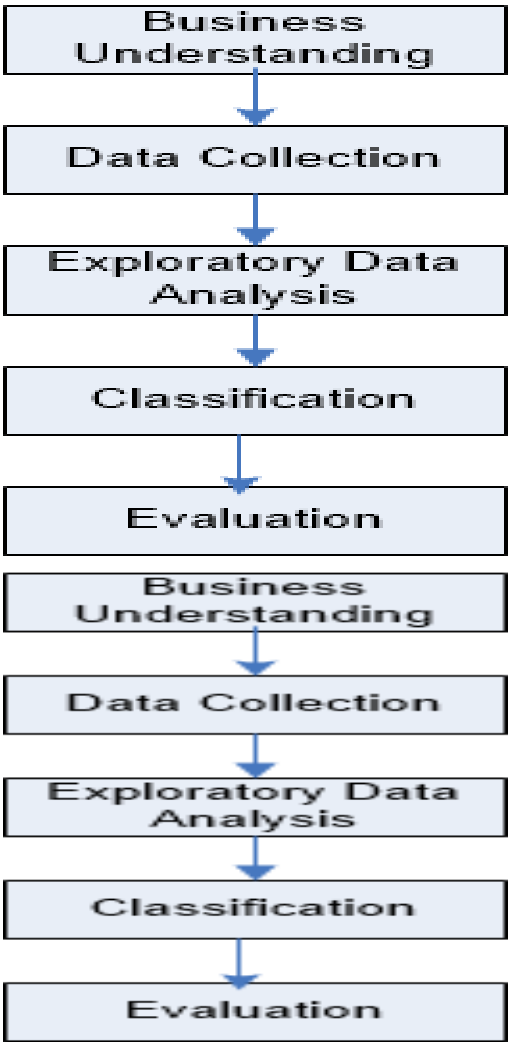


Figure 2: Work Methodology Flowchart

Dataset Description

The data used in this work were collected from selected Nigerian Prisons in Katsina, Kano, Kaduna, Niger and Abuja. Crime is broad term that covers range of unpleasant activities ranging from simple to complex. In this work crime was categorized into three groups as low, average and high crimes. The data consists of five fields and a class attribute categorized as shown below:

Education

- No Education: Implies not having any educational qualification
- Low Education: includes Primary and Secondary graduates
- Average Education: includes ND, NCE
- High Education: includes Degrees, PGD, MSC, MA,PhD

Occupation

- Unemployed: implies no work at all
- Self Employed: includes farmers, Apprentice, traders, artisans
- Employed: includes private employed and government employed

Age

- Early: ages (18 : 34)
- Middle: ages (35 : 50)
- Late: ages (51 : 150)

Crime

- Low Crime: Breach of Trust, Conspiracy, Assult
- Average Crime: Rape, Kidnapping, Drug
- High Crime: Homicide, Armed Robbery, Theft

Sex

- Male
- Female

Class

- Vulnerable
- Non Vulnerable

5. Performance Measures

The common measures for evaluating performance of data mining models are sensitivity, relevance, specificity, kappa statistics, area curve, time and accuracy.

Sensitivity: It is a statistics that shows the records that are correctly labelled by the classifier. It can be defined as: $\text{Sensitivity} = {TP}/{N}$

$\text{Sensitivity} = {TP}/{N}$

Specificity: It is simply a report of instances incorrectly labelled as correct instances; it can be defined as: $\text{Specificity} = {FP}/{N}$

$$\text{Specificity} = {FP}/{N}$$

Precision: Simply measures exact relevant data retrieved. High precision means the model returns more relevant data than irrelevant data.

$$\text{Precision} = \frac{TP}{\text{TP} + FP}$$

Kappa: measures the relationship between classified instances and true classes. It usually lies between $[0, 1]$, the v $[0, 1]$ alue of 1 means perfect relationship while 0 means random guessing.

Accuracy: this shows the percentage of correctly classified instances in each classification model

Time: Implies time taken to perform the classification (Milan and Sunila, 2011; Hong etal, 2006)

6. Exploratory Analysis of Crime Data

In this work WEKA mining tool was applied. WEKA is a machine learning software that has gained recognition in data mining because it implements many different data mining algorithms and also has potent tools for data pre-processing and visualization. It is an open source and accepts its data in Attribute Related File Format (ARFF). The sample of the converted ARFF file for this work is shown in figure 4.

```
@relation 'Formated Prisons2-
weka.filters.unsupervised.attribute.Remove-R1-2,11-19
-weka.filters.unsupervised.attribute.Remove-R7-
weka.filters.unsupervised.attribute.Remove-R3-
weka.filters.unsupervised.attribute.Remove-R3'
@attribute Offence {low,high,average}
@attribute Age {early,late,middle}
@attribute Sex {M,F,'M '}
@attribute Edu-Qualification {low,average,NONE}
@attribute Occupation {'Self Em-
ployed',unemployed,employed}
@attribute Class {vulnerable,'non vulnerable'}
@data
low,early,M,low,'Self Employed',vulnerable
high,early,M,average,'Self Employed',vulnerable
low,early,M,average,'Self Employed',vulnerable
average,early,M,low,'Self Employed',vulnerable
average,early,M,low,'Self Employed',vulnerable
high,early,M,average,'Self Employed',vulnerable
average,early,M,low,'Self Employed',vulnerable
low,early,M,low,'Self Employed',vulnerable
high,early,M,low,'Self Employed',vulnerable
high,early,M,low,'Self Employed',vulnerable
low,early,M,low,'Self Employed',vulnerable
low,early,M,average,unemployed,vulnerable
average,late,M,NONE,employed,'non vulnerable'
low,early,M,average,'Self Employed',vulnerable
average,early,M,average,'Self Employed',vulnerable
```

Figure 4: A Sample ARFF for the Crime dataset

When preparing data for data mining seeing the data pictorially provides insight into what is happening and this insight can help improve model building. The data mining tool chosen for this work has the features for exploratory data analysis. The relative densities of the various attributes in the data set are as shown in figure 5.

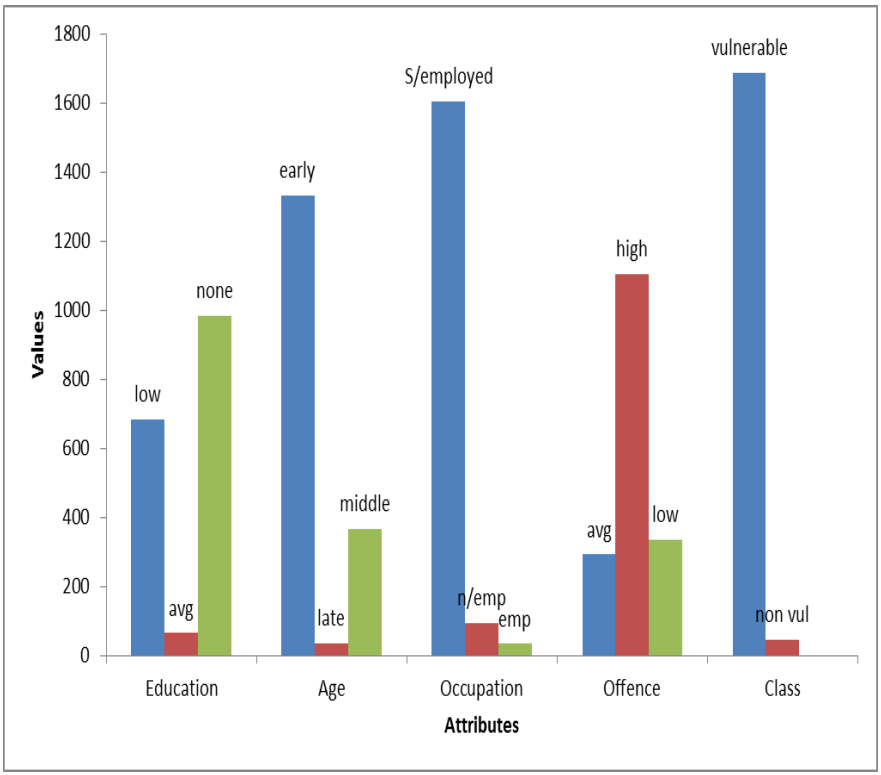


Figure 5: Densities of Attributes in the dataset

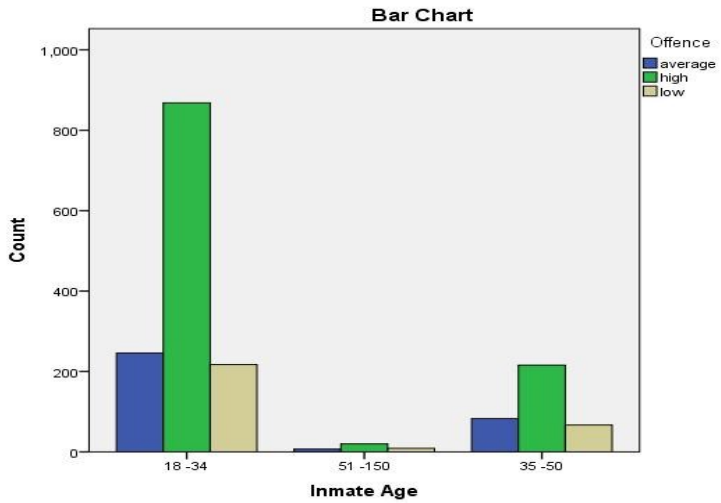


Figure 6: Association of Age and Offence

The visualization of the relationship between the age and offence reveals that majority of the offenders are within the ages of 18 -34 (early age) and commit high crime.

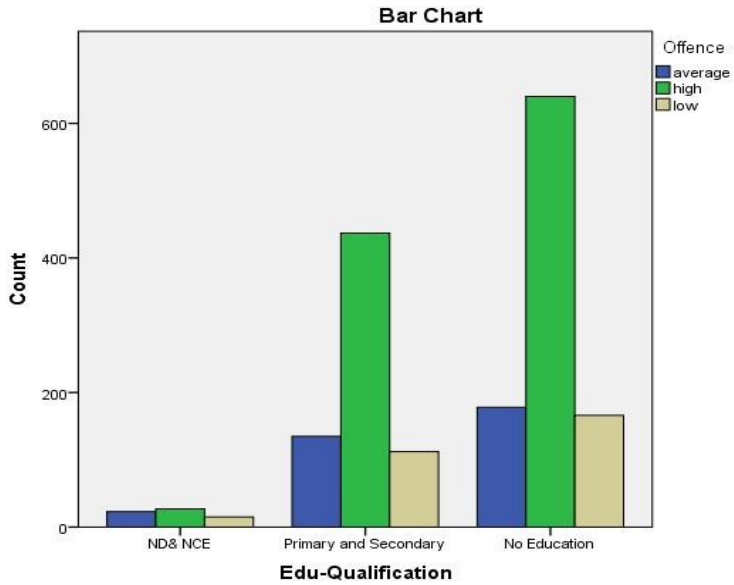


Figure 7: Association of Educational Qualification and Offence

The visualization of the offence versus educational qualification reveals that majority of the offenders have low education qualification (primary and secondary) or no education qualification at all and that they commit high crime and are vulnerable groups.

7. Classification of Crime Dataset

In this work k-Fold cross-validation method has been applied In order to ensure good performance of the classification model. The method was used to train and test the classifier. This method usually divides dataset into k folds; it trains the model with k-1 folds and tests the built model with the remaining k fold. It usually obtained k different results and takes the average to obtain the model accuracy. In this work, 10 fold cross validation was used. This method is better than the random sampling method because it takes care of the bias usually associated with random sampling method. The Classification was done using C4.5 classifier which is a decision tree classifier; Naïve Bayesian which is a probability based classifier de-

veloped to handle categorical dataset was used for reliability test. Table 1 is the tabulation of the result obtained from the two classifiers using WEKA mining tool.

Table 1: Tabulated Results

Evaluation Metrics	NB	C4.5
Time	0.05 secs	1.06 secs
Accuracy	93	97
TP Rate	0.935	0.971
FP Rate	0.067	0.027
Kappa	0.8696	0.9409
Precision	0.935	0.971
Recall	0.935	0.971
ROC curve	0.989	0.986

The result above revealed that the C4.5 classifier has better accuracy of 97 in comparism to the accuracy of Naïve Bayes. C4.5 though took more time of 1.06 seconds to build the model compare to 0.05 seconds taken by the Naïve Bayesian still handles the data better. In terms of classifier performance using the ROC curve the C4.5 classifier performed comparably well against Naïve Bayes on the dataset.

ROC curve is used to visualize classifiers performance. It is usually plotted using sensitivity at the y axis and specificity at the x axis. If the area under the curve is 1, it indicates perfect prediction while 0.5 implies random guess. The areas under the curve for the naïve Bayesian and C4.5classifiers are close to 1 which indicates the classifiers performed well.

8. Deductions from C4.5 Classification Tree

i. Offence = high implies vulnerable
ii. Offence = low and Education Qualification = high implies non vulnerable
iii. Offence = low and Education Qualification = average implies non vulnerable
iv. Offence = low, Education Qualification = NONE, Age = early, and Sex = male implies vulnerable
v. Offence = low, Education Qualification = NONE, Age = middle implies non vulnerable
vi. Offence = low, Education Qualification = NONE, Age = late implies non vulnerable

vii. Offence = low, Education Qualification = NONE, Age = early, Sex = F and Occupation = employed or unemployed implies vulnerable
viii. Offence = low, Education Qualification = low, Age = middle implies non vulnerable
ix. Offence = low, Education Qualification = low, Age = early and Sex = M implies vulnerable
x. Offence = low, Education Qualification = low and Age = late implies non vulnerable

9. Conclusion

Data mining has the capability that makes it simple convenient and suitable for data extraction from large databases. It employs different mining algorithms for its work. Many agencies gather data for its operational purposes, such data can be mined to discover some relevant patterns that can aid in decision making. The analysis of crime data will help to unravel crime pattern and nature of those who commits such crime so that appropriate strategies and rules will be put in place to control such crimes. The work reveals that the majority of the inmates that commit crime are between the ages of 18 to 34 and have low or no educational qualification and are either self employed or not doing anything at all. The classification result reveals that 98 percent of these groups of people are threat to the society. Thus, the researchers are of the opinion that government should encourage education and our youths should be gainfully employed.

Correspondence
Obuandike Georgina N.
Department of Mathematical Sciences and IT
Federal University Dutsinma
Katsina state, Nigeria

References

Barnett, V., Lewis T., Outliers in Statistical Data. John Wiley, 1994.

Brown, D. (2003). The Regional Crime Analysis Program (RECAP): A Framework for Mining Data to Catch Criminals. http://vijis.sys.virginia.edu/publication/RECAP.pdf

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., and Chau, M. (2004). Crime Data Mining: A General Framework and Some Examples. *Computer,* 37(4), 50-56.

Fayyad, U.M. and Uthurusamy, R. (2002), "Evolving Data Mining into Solutions for Insights", Communications of ACM*,* 45(8), 28-31.

Gartner Group (1995), Gartner Group Advanced Technologies and Applications Research Note http://www.gartner.com.

Hawkins, D., Identification of Outliers, Chapman and Hall, 1980.

Hong, H., Jiuyong, L., Ashley P., (2006) "A Comparative Study of Classification Methods for Microarray Data Analysis", published in CRPIT, Vol. 61.

Jiawei, H., Micheline, K., and Jian P. (2012)"Data mining: Concept and Techniques" 3rd edition, Elsevier,

Johnson, R., (1992) "Applied Multivariate Statistical Analysis", Prentice Hall.

Julio, P. and Adem, K. (2009) Data Mining and Knowledge Discovery in Real Life Applications, ISBN 978-3 -902613-53-0, pp. 438, I-Tech, Vienna, Austria.

Kurgan, L. and Musilek, P. (2006); "A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review. Volume 21 Issue 1, pp 1 - 24, Cambridge University Press, New York, NY, USA.

Megaputer Intelligence, Inc. (2002). Crime Pattern Analysis: Megaputer Case Study. http://www.elon.edu/facstaff/mconklin/cis230/cases/crime_pattern_case.pdf

Milan, K., Sunila, G.,(2011) "Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction", IJCST, Vol. 2, Issue 2, pp. 304-308,.

Naisbitt J. (1986) " *Megatrends"*, 6th ed., Warner Books, New York.

Otto, G. and Ukpere, W. I., (2012) "National security and development in Nigeria", African Journal of Business Management Vol.6 (23), pp. 6765-6770.

Taheri, S., Year Wood, J., Mammadov M. Seifollahi S. (2014) "Attribute Weighted NaïveBayes Classifier Usinga Local Optimization", *Neural Computing and Application*, Volume 24, Issue 5, pp. 995–1002.

Williams, G. J., Baxter, R. A., He H. X., Hawkins S., Gu L.,(2002) "A Comparative Study of RNN for Outlier Detection in Data Mining," IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102.

Wilson , O.W. (1963). *Police Administration*. USA, McGraw Hill Company.

Witten, I. and Frank, E. (2000). *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann publishers.

ZhaoHui, T. and Jamie, M. (2005), "Data Mining with SQL Server 2005",Wiley Publishing Inc, Indianapolis, Indiana, 2005.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.68-83

Crime Control Using National Social Security Numbering System

ADEJORO CORNELIUS ONIMISI
University of Nigeria, Nsukka, Nigeria

OGBUAGU-UJAH BRIDGET
Nigeria Defense College Abuja, Nigeria

ABSTRACT This paper is concerned with how National Social Security Numbering System can be used to control crime. A typical Social Security Number (SSN) is a 9-digit number configured and assigned to citizens and legal migrants of a country for the purpose of uniquely identifying them to aid easy administrative purposes of a country. In principle, a National Social Security Number can be used to track citizens, permanent residents, temporary residents, legal emigrants for the purpose of work, taxation, government benefits and other governmentally-related functions, health care, open a bank account, obtain a credit card, drive a car, facilitate payment procedures, etc. The motivation for carrying out this research is to checkmate an orchestration of illegal activities in the country such as insurrection, advance free fraud, white collar theft, irregularities in admissions and recruitment. This paper aim is to develop a security measure that will make it almost impossible to falsify one's Personal Identifiable Information (PII); Provide a unified and centralized database management system for the National Social Security Numbering project for the Federal Republic of Nigeria that will make it difficult for illegal immigrant that comes into Nigeria with the aim of committing crime and for Nigerians who will want to commit fraud by hiding their identity. The Object Oriented System Analysis and Design Methodology (OOSADM) was used for the analysis, design and development of the National Social Security Numbering system with the aid of these programming technologies: JavaScript, Cascading Style Sheet version 3, Hypertext Markup Language version 5, PHP Hypertext Pre-Processor version 5, Asynchronous JavaScript and Extensible Markup Language (AJAX), PHPMyAdmin and MYSQL database server version 5.5. Performance of the system was evaluated with a sample of one hundred National Social Security Numbers which were randomly generated and assigned to one hundred fictitious names that were assumed to have formally applied for NSSN by filling out a short electronic form with their Personally Identifiable Information (PII). Each NSSN was later entered into a search bar in a user graphical interface designed for law enforcement agents and the system was able to display a complete personal profile of the owner of a test NSSN. Subsequently, a search was conducted with a wrong NSSN and the system reported an invalid NSSN. A system that can instantly generate an NSSN in accordance with the proposed model following an application made for NSSN by citizen/immigrant has been developed and further work is ongoing.

Keywords: National Social Security Number, Numbering System, Immigrant, Citizen, Security

Introduction

Nigerian governments need to assign a unique and National Social Security Number (NSSN) to each citizen of the country, and legal migrant by putting his or her Personally Identifiable Information (PII) on a central database as an efficient method of identification for several purposes such as: security intelligence services, orderly nation administration, to mention but a few. This study focuses on developing a database-driven Social Security Numbering scheme for the Federal Republic of Nigeria whose intent and purpose are to randomly generate and assign Social Security numbers to millions of existing and future generation of Nigerians for orderly administration and crime control using computers, tablets and phones.

The use of the Social Security Number (SSN) has expanded significantly since its inception in 1936. An SSN as it is called in the some countries is a 9-digit number consisting of a 4-digit serial number, a 2 digit year of birth indicator, and a 3-digit number indicating the geographic area of registration. Alessandro and Ralph (2009). This research focuses on developing a National Social Security Numbering System (NSSN) that can uniquely identify each bona-fide citizen of Nigeria and legal migrant and hold their Personal Identifiable Information (PII) in a centralized database server.

The current attempt by the federal government of Nigeria to identify Nigerian citizens with their Driver's Licenses, National Identification Cards, Independent Electoral Commission (INEC) Voters Registration Cards and ECOWAS International Passports is characterized by the following problems:

- There is no particular unified method of identifying millions of Nigerian citizens and migrants, and consequently, capturing their individual records or profiles in bits from the Driver's Licenses, National ID cards, International Passports, etc. is not comprehensive.
- Many Nigerians do not readily have all the above mentioned means of identification, and thus, it is not easy to identify each person in the country;
- None of these existing means of identification holds complete information about a citizen such as bank records, academic qualifications, cultural background information (nativity, village or town name), etc.;
- There are some discrepancies in the Personal Identifiable Information (PII) entered on some of the above mentioned means of identification for some people owing to falsification or hoarding of information;
- There is no unified central database anywhere in the country for official reference.

The objectives of this project are to:

- Develop a model for the generation of the National social security Number.
- Develop a piece of software for National Social Security Numbering that can uniquely identify each bona-fide citizen of Nigeria and legal migrant and hold their Personal Identifiable Information (PII) in a centralized database server;

- Develop a security measure that will make it almost impossible to falsify one's Personal Identifiable Information (PII);
- Provide a unified and centralized database management system for the National Social Security Numbering project for the Federal Republic of Nigeria.

The scope of this study covers the entire Nigerian population for which it is developed to generate national social security numbers (NSSNs) for its citizens and legal immigrants. A unified and centralized database management system with the capacity of hosting nearly 1 billion randomly generated National Social Security Numbers will be quite instrumental to government officials not only for citizenship identification purposes, but for orderly national administration. With this software engineering effort, tracing the identity of any person on Nigerian soil can be done swiftly and simultaneously from any part of the world over the web using computers, mobile phones, tablets and any internet-enabled device.

Literature Review

This part of the paper reviews related literatures written by different authors on Social Security Numbers with a view to giving this research the all-important knowledgebase required for its design and development. The Social Security number (SSN) was created in 1936 for the sole purpose of tracking the earnings histories of U.S. workers, for use in determining Social Security benefit entitlement and computing benefit levels. Since then, use of the SSN has expanded substantially. Today the SSN may be the most commonly used numbering system in the United States. As of December 2008, the Social Security Administration (SSA) had issued over 450 million original SSNs, and nearly every legal resident of the United States had one. The SSN's universality has led to its adoption throughout government and the private sector as a chief means of identifying and gathering information about an individual Carolyn (2009). Creating the SSN scheme and assigning SSNs to U.S. workers was no easy task. Passage of the Social Security Act in August 1935 set in motion a huge effort to build the infrastructure needed to support a program affecting tens of millions of individuals. Carolyn (2009).

A Social Security Number encompasses Personal Identifiable Information (PII) of a citizen and demands strict security measures to be put in place for avoidance of identity theft. Buttressing this fact, the escalation of security breaches involving personal identifiable information (PII) has contributed to the loss of millions of records over the past few years. Erika, Tim and Karen (2010) . Though the Social Security Number was originally established for the Social Security program, other government agencies soon realized that it could be used as a convenient identifying number for tracking other government programs. Use of the SSN as a federal government identifier was based on Executive Order 9397, issued by President Franklin Roosevelt. Erika, Tim and Karen (2010)

It is very important that you do not share your Social Security Number with anyone else as this can lead to Identity Theft. According to Fact sheet (2016), the Social Security Administration (SSA) uses the Systematic Alien Verification for

Entitlements (SAVE) Program's Verification Information System (VIS) of the Department of Homeland Security (DHS) as its primary data source to verify legal entry into the United States and, in conjunction with travel documentation. Erma and Felix (1982)

Many countries have different names for a Social Security Number, criminals may make use of enhanced predictability to generate someone's Social Security Number: In principle, a Social Security Number (or an equivalent such as a National Insurance Number in the United Kingdom) is an identifier, not an authenticator. It would be unsuitable for a password, because it isn't secret David (2009). In fact, an SSN is essentially a database primary key, an identifier that is unique to you and to your individual personal records in the Social Security Office's database. The most practical way of generating such a key is often to enhance predictability, not to reduce it, in keeping with words of the writer. David (2009). Quoting the Social Security Office, David (2009) agrees with the other author Erma and Felix (1982) that the nine digits of the Social Security Number are grouped as follows:

- The first three digits represent the Area Number.
- The next two digits represent the Group Number.
- The four digits at the end are called the Serial Number,

Detailed explanation that Social Security Number (SSN) is a unique, 9-digit identification number, issued by the U.S. Social Security Administration (SSA) to U.S. citizens, permanent residents, and qualified foreign nationals - including those who meet the eligibility requirements for a Social Security Number such as students with on-campus employment, students with CPT (Curricular Practical Training) authorization, students, scholars, and dependent family members with EAD (Employment Authorization Document). Fact Sheet (2006). However, it was not until the 1960's that federal agencies began to adopt the SSN as a general government identifier in other contexts, and this effort was to devise a method for uniquely identifying the earnings records for the millions of persons covered by the new law Since entitlement to Social Security and the benefit amount were to be determined from a person's earnings over many years, a method was needed for maintaining permanent and accurate earnings records for each person working in employment covered by the Social Security program. Fact Sheet (2006). Social Security number was developed for this purpose. This unique configuration, plus the fact that an SSN is used for many purposes besides employment (income tax returns, bank accounts, drivers' licenses, and so forth), makes the number easily recognizable. Although most people believe that each part of the number has a special significance, few know what that significance is. Fact Sheet (2006).

Until 1972, the area number indicated the location (State, territory, or possession) of the Social Security office that issued the number. When the Social Security numbering system was developed, one or more area numbers were allocated to each State based on the anticipated number of issuances in the State. Because an individual could apply for an SSN at any Social Security office, the area code did not necessarily indicate where the person lived or worked. Since 1972, the Social Security Administration has been issuing SSN's centrally from its headquarters in

Baltimore. The area code now indicates the person's State of residence and shows on an SSN application. Fact Sheet (2006).

In supporting the use of Social Security Number as a tool for curtailing and controlling the movement of non-immigrants into a country, Fact Sheet (2006) reveals that the Social Security Administration (SSA) uses the Systematic Alien Verification for Entitlements (SAVE) Program's Verification Information System (VIS) of the Department of Homeland Security (DHS) as its primary data source to verify legal entry into the United States and, in conjunction with travel documentation, to verify the immigration status of non-citizen Social Security number (SSN) applicants.

A Social Security Number (SSN) is different from a Social Security program. The former is a 9-digit identification number issued by the US Social Security Administration (SSA), a requirement for all persons who work and receive pay in the US as well as used to report wages to the government, SSN bulletin (2014) whereas SSA bulletin (2014) sees the Social Security as government program to help not only older Americans, but also workers who become disabled and families in which a spouse or parent dies. Your link with Social Security is your Social Security number. You will need it to get a job and to pay taxes. We use your Social Security number to track your earnings while you are working and to track your benefits after you are getting Social Security. SSN bulletin (2014). However, in the opinion of the Chris (2011) who criticized the use of Social Security Numbers as primary keys in database design, some of the issues in the decision of database for National Social Security Numbering system are: Uniqueness, Universality, Identification, and Security.

On the issues that surround getting a new SSN, SSA does not routinely assign new numbers, they will do so when a victim requests a new SSN and provides evidence that he/she has tried to resolve the problems brought on by identity theft but continues to be disadvantaged by the SSN misuse. Disadvantaged by misuse of the SSN means that the misuse has caused you financial or personal hardship within the past year. In the author's assertion, Fact Sheet (2013), Although, no single law comprehensively governs the use and disclosure of SSNs, certain federal laws restrict the use and disclosure of personal information, including SSNs, by government agencies or private sector entities. Cynthia (2006). Federal laws that require the use of an SSN generally limit its use to the statutory purposes described in each of the laws. For example, the Internal Revenue Code, which requires the use of SSNs for certain purposes, declares tax return information, including SSNs, to be confidential and prescribes both civil and criminal penalties for unauthorized disclosure. Barbara, Roland and Jacquelyn (1999).

Effort for SSN was to devise a method for uniquely identifying the earnings records for the millions of persons; they however differs in the number of digit as compare to that of Greece called AMKA. AMKA has the following 11-digit format: YYMMDDxxxyz, where the first 6 digits encode the person's date of birth (YYMMDD), the following 4 digits are a sequence number for people born on that date (xxxy) and the last digit is a control digit (z). The sex of the person is encoded in the last digit of the sequence number (y of xxxy): even digits are assigned to women and odd digits are assigned to men. This results in disclosure of both the

date of birth and the sex of a person by solely looking at their AMKA. Eleni, Alexandros, Sotiris, (2009).

A publication of California-Berkeley School of Law (2007) posited that at least 36 states have enacted legislation requiring organizations that possess sensitive personal information to warn individuals of security breaches. California led the way in the creation of these laws, driven by concerns about identity theft and tax information security.

On the possibility of predicting an SSN, the authors, Alessandro and Ralph (2009) seems to put Greek's style of SNN otherwise known as AMKA as stated by Eleni, Alexandros, Sotiris, (2009) under serious vulnerability. The authors, Alessandro and Ralph (2009), Maintained that Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. Our prediction algorithm exploits the observation that individuals with close birthdates and identical state of SSN assignment are likely to share similar SSNs. Alessandro and Ralph Gross (2009).

The author Kristin (2014) pointed out that Policymakers continue to be concerned with securing the economic health of the United States, including combating those crimes that threaten to undermine the nation's financial stability. Identity theft, for one, poses both security and economic risks. He further stated that an increase in globalization and a lack of cyber borders provide an environment ripe for identity thieves to operate from within the nation's borders as well as from beyond.

Today the SSN has a unique status as a privacy risk. No other form of personal identification plays such a significant role in linking records that contain sensitive information that individuals generally wish to keep confidential. And an identity thief armed with a name and an SSN can often open new credit or bank accounts, rent an apartment, get a job, get arrested and create a criminal record for someone else, or even have surgery and pollute the victim's medical records. SSN publication (2014).

Buttressing the issue of restriction, Kathleen (2008) narrated that as early as the 1970s, concerns regarding increased uses of the SSN by both government and private entities prompted studies and subsequent congressional action limiting government uses of the SSN. The Social Security Administration created a task force in 1970 to investigate "non-program" uses of the SSN, and the task force's report the following year stated: any examination of SSN policy must begin with the recognition that the number has ceased to be merely a "social security number." Especially in the past few years, the number has come into increasingly wide use as a numerical identifier throughout society, to the point where the adult American citizen is beginning to need a number to function effectively even if he is among the very small minority of people who never work in covered employment. In any situation where a government agency asks for a person's SSN, the agency, under section 7(b) of the Privacy Act, is required to tell the person whether the request is

mandatory or voluntary, and what uses the agency will make of the SSNs that are collected. Kathleen (2008).

Best practices to protect SSNs from disclosure include reducing the unnecessary use and collection of SSNs, securing electronic records systems, using management controls, and reducing or protecting the transmission of records containing SSNs. John G. Morgan (2008).

System Analysis and Design

We undertake the analysis and design of a national security numbering system with the aid of Object-Oriented System Analysis and Design methodology and the following Unified Modeling Language (UML) techniques: use case diagram, class diagram and two collaboration diagrams.

Analysis of Existing and Proposed System

Traditional means of identifying Nigerian citizens and immigrants with legal rights to live in Nigeria by local, state or federal government authorities include one of the following: National Identification Card, Permanent Voter's Card, Driver's License, Employment Identification Card, School Identification Card or International Passport. But unfortunately, none of these means of identification has been efficiently computerized to provide a quick and reliable on-the-spot details of persons under security scrutiny. There are many doubtful situations in which people are being demanded to prove their identities on the spot: in the bank, at schools, in the street, on major highways by police, military or other uniformed men and women at check points, etc. Following is a listing of the activities that take place in the existing system. For instance the voter's card system and National Identification system.

- A citizen or immigrant may be requested to identify himself or herself by either law enforcement agents or for officials purposes;
- A citizen or immigrant would most likely tender his or her National Identification Card, Permanent Voter's Card, Driver's License, Employment Identification Card, School Identification Card or International Passport as means of identification;
- If the person who received an identification document, presented by a citizen, is satisfied, then there would not be any issues.
- A citizen citizen who hard registered or applied for voter's card or National Identification card but misplaced it would be requested to go to court, get affidavit and repeat the registration again.

Some of the problems inherent in the existing system are:

- Lack of any central database of NSSNs for reference purposes;
- Some of the activities regarding identification of citizens and/or immigrants are paper-based, and therefore, done manually.

The Proposed System

Summary of Model for National Social Security Numbering System

The mathematical model for the proposed system is formulated to be generated at three different stages. The first stage will be the generation of the state code. The states code will use a two digit significant value, the local governments for all Nigerian states will also have two digit significant value. Thirdly, the individual number of person will be contained in five digits and this is serial. Because of the requirement of the internationally accepted standard of nine digits, the NSSN for any applicant will contain nine digits. The model shows how the NSSN how is generated.

$NSSN_L = X_{Li+} X_{Lj}+S_N code$
$NSSN_N = X_{Ni+} X_{Nj} + S_N code$
$NSSN_L$ Stands for NSSN generated in state with large population while $NSSN_N$ stands for those generated in state with normal population. $S_N code$ means serial number in each NSSN.

Discussion

The proposed multi-functional national social numbering system will operate in the following ways after development:

- The social security numbering information software application will run from a central server on network mode over the Internet;
- The server will provide access to multiple users via computers and other mobile devices such as tablets and Personal Digital Assistants (PDAs) from remote locations simultaneously;
- The NSSN software application will support cross-referencing of different identity information databases in the country running on database drivers, operating systems and hardware utilities such as the electoral database of Independent National Electoral Commission (INEC), National Driver's Licensing database, National Identity database, school portal databases, etc. that support the proposed platform with a view to providing a timely and accurate information unique to a Nigerian citizen or immigrant;
- The operation of the software works in such a way that upon confirmation of an NSSN, a citizen or immigrant can print his or her NSSN in hard copy by clicking on a print link;
- As a built-in security structure, the following data cannot appear twice in the NSSN database to avoid double application for NSSN: national ID card number, international passport number, driver's license number, permanent voter's card number, mobile phone number and email address.
- The software allow an applicant who has just applied for an NSSN or a citizen/ immigrant that has already obtained an NSSN to check whether his or her NSSN is ready or not, or just to confirm his or her NSSN should he or she had forgotten it.

Use Case Diagram of the System

The UML Use Case diagram shows all the actors and the use cases, actions or activities they perform in the proposed system. Figure 3.1 shows the use case diagram.

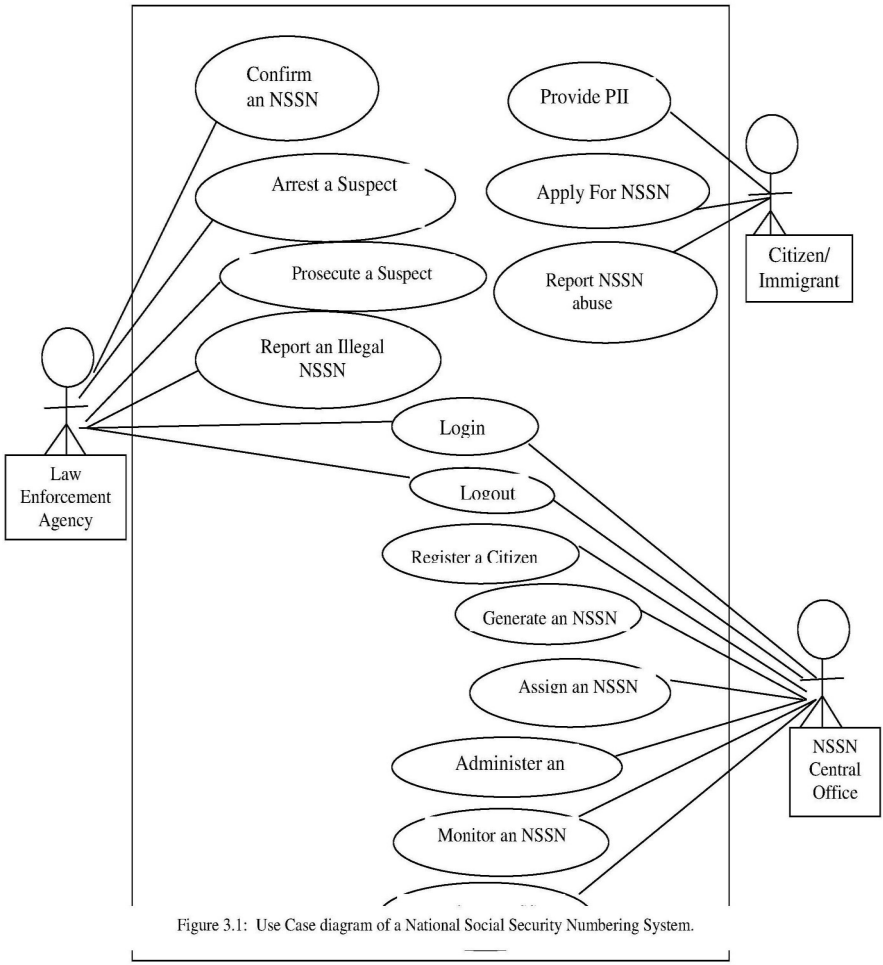


Figure 3.1: Use Case diagram of a National Social Security Numbering System.

3.1 can be described thus:

- The large central rectangular box represents the National Social Security Numbering system under development and contains the respective use cases of the software application. A use case is an activity, operation or function that must be undertaken by an object. An example of a use case is Check NSSN, which depicts the

action a law enforcement officer (i.e. an object of the NSSN system), has to take in an attempt to search for NSSN of a citizen;

- Law enforcer, citizen or immigrant, NSSN office, and MYSQL relational database engine are all users or objects of the system.
- The small oval shapes depict use cases or operations of various objects of the NSSN system. There are twenty use cases in the proposed system and they are listed as follows: Confirm NSSN, Arrest A Suspect, Prosecute A Suspect, Report Illegal NSSN (for Law Enforcer object); Provide PII, Apply For NSSN, Cancel NSSN Application and Report NSSN Abuse (for Citizen or Immigrant object); Register, Generate, Assign, Administer, Monitor, De-assign and Report (for NSSN office object); Store Generated NSSN, Store Citizenship Record and Store Real-Time Activity Log (for MYSQL Database Engine object); Login and Logout.

System Design

Figure 3.1 can be described thus:

- The large central rectangular box represents the National Social Security Numbering system under development and contains the respective use cases of the software application. A use case is an activity, operation or function that must be undertaken by an object. An example of a use case is Check NSSN, which depicts the action a law enforcement officer (i.e. an object of the NSSN system), has to take in an attempt to search for NSSN of a citizen;
- Law enforcer, citizen or immigrant, NSSN office, and MYSQL relational database engine are all users or objects of the system.
- The small oval shapes depict use cases or operations of various objects of the NSSN system. There are twenty use cases in the proposed system and they are listed as follows: Confirm NSSN, Arrest A Suspect, Prosecute A Suspect, Report Illegal NSSN (for Law Enforcer object); Provide PII, Apply For NSSN, Cancel NSSN Application and Report NSSN Abuse (for Citizen or Immigrant object); Register, Generate, Assign, Administer, Monitor, De-assign and Report (for NSSN office object); Store Generated NSSN, Store Citizenship Record and Store Real-Time Activity Log (for MYSQL Database Engine object); Login and Logout.
- There are arrows with dotted lines seen within the use case diagram. They link one use case with another. An action, operation, function or activity that must take place first before the other is shown with an arrow head pointing towards it whereas the action at the tail of a dotted line signifies what follows after the first action. The following use cases are connected by the dotted lines: confirm NSSN and login; provide PII (i.e. Personally Identifiable Information) and register; apply for NSSN and assign; monitor and store real-time activity log; de-assign and store citizenship record; administer and report NSSN abuse; generate and store generated NSSN.

The following Unified Modeling Language notations and models will be employed in the development of the proposed national social security numbering system: use case diagram, class diagram, collaboration diagrams, activity diagram and database design diagram.

UML Collaboration Diagram of the System

UML collaboration diagram (See figure 3.3) shows how an operation within the system is executed by some programmed sequence of logic.

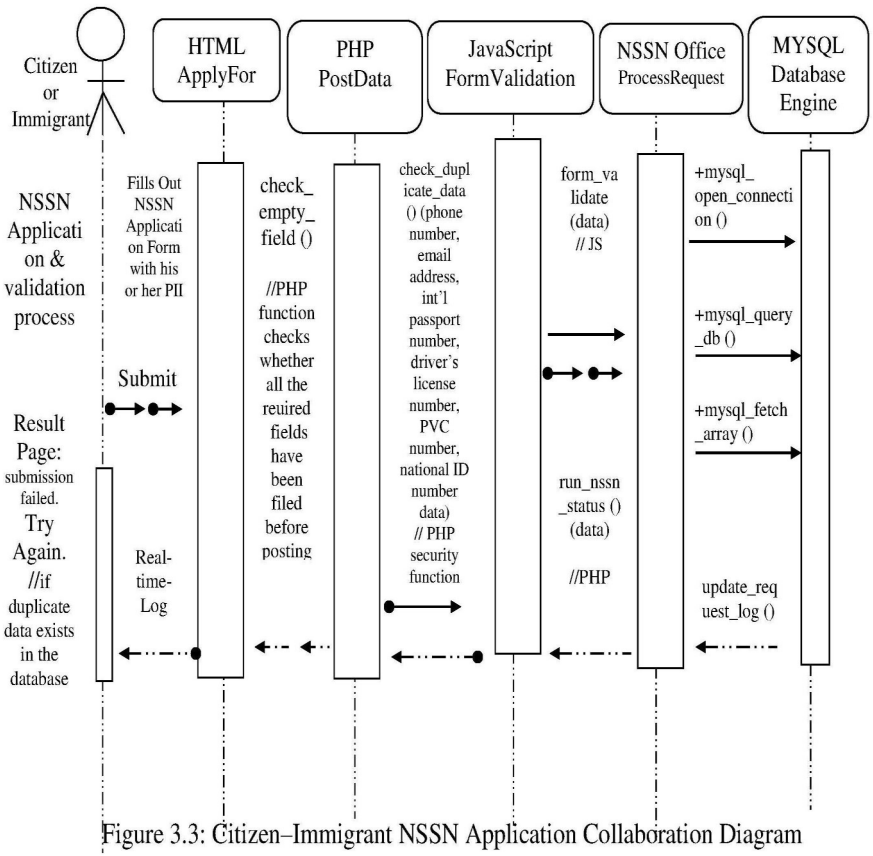


Figure 3.3: Citizen–Immigrant NSSN Application Collaboration Diagram

Figure 3.3 above shows a collaboration diagram for real-time process during an NSSN System (NSSN) headquarters. The first item from the left hand side labeled 'citizen or immigrant' represents a citizen or immigrant who will use this National Social Security Numbering System (NSSN) headquarters. The first item from the left hand side labeled 'citizen or immigrant' represents a citizen or immigrant who will use this National Social Security Numbering (NSSN) system. The NSSN application sub-system includes is a simple HTML NSSN application form with embedded JavaScript code on it. A citizen or immigrant will have to fill out an HTML NSSN application form with his or her Personally Identifiable Information (PII) before submitting it. A JavaScript code will run some validation routine checks on

the HTML NSSN application form prior to its submission. Once the user clicks on the submit button, PHP activates a subroutine written to check against an attempt to post duplicate entries in the database. This PHP subroutine will not allow a citizen or immigrant to request for an NSSN using another person's phone number, email address, international passport number, national ID card number, permanent voter's card number, driver's license number, etc. Upon submission of the NSSN application form, the system will check the citizen's status in the database. If a citizen or immigrant already has an NSSN, the system will simply display his or her NSSN on a page, otherwise, the system will return a message: "Your NSSN is not available. Please check back later". However, it must be known that no user can check his or her NSSN without logging into the system using a Security Access Code (SAC) which the NSSN system will have generated and emailed to him or her during an NSSN application process.

Design of the Proposed System

The operation of the proposed national social security system is as shown (Table 3.1) below in the database design, system architecture and input/output design.

Database design

NSSN Central Office

Table 3.1 NSSN Central Office

Field	Data Type	Size	Description	Action
-nssnofficeid	Long	254	Defines auto-increment integers for new records	Primary Key
-nssnservername	Varchar	30	Defines an NSSN server name	
-nssnservernicserial	Varchar	30	Defines a serial number of a Network Interface Card of NSSN server	
Nssnserverwlanicserial	Varchar	30	Defines a serial number of Wireless LAN card on the server	
Nssnservertype	Varchar	50	Defines the type of Operating System running on the server hosting NSSN application	
nssnservermodel:	Varchar	30	Defines a model of NSSN server	

System Architecture

The figure 3.5 below shows the system architecture of a National Social Security Numbering software application.

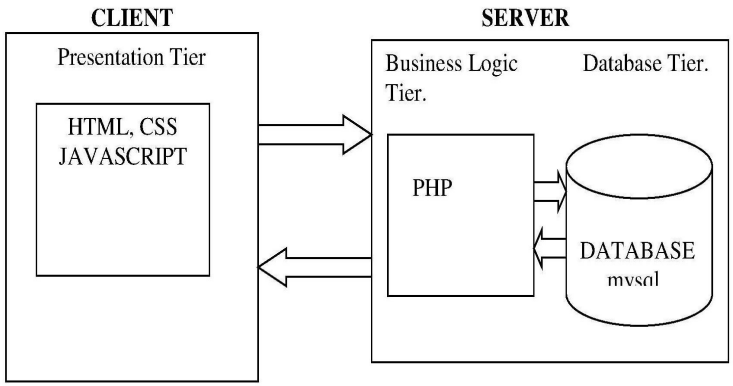


Figure 3.5: System Architecture

Presentation Tier: Responsible for rendering HTML requests such as apply for nssn, cancel nssn application, login, logout, track nssn, report nssn abuse or theft, etc. in the graphical user interface.

Business Logic Tier: Business rules and logical cum security constraints that prevent a citizen from applying for NSSN using another person's email address, phone number, international passport number, national ID card number, permanent voter's card number, driver's license number, etc are implemented here as well as enforcement of Referential Integrity to provide for data security and integrity.

Data Access Tier: Data access tier is responsible for database manipulation and communication. This layer consists of the Database for storing data. The database system used is MYSQL.

Results

i. We have successfully designed and developed a software system that can generate and assign a national social security number for Nigerians and legal immigrants.

ii. We have been able to create a platform by means of this piece of software application that is capable of holding hundreds of millions of records for each Nigerian citizen and legal immigrant within a centralized database.

Conclusion

We already have a system that can instantly generate an NSSN in accordance with the proposed model following an application made for NSSN by citizen/immigrant. But this research is still open for further investigation, in particular to developed a most suitable model that can comfortably show how 9 digit unique but serial number accommodates over 170 million Nigerians.

Limitation of the Proposed System

This research is limited to the adoption of Voters card, National Identity Card and the likes as a means of identifying who is legible to get an NSSN, but there is need to introduce a biometrics to guide against non Nigerians or illegal immigrants from getting to do registrations as the above means cannot be said to be sufficient enough.

Correspondence
Adejoro Cornelius Onimisi
Department of Computer Science
Faculty of physical sciences
University of Nigeria, Nsukka, Nigeria
Email: specialcornel@gmai.com

Ogbuagu-Ujah Bridget
Nigeria Defence College Abuja, Nigeria
Email: bridgechi2000@yahoo.com

References

Alessandro Acquisti and Ralph Gross. “Predicting Social Security Numbers from public data”. Carnegie Mellon University, Pittsburgh. May 2009. Pages 1, 3.

Barbara Bovbjerg, Roland H. Miller, Jacquelyn Stewart, “ Social Security: Government and commercial use of Social Security Number is Widerspread”. A report to the Chairman subcommittee on Social Security, House of Representative, USA, 1999. Page 8.

Carolyn Puckett, “The Story of the Social Security Number” Social Security Bulletin Vol. 69, No. 2 2009. Pages 1, 2 http://www.ssa.gov/policy/does/ssb/vb9n2

Chris Hibert, "What do you do when they ask for your Social Security Number". Available at http://www.biuldfreedom.com/tl/tl17b.shtml. page 4. 2011

Cynthia M. Fagnoni, "Social Security Numbers: More Could Be Done to Protect SSNs". United States Government Accountability Office. March 30, 2006. Page 11.

Eleni Gessiou, Alexandros Labrinidis, Sotiris Ioannidis, "A Greek (Privacy) Tragedy: The Introduction of Social Security Numbers in Greece". 2009. Page 2.

Erika McCallister, Tim Grance, Karen Scarfone, "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)". Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg. April 2010.

Erma Barron and Felix Bamberger, "Meaning of the Social Security Number". Division of OASDI Statistics, Office of Research and Statistics, Office of Policy, Social Security Administration. Social Security Bulletin, Vol. 45, No. 11. November 1982. Page 1.

David Harley, "Social Security Numbers: Identification is Not Authentication". Social Security Numbers White Paper. August 2009. Page 5.

"Fact Sheet" A U.S. Department of Homeland Security bulletin available at www.socialsecurity.gov. May 15, 2006. Page1.

"Fact Sheet 113: Changing a Social Security Number". An article of Identity theft Center, US. 2010. Page 2. www.protectmyid.com 2013

John G. Morgan "Safeguarding Social Security Number in Tennessee Government Records" A report of Offices of Research and Education Accountability. Tennessee Comptroller of the Treasury. October 2008. Page 22.

Kathleen S. Swendiman, "The Social Security Number: Legal Developments Affecting Its Collection, Disclosure, and Confidentiality". A report by Congressional Research Service. February 21, 2008. Page 8.

Kristin Finklea."Identity Theft: Trends and Issues", A report by Congressional Research Service. January 16, 2014. Page 1.

"Security Breach Notification Laws: Views from Chief Security Officers". A Study Conducted for the Samuelson Law, Technology & Public Policy Clinic, University of California-Berkeley School of Law. December 2007. Pages 3, 9.

"Social Security Number (SSN) Application Instructions For International Students, Scholars and Visitors". (2014) A publication of US Social Security Administration. Page1.

"Social Security: Understanding the Benefits" Social Security Administration SSA Publication (2014) No. 05-10024ICN 454930Unit of Issue - Package of 25. Page 4.

"The Use of Social Security Numbers in California Colleges and Universities" The California College and University Social Security Number Task Force. July 1, 2010. Page7.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.85-96

Students' Academic Performance Modeling and Prediction: A Fuzzy Based Approach

ETUK, S. O., OYEFOLAHAN, I. O., ZUBAIR, H. A., BABAKANO, F. J. & BIMA, M.E.
Federal University of Technology, Minna, Nigeria

ABSTRACT. In higher Institutions of learning, importance is placed on the quality of students admitted as this has direct effect on the quality of graduates been produced by the Institutions, and thus affect the National man-power quality at large. One of the challenges facing the Universities is admitting students on merits and surprisingly, the academic performance of the students admitted on merit begins to drop. Therefore, it is important to predict students' academic performance early enough so as to help instructors take appropriate action in adjusting teaching style and improve greatly on Students' success. In this paper, a fuzzy logic model is used to model data of students and predict their academic performance. Factors like students Ordinary level(O' level) grades, motivation to study in their given course and parents' academic background were used to predict students' academic success level prior to the end of their first academic session. The results when compared with the actual result for the semester examination show 75% accuracy. This early academic performance prediction serves as a guide to the instructor. A good understanding of the students help the instructors to take appropriate steps for effective teaching and learning, and thus improving students' academic performance. Hence, the percentage of students withdrawn from the University after their first academic session due to poor performance (cumulative grade point average of CGPA below 1.5) is expected to be reduced.

Keywords: Student performance, Fuzzy logic, Predictive model

1. Introduction

Education, no doubt is the bedrock on which the development of a Nation lies. A society with quality graduates may bring about higher rates of innovations and faster acquisition of new technology capabilities. Thus, according to (Oladokun, *et al,* 2008), the admission process into any higher institution of learning aims at admitting candidates whose performance would be satisfactory academically in the University. The quality of candidates admitted into any higher institution has a direct impact on the quality of research within the institution and thus has an overall impact on the Nation's development (Oladokun *et al,* 2008). Recently, the decline in the performance of students at the end of their year one in the University is alarming despite that students were all admitted base on merits (Arora and Saini,

2013). Evaluating students' academic performances or achievement using the cumulative grade point average (CGPA) as a pointer is a typical practice in every tertiary academic environment. A percentage of fresh students usually find themselves below the minimum grade point required at the end of their first year (Adedeji, 2001).

This decline can be attributed to several factors affecting effective teaching-learning system amongst which are non- conducive environment of learning and students' lack of motivation to learn (which can be tied to the students' lack of interest in the offered course of study), amongst others. Clearly, the factors considered for admission process is not enough to give teachers a right idea about the level of academic capability of the students they are meant to teach. A very high standard of instruction may not suit a class of more "below-the-average" students while on the other hand, a reasonably lower standard of instruction may be perceived as boring to a class of "more- intelligent-students".

Teachers are considered to be most responsible observers who not only engage class but also monitor the behavior and understanding of students (Arora *et al,* 2013). Hence, if an instructor can properly assess and predict student performance early enough or half way into their first session; then the instructor can take appropriate action to greatly improve the teaching-learning processes geared at improving students' performances. Due to the probabilistic nature of predicting students' performances, fuzzy-based models have been found very effective due to their capabilities to account for fuzzy measures.

In this study, a soft computing approach (fuzzy model)was used for predicting academic performance of students (considering factors like students' O'level grades, UME score, parents' academic background and motivation to learn). The fuzzy model developed was tested using the year one student of the department of Information and Media Technology, Federal University of Technology, Minna as participants.

1. Background of study

A. Student Performance

Academic Background. Students who perform very well in a particular field in an Institution of learning must have had the basic knowledge of the field from their previous learning experiences. In Nigeria, students are grouped into classes of Sciences, Social Sciences and Art while in secondary school based on their performance in related subjects. Consequently, this might translate into the field to which students will be admitted for tertiary education. A student without prior knowledge of the basis of a field or with a weak grade score in the O' level subjects will likely find it challenging to cope with such a field in the University. It is therefore of paramount importance to this research work to use students' academic background as one of the factors in predicting the performance of students in the University.

In previous study, (Adedeji, 2001) used a correlation and regression analysis to investigate the relationship between students' UME scores and their academic performance in the University but due to the nature of UME test as a purely objective

based examination, a combination of Ordinary level academic performance and UME score will provide a better representation of students' prior knowledge.

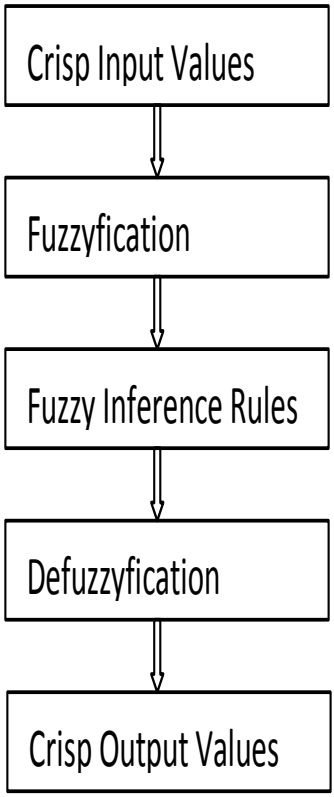
B. Motivation to learn

A driving force to succeed cannot be removed from a success story of any scholar. This also has a very large role to play in students' academic performance. This motivation to succeed academically could be based on factors like interest in a particular course or subject, having a scholar as a role model and wanting to be same, not wanting to disappoint someone who has been part of the students' academic growth, responsibilities of student to alleviate poverty in the family etc. With the right motivation and zeal to succeed, great effort will be put into learning and directly improves academic performance.

C. Fuzzy Logic

Fuzzy logic had long being used to construct better models of reality. Its advantage lies in its ability to provide foundations for approximate reasoning using imprecise propositions base on some fuzzy inference rules. As with sets, fuzzy rules of inference were devised a few decades ago, based on the much older crisp rules. Fuzzy logic shows that truth itself is fuzzy. Rules of inference are rules for deriving truths from stated or proven truths and thus, fuzzy logic are known as efficient tools used to overcome uncertainties related to vagueness, ignorance and imprecision (Yusof *et al,* 2009). Fuzzy Logic involves three main stages namely; fuzzification, rule evaluation and defuzzification. Fuzzification is the process of translating the measured numerical values into fuzzy linguistic values. It is a stage where the degree of membership is determined by applying membership function. Rule evaluation is where knowledge provided by experts is formed, which is then called fuzzy rules (Yusof *et al,* 2009). The fuzzy inference rule will output a fuzzy result, described in terms of degrees of membership of the fuzzy sets.
Defuzzification interprets the membership degrees in the fuzzy sets into a specific action or real-value. This is illustrated in Figure.1

Fuzzy logic has been applied to all fields of life like health to predict cancer, to environment to predict flood detection, and of course, to education to predict students' academic performance.



2. Related Works

A number of works have been done by researchers in an attempt to predict Universities students' academic performance before graduation. This is aimed at having an academic performance assessment of the students to be tutored in order to categorize students and pay more attention on the "very good" students while devising means of supporting the "not-so-good" students.

Existing methods for predicting students' performance include statistical methods. In the case study of (Golding and McNamarah, 2005), they used stepwise multiple regression analysis to predict how factors like students' demographic attribute, qualifications on entry, aptitude test score etc affect the students overall performance. They conclude by suggesting base on their predictions that students with satisfactory predictive performance be allowed to continue their registered program while those with lower performance be channeled to another related program (Golding and McNamarah, 2005).

Other methods researchers used for academic performance prediction is the Data mining algorithms like the decision tree. The work of (Kabakchieva, 2013) used students' personal and pre-University characteristics like gender, birth year,

place of living, total score from previous education etc to predict and classify students into 5 classes which are Excellent, Very good, good, average and bad. They obtained 66.3% accuracy.

Artificial Neural Network (ANN) is another method. (Oladokun *et al,* 2008) developed an Artificial Neural Network (ANN) model considering various factors like ordinary level subjects' scores and subjects' combination, matriculation examination scores, age on admission, parental background, types and location of secondary school attended and gender, among others, to predict the likely performance of a candidate being considered for admission in the University and he achieved an accuracy of 74%. These methods leave a question unanswered: How does it deal with environmental changes and vagueness of reality? This unanswered question lead researchers to using fuzzy based models to predict students' academic performance due to its ability to accommodate uncertainties related to vagueness and imprecision (Osman *et al* 2009) and (Osman *et al,* 2012). In (Yildiz *et al,* 2014), a fuzzy RFM-Model was developed to predict distant learning students' performance. They considered factors like recency, laying importance on the length of time taken before a registered student is admitted on the system; frequency, stating how often an admitted student log on to the system and monetary, considering the length and period of time spent online on the system. The rules of the fuzzy system have been according to expert opinions and the prediction accuracy was 74.7%. The question is; would this model built for distant learning platform generalize to a classroom learning environment?

3. Methodology

The Fuzzy Logic Process

A. Crisp Input Values

In this study, data of year one students of the department of Information and Media Technology, 2015/2016 session were used. Due to the preliminary nature of the study, 36 students in year one were randomly sampled. Out of the 36 questionnaire distributed, 28 were returned, but only 20 responses were valid. This represents 55.60% valid responses. The data for this model was gotten from the students' files and response to the questionnaire on their parents' academic background and what motivates the students to undertake learning in their present course of study. The data were rated "below average", "average" and "good" and a classical fuzzy model was used to predict students' class of degree by predicting their CGPA based on experts opinion.

The crisp input values are the students' data gotten from the students' record and their response to the questionnaire. The process of transforming or normalizing this input values are shown in Table 1.

1. O' level result
2. Motivation to learn
3. Parent's literacy level

S/N	Input Variable		Score	
1	O'level Results	Mathematics	A B C	Good Average "Below average"
		English	A B C	Good Average "Below average"
		Physics	A B C	Good Average "Below average"
		Chemistry / Biology	A B C	Good Average "Below average"
		Geography / Agric science or Economics	A B C	Good Average "Below average"
2	Course Applied for in UME	Present Course Related Not Related	5 – 4 3 1-2	Good Average "Below average"
3	Parent's academic background	Illiterate – pri sch Secondary sch Tertiary Sch	5 – 4 3 1-2	Good Average "Below average"

Table 1: Input Data Transformation

B. Fuzzy Inference Rules

Fuzzy inference rules of the model were created to predict student's academic performance based on expert opinion using O' Level Result, student motivation to undertake their course of study, and their parents Literacy level.

1. If (Result is "below average") and (Motivation is "below average") and (Parent's Literacy is "below average") then (Performance is "below average")
2. If (Result is "below average") and (Motivation is "below average") and (Parent's Literacy is average) then (Performance is "below average")

3. If (Result is “below average”) and (Motivation is “below average”) and (Parent’s Literacy is Good) then (Performance is “below average”)
4. If (Result is “below average”) and (Motivation is average) and (Parent’s Literacy is “below average”) then (Performance is average)
5. If (Result is “below average”) and (Motivation is average) and (Parent’s Literacy is average) then (Performance is average)
6. If (Result is “below average”) and (Motivation is average) and (Parent’s Literacy is good) then (Performance is average)
7. If (Result is “below average”) and (Motivation is good) and (Parent’s Literacy is “below average”) then (Performance is average)
8. If (Result is “below average”) and (Motivation is good) and (Parent’s Literacy is average) then (Performance is good)
9. If (Result is “below average”) and (Motivation is good) and (Parent’s Literacy is good) then (Performance is good)
10. If (Result is average) and (Motivation is “below average”) and (Parent’s Literacy is “below average”) then (Performance is “below average”)
11. If (Result is average) and (Motivation is “below average”) and (Parent’s Literacy is average) then (Performance is average)
12. If (Result is average) and (Motivation is “below average”) and (Parent’s Literacy is good) then (Performance is average)
13. If (Result is average) and (Motivation is average) and (Parent’s Literacy is “below average”) then (Performance is average)
14. If (Result is average) and (Motivation is average) and (Parent’s Literacy is average) then (Performance is average)
15. If (Result is average) and (Motivation is average) and (Parent’s Literacy is good) then (Performance is good)
16. If (Result is average) and (Motivation is good) and (Parent’s Literacy is “below average”) then (Performance is average)
17. If (Result is average) and (Motivation is good) and (Parent’s Literacy is average) then (Performance is good)
18. If (Result is average) and (Motivation is good) and (Parent’s Literacy is good) then (Performance is good)
19. If (Result is good) and (Motivation is “below average”) and (Parent’s Literacy is “below average”) then (Performance is “below average”)
20. If (Result is good) and (Motivation is “below average”) and (Parent’s Literacy is average) then (Performance is average)
21. If (Result is good) and (Motivation is “below average”) and (Parent’s Literacy is good) then (Performance is average)
22. If (Result is good) and (Motivation is average) and (Parent’s Literacy is “below average”) then (Performance is average)
23. If (Result is good) and (Motivation is average) and (Parent’s Literacy is average) then (Performance is average)
24. If (Result is good) and (Motivation is average) and (Parent’s Literacy is good) then (Performance is good)
25. If (Result is good) and (Motivation is good) and (Parent’s Literacy is “below average”) then (Performance is good)

26. If (Result is good) and (Motivation is good) and (Parent's Literacy is average) then (Performance is good)
27. If (Result is good) and (Motivation is good) and (Parent's Literacy is good) then (Performance is good)

C. Crisp Output Values

The Output values illustrated in Table 2 represents the students' Academic Performance based on the University's classification of degrees into '1st Class', '2nd Class Upper', '2nd Class Lower' and '3rd Class'.

S/N	Output Variable	Class	CGPA
1	Good	1st Class – 2nd Class Upper	4.5 – 5.0 3.5 – 4.49
2	Average	2nd Class Lower	2.5 – 3.49
3	"below average"	3rd Class	1.5 -2.49

Table 2: Output Data Transformation

D. The fuzzy based Model

The model below was developed base on students' records to determine their academic performance.

$$\frac{J}{JA_{ve}S} + 0.6\,(Wol1 + Wol2 + \cdots + Wol5) + 0.1(d.f)$$

Where:
J: Jamb score
$JA_{ve}S$: Average Jamb Score.
Wol: Weight of O' Level Result (minmum 5 credit passed courses).
d.f: Dynamic factor.

To account for the dynamic factors, certain questions were asked from the student.

4. Results and Discussion

The Model was tested using feedback from year one students of the department of Information and Media Technology based on the valid responses received. The

responses were used to predict students' CGPA and compared with their actual CGPA as shown in figure 2 below.

The model was 75% accurate in predicting the students' likely class of degree by predicting their CGPA with absolute error displayed in Table 3. This accuracy is of fair performance when compared with existing work in literature of (Oladokun *et al,* 2008) that had 74% accuracy in predicting students' performance.

S/N	Actual CGPA	Predicted CGPA	Absolute Error
1	2.41	2.53	-0.12
2	2.5	3.03	-0.53
3	2.95	2.65	0.3
4	1.5	2.49	-0.99
5	1.05	2.01	-0.96
6	2.77	2.62	0.15
7	2.50	2.69	-0.19
8	1.95	1.7	0.25
9	1.73	2.7	-0.97
10	2.5	2.58	-0.08
11	3.82	3.24	0.58
12	2.64	2.73	-0.09
13	1.55	2.4	-0.85
14	2.91	2.87	0.04
15	3.18	3.17	0.01
16	3.27	3.28	-0.01
17	3.77	3.07	0.7
18	2.05	2.41	-0.36
19	2.77	2.61	0.16
20	2.14	2.43	-0.29

Table 3: Result findings

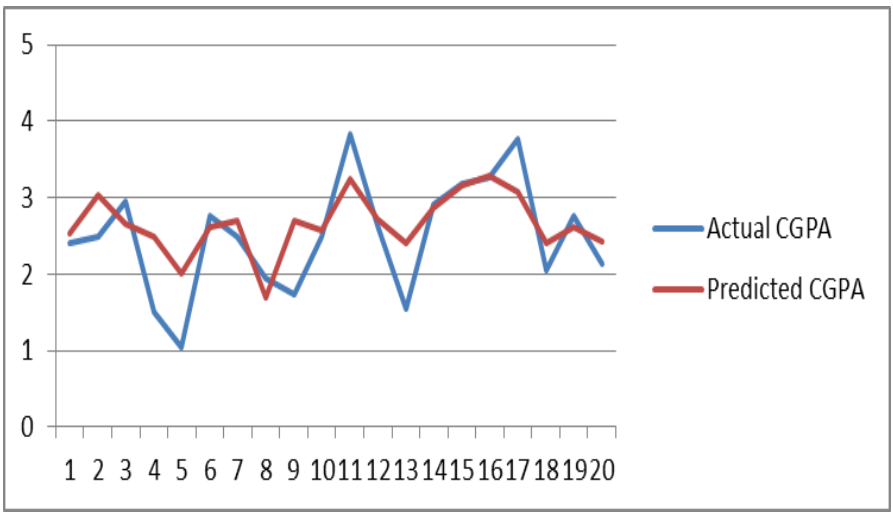


Figure 2: Testing of the Model

Implications of Findings

The result from this study helps the Universities and Instructors in particular to better understand the academic capability of each and every student early enough and adjust the teaching methods so as to better prepare the students to survive the academic pressures. It can be deduced that students whose actual performance is lower than the predicted performance need a special attention in order to live up to the academic requirements. If necessary actions are taken to improve teaching and learning, this will motivates the students to do better. This will have direct positive effect on the percentage of students being withdrawn from the University after their first session for poor academic performance of below 1.5 CGPA. A fuzzy based model was used because of its ability to account for uncertainties and it's solely based on expert opinions. The CGPA predicted are the lower-bound of the degree of class which means the students can achieve the predicted CGPA or higher which is actually fuzzy in nature.

Conclusion

In conclusion, a fuzzy-based mathematical model was developed to predict the academic performance of students'. The test of the model was performed on year one students because it is of paramount importance to begin to understand the students' academic capability from their year one so that all effort can be put in place to assist each students achieve his academic potentials before graduation. The result from the model when compared with the students' 1st semester result shows 75% accuracy and absolute error less than 1.0 in each case. This handy information to the instructors as early as mid session of year one still gives the instructors time to decide on the best teaching techniques for the students. This will improve teach-

ing and learning before the end of the session when students with below 1.5 CGPA are withdrawn.

Limitation of Study

The major limitation of this study is the students' unwillingness to answer and submit the questionnaire.

Future Work

This research study is on-going and an oral interview of students is recommended so as to tackle the limitation of questionnaire and this may likely be an improvement over this model Also, this model can be extended to other department and faculty of the University.

Correspondence
Etuk, S. O.
School of Information and Communication Technology
Federal University of Technology, Minna, Nigeria
Email: abiolastella@futminna.edu.ng.

References

Adedeji, O.B. (2001). "A study of the relationship between students UME results and their undergraduate performance". unpublished

Arora, N., and Saini, J. R. (2013). "A fuzzy probabilistic neural network for students' academic performance prediction". International Journal of Innovative Research in Science, Engineering and Technology, vol 2(9), pp. 4425-4432.

Arora, N., and Saini, J. R. (2013). "Predicting students academic performance using fuzzy ARTMAP network". International Journal of Advances in Engineering Science and Technolody, vol 3(3), pp. 187-192

Golding, P. and McNamarah, S., (2005). "Predicting Academic Performance in the School of Computing & Information Technology (SCIT),‖ Proceedings of 35th ASEE /IEEE Frontiers in Education Conference.

Kabakchieva, D. (2013). "Predicting Student Performance by Using Data Mining Methods for Classification, Cybernetics and Information Technologies, vol. 13, No. 1, pp. 61-72, 2013.

Oladokun, V.O., A.T. Adebanjo, and O.E. Charles-Owaba. (2008). "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course". Pacific Journal of Science and Technology. 9(1):72-79.

Osman T. and Bahattin K., An adaptive neuro-fuzzy model for prediction of student's academic performance, Computers& Industrial Engineering, No. 57, pp. 732 –741, 2009.

Osman, Y., Abdullah, B., Sevinc, G. and Fulya, D. K. (2012). "A genetic-fuzzy based mathematical model to evaluate the distance education students' academic performance", Procedia-Social and Behavioural Science, vol 55, pp. 409-418

Yildiz, O., Gulsecen, B. And Fulya, D. (2014). "Rules optimization based fuzzy model for predicting distance education students' grades". International Journal of Information and Education Technology, vol 4(1), pp. 369 – 372

Yusof, N., Zin, N. A., Yassin, N. M., and Samsari. (2009). "Evaluation of students' academic performance and learning efficiency based on ANFIS." International Conference of Soft Computing and Pattern Recognition. Pp. 460-465. IEEE Computer Society.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.98-108

A Framework for Unified Distributed System for Crime Prevention and Detection (UDSCPD)

PETER E. AYEMHOLAN
National Defence College, Abuja, Nigeria

GARBA SULEIMAN
FCT College of Education, Zuba-Abuja, Nigeria

OSAIGBOVO TIMOTHY
Aduvie International School, Abuja, Nigeria

ABSTRACT In recent times, the economic recession in most countries of the world has led to a drastic upsurge in crime rate and other related vices. The rise in crime and the frequency with which they are perpetrated has left the relevant security agencies with the challenge of proactively responding and preventing crime incidents. Developing nations, for instance, are the worst hit by the rising wave of criminal activities. In these nations, the traditional methods of crime reporting, which are usually centralized in nature, have proven to be very slow and incapable of delivering a good response time required to detect and prevent crime. This study attempts to develop a framework for the implementation of a crime detection and prevention system. The system presents a unified distributed architecture, allowing public users to provide tip-off on crime incidents. Multiple data collected from different sources are warehoused in a datacentre. It is redistributed to other participating datacentres for data analytics and backup processes aimed at ensuring reliability and availability. The system offers a web interface to facilitate quick gathering and escalation of crime incidents data in form of text messages, video and images using smart mobile devices. The data are processed to derive useful information which is further transmitted to relevant authorities to enable them mobilize into action. The proposed system would be of immense benefits to the relevant agencies saddled with the responsibilities of curbing crimes in Nigeria. It would not only assist these agencies in identifying localities which are highly vulnerable to crime, but will also largely reduce the Response time (RT) to crime situations with a view to preventing any possible future occurrence. This system will simplify ways of crime reporting by the public through the web interface. It also provides crime prevention and detection mechanism for the security agencies through the profiled crime database at the datacentre; thereby assisting in identifying localities which are vulnerable to crime and reduce response time to crime situations.

Keywords: Crime Detection, Crime Prevention, Distributed System, Datacentre, Cloud Computing

1. Introduction

The growth and development of a nation is adjudged largely by her capacity to guaranty the safety of lives and property within her territorial integrity. According to Plant and Michael (2009), the safety of lives and security of property are viewed as basic human rights which essential to the community's overall quality of life. When the citizenry is not, and does not feel reasonably safe as a result of security threats, other critical government functions such as economic development, government finance, public education, stable housing, and other basic services become much more difficult to provide. Crime prevention refers to efforts to prevent crime or criminal offending in the first instance – before the act has been committed (Welsh and Farrington, 2010). Various nations of the world today are taking drastic measures to curtail the rising wave of crime in their society, in fulfillment of their statutory responsibilities to provide safe and stable environment for the enhancement of social and economic interactions. Some of the notable efforts are the setting up of agencies with the mandate to deal with crime incidents and the deployment of technology for effective crime management. Welsh et al (2010) identified the profound influence technological advances over the years have made on the way we think about crime and the efforts that are taken to prevent it. Since the fall of 21st century, technology has become the driving force leading to reform of crime prevention and crime control strategies, both by individual citizens and concerned groups, and by formal police agencies (Reichert, 2001; Chan, 2001; Harris, 2007).

According to a recent review of police technology by Harris (2007), the first technology revolution in the United States that changed the way police were organized and how they operated centred around three technological innovations that were incorporated into policing: the telephone, the two-way radio, and the automobile. With the proliferation of telephones in the early twentieth century, policing changed. Citizens were encouraged to call the police to deal with a multitude of problems, and the police responded to those calls from dispatch through a two-way radio, and sped quickly to locations using patrol cars. The advents of Internet and social media platforms have also offered new opportunities for easy crime reporting, enabling security agencies to quickly respond to crime situations.

Following the economic recession facing most countries of the world today, there is now a drastic upsurge in crime rate and other related vices. The rise in crime and the frequency with which they are perpetrated has left the relevant law enforcement agencies with the challenge of proactively responding and preventing crime incidents. Developing nations, for instance, are the worst hit by the rising wave of criminal activities. In Nigeria, for instance, the traditional methods of crime reporting, requires the public to provide tip-off on crime through phone calls or visit the closest police posts to report crime incidents.

Literature Review

In order to provide insight to the proposed system, it is important to review similar solutions offered by other authors with a view to identifying their advantages and

shortcomings. First, is the work of Oludele et al (2015), where the author asserted that the present world is technologically driven as it is employed by many areas of profession in carrying out their operations. In view of this the researcher designed Crime Record Management System CRPMS for law enforcement agency (NPF) to keep records of crime and criminals. This Crime Record Management System CRPMS was designed to replace the manual method which involved the use of pen and paper. This CRPMS will enhance proper and efficient management of criminal records by the NPF which will help in taking decisions and ensure reliability of their operations.

Since security issues have been a major issue of governments and nations in general, there has been a public outcry by the government and the public in finding solution to this menace of crime rate. To this end, Ogunleye, et al (2011) presented a computer based security framework for crime prevention in Nigeria whereby it shows how CCTV system can be used in operational, decisions, performance stands and provision of evidence. It also shows its usage for incidence reduction or post incidence analysis which acts as a deterrent or provision of valuable support to the security agencies. The usage of CCTV and its importance in crime prevention and control was corroborated by Jimoh et al (2013). The study shows that with the use of CCTV and introduction of solar power system as an alternative power source helps to reduce the cost incurred while installing such CCTV system in different locations.

Jong-moon (2011), on the other hand, asserted that due to the increased usage of printable devices and the spread of wireless networks in areas where physical access cannot be reached, the researcher looked at the signature–based detection with smart wireless and IP tables using wireless attack detection and prevention measures to implement a wireless intrusion prevention system in order to exploit crime and damage. Adigun (2013) correlates the residents response to crime in Nigeria cities by examining the socio-economic attributes of residence building and environmental features, residence crime magnitude, fear of crime events, fear of neighborhood and household, safety measure in some selected cities in order to establish a relationship between them, which shows that meaningful intervention of crime and its control must first begin with detection on building and environment features that discourages crime incidence and thereby reduces fear of crime.

Canter (1997), explored the use and possibilities of GIS by Baltimore County Police in describing and analyzing crime activity. He noted that as this technology gains greater acceptance and use within police departments, it will become clear that the ability to produce automated pin maps is only one of many possible applications. He concluded that GIS should be viewed as a tool for which police analysts could obtain a better understanding of criminal activity from a geographic perspective.

Kawai and Samson (2011) developed a user-friendly system, which enabled an efficient exchange of information on previous convictions of criminals. Consequently, Information about previous convictions is circulated between judges and prosecutors as well as police authorities. This is vital in order to provide adequate responses to crime and also to prevent new crimes from being committed.

Ahmed and Salihu (2013) examined the use of Geographic Information Systems (GIS) and spatial database of crime characteristics which helped in the determination of crime prone spots in Dala LGA of Kano State. The study specifically revealed the challenges facing police departments that sought to adopt computerized crime mapping systems. The researchers used ArcGIS version 9.3 to analyze different data sources obtained from the Nigerian Police Force. The result showed that crime rate outside the city wall was higher, while the rate increased with time particularly in the year 2010.

Singh (2014), in his work, "Mobile Application and Public Safety System for Crime Tip-Off and crime prevention by communities" took a step further to identify that the usage of mobile phones is rapidly increasing from basic SMS applications, to snapping photos and sending via multimedia applications, using social media application, and using specific smart phone applications intended for community needs. He therefore proposed a public safety system which would be on a central server enabling chatting with public safety officers, filling in the tip-off form, submitting via the mobile application and also uploading photos relevant to the tip-off. The public safety was intended to: store, route the tipoff to the officer(s), track the jobs assigned to the officer(s), update the community user who had provided the tip-off, update and report to relevant authorities.

This study, however, revealed that most of the previous solutions are centralized and sometimes, not real-time in nature. Additionally, they are not enhanced with the capabilities to allow a witness at a crime scene to instantly provide a tip-off through SMS, voice and video calls, social media platforms and also captures the crime location using Global Positioning System enabled smart mobile devices to the appropriate authorities for necessary actions. Although Singh (2007) had well addressed most of the identified shortcomings, the solution did not adequately cater for failure of the central server by providing redundant datacentres which would become active in case of failure of the central server. This study, therefore, attempts to develop a framework for the implementation of a unified distributed system for crime prevention and detection. Aspnes (2006) identified distributed system as one comprising of a collection of autonomous computers linked by a network and software, enabling computers to coordinate their activities and share the resources of the system so that users perceive the system as a single, integrated computing facility.

The distributed architecture of the UDSCPD, as proposed by the researchers, would allow crime data to be warehoused and processed by distributing it among participating datacentres. The purpose is to fast track the processing of crime data thereby reducing the amount of time required by the law enforcement agency to respond to crime incident. The system would offer a web interface to facilitate quick gathering and escalation of crime information in form of texts, images or video formats. Big Data analytic procedures are applied to the data to derive a useful pattern/trend which is forwarded to relevant authorities to enable them mobilize into action. In order to guide against loss of crime information, the proposed system would cater for data redistribution to all participating datacentres to ensure reliability and availability.

3.0 Methodology

The proposed method identified four (4) major activities in achieving the research aim. According to Erhan (2014), Information systems and artificial intelligence technology is a new alternative to manual information processing or computer aided information processing. It offers the advantage of delivering an optimized result obtained quickly by a deterministic approach. The proposed method presents the researchers framework for achieving an efficient and real time crime prevention and detection techniques. Fig. 1 represents a physical working scenario showing the communication established in actualizing the crime prevention and detection functionality. The public users uses a smart mobile devices or Internet enabled personal computers to access the web server which is linked to the databases at the datacentres of each security agencies where all records of any reported crime or actions concluded on such crime are profiled.

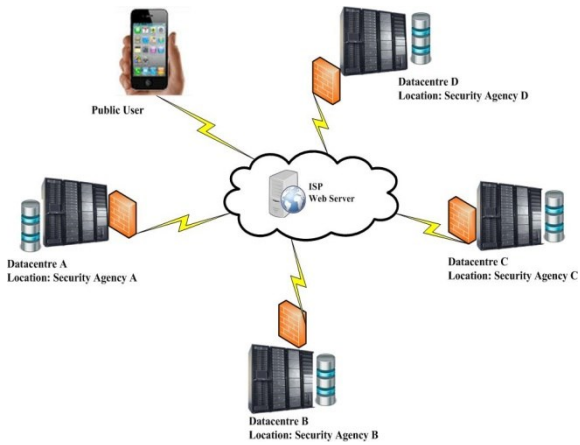


Figure 1: Physical components of proposed system

3.1. Public Interface

The public interface consists of the Public User and the Web Server. The primary aim of this unit is to provide a medium for the general public to report any crime committed or provide a tip-off on crime incidents within their vicinity to the security agency for profiling and necessary action. The web server, which is hosted in the cloud, provides a web interface (website) with which the users can interact and fill out online tip-off form and also download mobile phone application for android and IOS. The interface allows users to capture and upload additional information such as images, voice and video that could assist the security agency in tracking down the criminals. A public user who is using his/her smart mobile phone to provide a tip-off is also required to turn on the GPS facility in the phone to allow instant capturing and transmission of the location from which the crime is being re-

ported. The responsibility to report crime to the Crime Detection and Prevention System (CDPS) is performed by both public user and security agents who may be present at the crime scene. Apart from allowing public users to provide information on crime incidents, the web application, which is hosted on the web server, helps the security agencies to have an overview of crime and criminal records and activities as reported by the public users and for every action performed by each of the security agencies. The web server also provides functionalities of separating crime to each agencies that are tasked to prevent, control or provide immediate response to such incidence.

3.2. Security Agency

Security Agency is any organization authorized to deal with crime situation with a view to preventing any further occurrence. In this work, some of the recognized agencies include the NPF, DSS, NSCDC and other agencies linked to the system to provide security services. Their function is to receive tip-off on any reported crime incidents in form of SMS notification through their smart mobile devices. They swiftly mobile the concerned unit/s to deal with the situation, report action performed on the reported/detected crime and profiles such cases for record purpose. In addition, the security agency shares real-time information on crime incidents with sister agencies for prompt action.

3.3. Datacentre

The web application establishes a connection in such a way that every reported crime is profiled into the database hosted in the active (main) datacentre and as well updates other (redundant) datacentres accordingly. The datacentre serves as the warehouse for all crime data collected as well as provides the facilities for data analytics. In other word, datacentre provides facilities for storage and processing of crime data. Each of the security maintains a datacentre which is connected to the datacentres of other security agencies for data sharing and to support distributed processing. It is designed with the functionalities to enable each security agency to have all records/information of crime and criminal existing in a locality. All datacentres are interconnected by VPN for real-time sharing of information and update of crime databases in each datacentre. It allows other datacentres to have access to that information to update their database. If there is any break down or loss of communication from the active datacentre, other datacentres may be consulted for update, thereby providing backup service to the active datacentre.

3.4. Crime Database

The Crime Database comprises of a cluster of databases of similar attributes stored in multiple datacentres. This is where data obtained from public users via the web server are stored.

A. Structure of the Crime Database. In this study, the crime database is structured into 4 databases and hosted in 4 Datacentres identified as Datacentre (A – D). They are hosted at the locations of 4 security agencies also identified as Security Agency (A-D) as shown in fig 4. One of the Databases is selected as the active Database by the Web server for its data transactions. The web server selects any of the databases as the active database using certain criteria such as congestion, bandwidth and availability. When the active database becomes unavailable, one of the remainder 3 databases is promoted to active database. At a certain time interval, the active database ensures that the contents of the other 3 databases are updated.

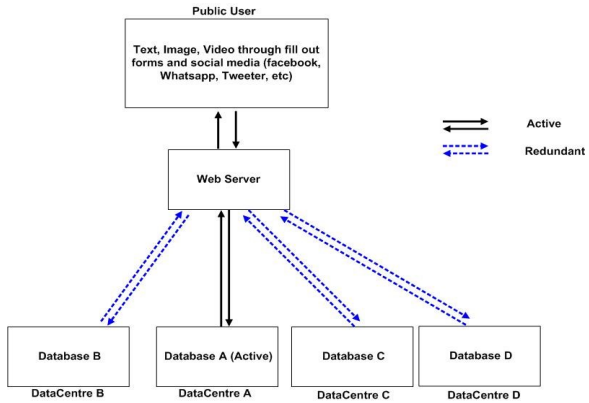


Fig 2: Web server -Database Interaction

B. Crime Data Analytics. Crime data collected from public users which are both structured and unstructured, are voluminous in nature. The data, therefore, need to be processed with a view to identifying patterns such as trend, frequency of occurrence and crime location which could assist security agencies in tracking down the criminals. The process requires the application of big data analytics to support the processing of data across the 4 datacentres as illustrated in figure 3.0. The setup is equipped with the functionalities handle both structured and unstructured terabytes and petabytes of data using Hadoop File System (HDFS) with adequate support for accelerated parallel mechanism.

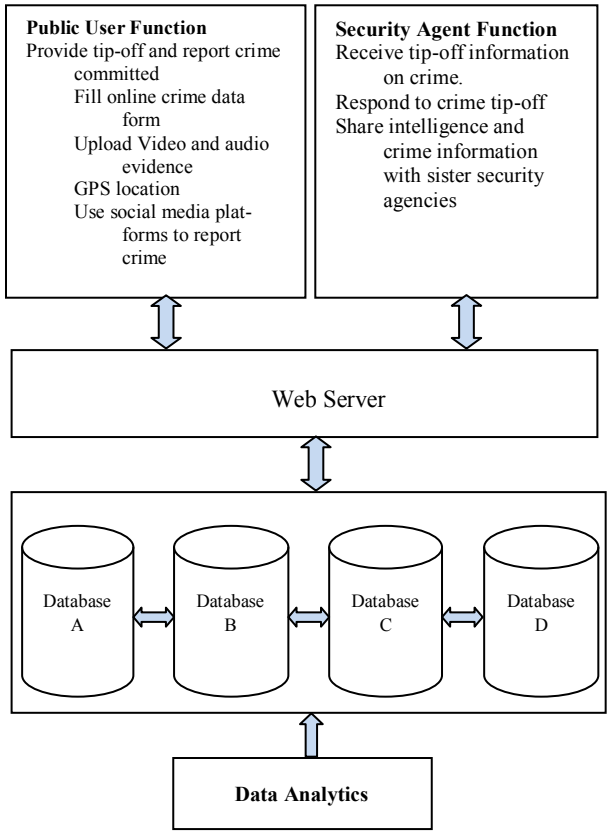


Fig 3: Functional Framework of the UDSCPD

4.0 Resources and Manpower Capacity

The resources implication of this framework consists of an integrated and functional database using a powerful software tools such as c#, Java, Hadoop and MySQL for the database design. In addition a highly sophisticated hardware such as routers, storage, server computers with higher configuration, real time Internet access with high bandwidth need to be made available at each datacentres and the main active database. Also, in order to efficiently and effectively manage the system, qualified manpower need to be employed. Adequate and up-to-date training, workshop and seminars should be provided for the back-end personnel handling the system and the security operatives who are to utilize the data to prevent crime. Incentives such as wages/salary should be given to the personnel to prevent them from compromising the security of the system. Additionally, fund should be made available to maintain the software and hardware of the propose framework. Confidentiality of public user who reports any crime should be protected so as to give user confidence and encourage prompt reporting of any crime incidence.

5.0 Limitations of the study

Our proposed system has the following inherent limitations:

- One of the limitations of the framework is constant supply of power to the system at the datacentres.
- The system requires a high Internet bandwidth for real-time report of crime and prompt response from the security agencies.
- Not all communities are provided with adequate GSM network coverage. This may hamper the quick reporting of crime incidents to the UDSCPD.
- Most public users do not have smart phones that could support access to UDSCPD through social media platforms and provide location based information through GPS.

6.0 Discussion

The system consists of datacentres which house the facilities for data analysis. It also supports the use of mobile applications on smart devices to facilitate quick gathering and escalation of crime incidents data in form of text messages, voice, video and images. A web server, hosted in the cloud, provides an interface through which the user can provide tip-off to the system in form of fill-out online form and social media services. The concerned security agency receives a prompt notification from the web server on the reported crime requesting immediate action. The datacentres on the other hand, serves as the warehouse for all crime data collected such that each centre is connected to the other via VPN for sharing of information and update of crime database in the datacentres. Thus, in the event of failure of the active datacentre, any of the datacentres with optimal path can be upgraded to Active datacentre to provide a backup service to the web server until the failed datacentre for crime database is restored.

7.0 Conclusion

The issue of crime prevention and detection will be more simplified as the paper proposes an implementable framework for a unified distribution system for crime prevention and detection where the system is designed to allow tasks to be distributed among participating devices so as to reduce time and increase efficiency. The system offers a public web interface to enable users provide tip-off on crime incidents. The database profiles the crime data and notifies concerned security agency about the reported crime incident to enable them take appropriate action. Further research study on this subject would consider the development and implementation of the proposed framework for UDSCPD. Text data would be obtained from the NPF, NSCDC and some selected security agencies in Nigeria to evaluate the performance and reliability of the system. Also, further research can be done in terms of database security whereby encryption should be worked upon to enhance data security and integrity.

Correspondence
Peter E. Ayemholan
Department of ICT
National Defence College, Abuja, Nigeria
Email: ayemonlan2000@yahoo.com

Garba Suleiman
Computer Science Department
FCT College of Education, Zuba-Abuja, Nigeria
Email: sulgarba@gmail.com

Osaigbovo Timothy
ICT Unit, Aduvie International School, Abuja, Nigeria
Email:timothyosaigbovo@hotmail.com

References

Ahmed, M. & Salihu, R. S. (2013). Spatiotemporal Pattern of Crime Using Geographic Information System (GIS) Approach in Dala L.G.A of Kano State, Nigeria, American Journal of Engineering Research (AJER), Volume-2, Issue-3, pp-51-58

Anderson, R. (2001). Security Engineering, a Guide to Building Dependable Distributed Systems, John Weley & Sons Inc, New York City, 2nd Ed, pg 185-213.

Aspnes, J. (2006). Notes on Theory of Distributed Systems, CPSC 465/565: Spring, retrieved from http://cs-www.cs.yale.edu/homes/aspnes / classes/465/notes.pdf on 14 Aug 2016.

Byrne, J.M and Marx G.T. (2011), Technological Innovations in Crime Prevention and Policing.
A Review of the Research on Implementation and Impact, Maklu-Uitgevers, nr. 20, p. 17-40.

Canter, P.R (1997). Geographic Information Systems and Crime Analysis in Baltimore County, Maryland, Baltimore County Police Department's Crime Analysis Unit, retrieved from http://www.popcenter.org/library/ crimeprevention/volume_08/06-canter.pdf on 15 Sep 2016.

Chan, J. (2001). "The Technology Game: How
Information Technology is Transforming Police Practice" Journal of Criminal Justice, 1:139-159.

Erhan, B. P. and Muhammed, V. A. (2014). Information Systems and Artificial Intelligence
Technology Applied in Engineering Problems. Paper presented at International Conference on Educational Technology with Information Technology (ICETIS 2014), Bangkok, Thailand – November 8-9, 2014. Proceed-

ings are available *on International Journal of Information Technology & Computer Science (IJITCS)* . 17 (2).

Harris, C. (2007) "Police and Soft Technology: How Information Technology Contributes to
Police Decision Making" In: Byrne, J. and Rebovich, D.(2007). The New technology of Crime, Law and Social Control, Monsey, NY: Criminal Justice Press, p. 153-183.

Jimoh., N.B. & Abdulazeez, S. A. (2013). A Computer-Based Security System for Crime
Prevention and Control in Kaduna Metropolis. Research Journal of Information Technology. 5(4), 104-108

Jong-Moon, K., A-Yong, K., Jung-Soo, Y. & Hoe-Kyung, J. (2015): A Study on Wireless
Intrusion Prevention System based on Snort. International Journal of Software Engineering and Its Applications. 9(2). 1-12

Ogunleye, G.O, Adewale, O.S. Alese, B.K. & Ogunde, A.O. (2011). A Computer Based Security Framework fro Crime Prevention in Nigeria. Proceeding of Nigeria Computer Society (NCS) 10th International Conference July, 2011.

Oludele A., Onuiri, E.E., Olaore, O.A, Sowunmi O.O. and Ugo-Ezeaba, A.A. (2015). A Real-Time Crime Records management System for National Security Agencies. European Journal of Computer Science and Information Technology. 3(2), 1-2

Plant, J.B, and Scott, M.S. (2009). Effective Policing and Crime Prevention, A Problem-Oriented Guide for Mayors, City Managers, and County Executives, U.S. Department of Justice Office of Community Oriented Policing Services.

Reichert, K. (2001). "Use of information technology by law enforcement". Promising Approaches to Addressing Crime Series. University of Pennsylvania, Jerry Lee Center of Criminology, Forum on Crime and Justice.

Welsh,B.C, & Farrington, D.P. (2012). The Future of Crime Prevention: Developmental and Situational Strategies, Prepared for National Institute of Justice, Bethesda, Maryland, Retrieved from http://www.crim.cam.ac.uk/people/academic_research/david_farrington/nijprev.pdf on 16 Sep 2016.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.110-123

An Enhanced Conductance-Based Approach for Community Detection in Weighted Mobile Phone Networks

ELIZABETH N. ONWUKA, BALA A. SALIHU & PASCHAL S. IORNENGE
Federal University of Technology, Minna, Nigeria

ABSTRACT Community Detection has gained a lot of attention in recent years due to its applications in studying human behaviour in various spheres of life and most especially in the analysis of criminal networks. In this era of Big Data Analytics, community detection has been made easier by the availability of huge sources of data such as the Call Detail Records (CDR) of the telephone networks. Recently, focus in community detection is gradually drifting from unweighted networks to weighted networks, where the strength of the link between each pair of connected nodes is considered rather than just the existence of a link. However, existing algorithms for community detection have focused only on direct links between pairs of nodes in a network. In this work, an Enhanced Conductance-based Algorithm (ECBA) was develop to detect communities in a network. This was done by synthesizing the direct and indirect relationship strengths between all pairs of nodes on a weighted undirected graph. The algorithm was tested with CDR data using belonging degree and conductance as the decision metrics for community partitioning. Comparison with the original conductance-based algorithm shows significant improvement in quality of detection for communities of large sizes in terms of average shortest path distances, density, and how closely knit the connections are. Test results further show that using indirect relationships between pairs of nodes significantly reveals more information about community membership in large networks

Keywords: community detection; social network graphs; binary networks; weighted networks; conductance; belonging degree.

1. Introduction

The massive improvement in technology over the years, especially the technical and commercial success of the mobile phones and other handheld devices, has made the study of human behaviour or interaction patterns via this medium increasingly useful. Users of mobile phones, either for voice/SMS communication or for online social networks, leave digital traces that can be used to understand their behaviour and connections over time. Even criminal activities and criminal networks can be more easily understood and detected by analysing such data. Among the prominent ways this has been done is community detection.

A network is a group of nodes (or vertices) connected through edges or links. A community in a network is a group of nodes having more internal connections with

each other than external connections with the rest of the network (Fortunato, 2010). They are also called Clusters, Cliques or Cohesive groups (Borgatti, 2009) (Palla et al., 2005). Detecting communities of users on a network has gained significant growth due to applications such as warm containment in Online Social Networks (OSN), data forwarding in delay tolerant networks, routing strategies for MA-NETS, coding, automatic allocation of small LANs (Lu et al., 2013), detecting terrorist or criminal groups, Link prediction, information diffusion, and in biological or medical systems (Ahuja et al., 2016). This is commonly done on social network graphs which are made of nodes (users on the network), and edges which represent links between the users. Social Network graphs may be directed or undirected, weighted or unweighted. In a directed graph, the direction of communication between two nodes is considered, while Undirected graphs are made up of unordered pairs of vertices, i.e., direction of communication is not used in carrying out analyses. An undirected graph is unweighted (or binary) if a single edge connects each pair of vertices. An unweighted graph (or an unweighted version of a graph) is used for analyses when the goal is to simply know which nodes have communication links to each other. In this case there is no interest in the extent of communication in those links. However, for weighted graphs, there can be multiple edges connecting a pair of vertices, highlighting the extent of communication and hence, the relative strengths of the links (Lu et al., 2013).

In this paper, we approach the problem of community detected on weighted and undirected networks. Our main contribution is the development of an Enhanced Conductance-based Algorithm (ECBA) which not only uses direct relationship strengths but also indirect relationship strengths to improve the quality of community detection.

2. Related Works

The goal of community detection is to partition a network into dense regions of the graph. Each region represents a group of nodes that are closely related, and hence are in the same community. Most of the earlier algorithms for community detection were based on binary networks. Very prominent among them is one proposed in (Girvan and Newman, 2002) which focused on the boundaries of communities rather than their core. It is said to be the first algorithm in modern age community detection (Fortunato and Lancichinetti, 2009). In their approach edges are removed from the network based on Betweeness centrality values. The edges with the highest Betweeness centrality are removed, Betweeness is calculated again for the edges affected by this removal, and the process is repeated until no edges remain. However, the run time of the algorithm as the number of nodes increase makes it unsuitable for large graphs. Cfinder was developed to uncover the structure of complex networks by analysing the statistical features of overlapping networks (Palla et al., 2005). A community (a k-clique community) was defined as a union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing $k-1$

nodes) $k-1$. It was based on the fact that members can be reached through well connected subsets of nodes. This approach allowed overlapping in community membership. The community detection was done by setting a threshold weight for the links and ignoring links that were below this threshold weight making the network essentially a binary network. The RAK algorithm which is based on label propagation was also proposed in (Raghavan et al., 2007). In this approach, each node is first initialized to a unique label which represents the community it belongs to, and these labels then propagate through the network. A node would determine its community based on the labels of its neighbours. Each node joins a community which has the most of its neighbours as members and the labels of the nodes are updated at each iteration. As the propagation continues, dense connected groups of nodes finally settle for a unique label, and in the end, all nodes with the same labels are placed in the same community. This continues until each node in the network has the label to which the maximum number of its neighbours belong to. It is however possible for the iteration to end with two disconnected groups of nodes having the same label. It will require a breadth-first search on the subnetwork of each individual group to separate the disjointed communities thus increasing the computation time and complexity of the technique.

All of the methods briefly discussed earlier focused on binary networks. In such methods, attributes of nodes are emphasized instead of the edge content which represent the actual link between the nodes. Even though more challenging, edges provide a richer characterization of community behaviour (Qi et al., 2012). Most networks are weighted, so community detection is more reliable when the actual extent of interaction between nodes is considered (Ovelgönne et al., 2010). A notable algorithm for detection of communities in weighted networks is the COPRA algorithm (Gregory, 2010). This algorithm is based on the label propagation algorithm (RAK) discussed earlier. Label propagation is done just like in the RAK algorithm only that, in this case, a node can be a member of more than one community because of the use of community identifiers. A node is allowed to keep more than one community identifier in each label without retaining all of them. This algorithm can be used on weighted networks. However, it has the same convergence problem that the RAK algorithm had. In (Tiantian Zhang and Bin Wu, 2012) a method for finding communities of users by first identifying core nodes and finding cliques around those core nodes was proposed. It was argued that having global knowledge of the graph required by most algorithms is unrealistic for very large graphs. The *Strength* algorithm proposed in (Chen et al., 2010) used this strategy. It consists of finding an initial partial community (the node with the highest node strength). The community is expanded by adding tight nodes to the partial community until detection is complete for that particular community based on a set threshold for belonging degree of the neighbours of that community. The algorithm however, degrades in its performance when the overlapping increases. In (Lu et al., 2013), a conductance-based algorithm was developed. The algorithm is just like the *Strength* algorithm only that a new objective function, Conductance, is used in addition to the belonging degree, and here the initial community is a community of two nodes in the network with the highest edge weight between the two of them.

This algorithm had a dynamic threshold and could perform well on large networks, unlike the *Strength* algorithm. However, like all the previous algorithms discussed, indirect links between nodes were not considered.

Sometimes, friends (and also criminals like fraudsters) live in close proximity to each other. This reduces the amount of communication data available to study their relationships since most of their communication happen offline (Blackburn et al., 2014). Considering indirect connections can help to reveal more information about such relationships. According to Granovetter (1973), "The degree of overlap of two individual's friendship networks varies directly with the strength of their tie to one another." Thus, nodes with stronger ties to each other are more likely to have stronger indirect links or friends-of-friends. In an attempt to determine the distance in a communication network beyond which two nodes are no longer likely to be aware of each other's activities in (Friedkin, 1983), it was observed that two persons who were more than two steps away from each other in a network were unlikely to be aware of each other's work. Work by Christakis and Fowler, (2009) also led to a theory that social influence does not end with two people who are directly connected to each other but continues up to two or three hop relationships, though with diminishing returns. Work carried out by Blackburn et al., (2014) further verified this.

3. Community Detention with Synthesised Relationship Strengths

Here, we present a method for detecting communities using a synthesised relationship strength (direct and indirect) between pairs of directly connected nodes in a network.

A. Synthesised Relationship Strengths

In (XLin et al., 2014), a simple expression for calculating synthesized relationship strengths between pairs of nodes in a network was derived. The synthesized relationship strength is the weight of the link between any two nodes, it is derived from the combination of the weights of the direct the indirect paths between the two nodes. The synthesized relationship strength $RS(v_i, v_j)$ between nodes v_i and vj is written as

$$RS(v_i, v_j) = \propto RS_d(v_i, v_j) + \beta RS_{id}(v_i, v_j)$$

$$RS(v_i, v_j) = \propto RS_d(v_i, v_j) + \beta RS_{id}(v_i, v_j) \quad (1)$$

Where $\propto$ and β a $\propto$ re we β ighting coefficients for the direct and the indirect paths respectively. Selecting the experimental values for the attenuation coefficient and weight coefficient as used in (XLin et al., 2014), The synthesized relationship strength $RS(v_i, v_j)$ between nodes v_i and v_j

$$RS(v_i, v_j) = 0.6w_{i,j} + 0.4\frac{\sum_{i=1}^{n}\left(\prod_{j=1}^{d_j} w_j\right)}{n}$$

$$RS(v_i, v_j) = 0.6w_{i,j} + 0.4\frac{\sum_{i=1}^{n}\left(\prod_{j=1}^{d_j} w_j\right)}{n}$$

(2)

where $w_{i,j}$ repres $w_{i,j}$ ents the direct weight between the two nodes, and *d* is the length of their relationship strength along a given path (number of hops in between)

For *c* two-hop indirect paths with intermediary node v_k, where v_k v_k is a n v_k eighbour of both v_i and v_j, v_i the s v_j um of weights, P_1 across P_1 all such indirect paths was calculated as:

$$P_1 = \sum_{1}^{c}[(w_{i,k}] \times w_{k,j})$$

$$P_1 = \sum_{1}^{c}[(w_{i,k}] \times w_{k,j})$$

(3)

For *m* three-hop indirect paths with intermediary nodes v_k and v_l, v_k where v_l v_k is a n v_k eighbour of v_i, v_l is a v_i n v_l eighbour of v_k, and v_j i v_k s in a v_j neighbour of v_l; the s v_l um of weights, P_2 across P_2 all such indirect paths was calculated as:

$$P_2 = \sum_{1}^{m}[(w_{i,k}] \times w_{k,l} \times w_{l,j})$$

$$P_2 = \sum_{1}^{m}[(w_{i,k}] \times w_{k,l} \times w_{l,j})$$

(4)

Therefore,

$$RS_{id}(v_i, v_j) = 0.4\frac{(P_1 + P_2)}{(c + m)}$$

$$RS_{id}(v_i, v_j) = 0.4\frac{(P_1 + P_2)}{(c + m)}$$

(5)

Hence,

$$RS(v_i, v_j) = 0.6w_{i,j} + 0.4\frac{(P_1 + P_2)}{(c + m)}$$

$$RS(v_i, v_j) = 0.6w_{i,j} + 0.4\frac{(P_1 + P_2)}{(c + m)}$$

(6)

This was then used in place of $w_{i,j}$ in the $w_{i,j}$ conductance-based algorithm.

B. Metrics Used to Detect Communities

The following metrics were used as objective functions in our algorithm for detecting communities.

1) Conductance: It measures the fraction of total edge volume that point outside the cluster. That is, it measures how well knit a graph is. The lower the conductance value, the more connected the nodes are. This can be mathematically expressed as:

$$\phi(C) = \frac{cut\left(C, C/_G\right)}{w_c}$$

$$\phi(C) = \frac{cut\left(C, C/_G\right)}{w_c}$$

(7)

where $cut\left(C, C/_G\right)$ repres $cut\left(C, C/_G\right)$ ents the number of cut edges in the community (which means all edges leaving the community), and w_c is the w_c total weight of edges in the community.

2) Belonging Degree: Assuming C is a community in a network; for a node $u \in V$; k_u , N_u , a $u \in V$ re k_u n N_u ode degrees and neighbour sets respectively. And let w_{uv} be the w_{uv} weight of the link between nodes u and v (where *v* is already in the community). k_u can th k_u en be written as:

$$k_u = \sum_{u \in N_u} w_{uv}$$

$$k_u = \sum_{u \in N_u} w_{uv}$$

(8)

For the community C, and node u, the belonging degree $B(u, C)$ betwee $B(u, C)$ n node u and community C is defined as

$$B(u, C) = \frac{\sum_{u \in C} w_{uv}}{k_u}$$

$$B(u, C) = \frac{\sum_{u \in C} w_{uv}}{k_u} \quad (9)$$

C. The ECB Algorithm

The algorithm is made up of two stages: selecting the initial temporary community and expansion. It is basically the Conductance-based Algorithm (CBA) in (Lu et al., 2013), with synthesized relationship strengths used in place of direct weights between all links in the entire process. The algorithm is as follows

(a) Input Graph G

(b) Ca G lculate synthesized relationship strength between every pair of nodes in the network

(c) If edge set is not empty, select two nodes v_i and v_j w v_i ith t v_j he highest synthesized relationship strength

(d) Calculate the Conductance $\phi(C)$ of the $\phi(C)$ community C forme C d by v_i and v_j

(v_i e) Fi v_j nd all the neighbours (N) of C

(f) Pi C ck the neighbour with the highest belonging degree B(w, C) to C

(g) Ad C d to C and fo C rm a new temporary community C'

(h) Ca C' lculate conductance $\phi(C')$ of C'

(i $\phi(C')$) If C' $\phi(C') < \phi(C)$, the $\phi(C') < \phi(C)$ n $C' = C$, go to $C' = C$ (e)

(j) If $\phi(C') > \phi(C)$, then $\phi(C') > \phi(C)$ community C is det C ected

(k) Remove edge v_i - v_j , go t v_i o v_j (c)

(l) End

4. Performance Evaluation of the ECBA

The algorithm was tested on two datasets—i.e. the nodobo dataset and ground truth data from Zachary Karate Club.

1) Test with Nodobo Dataset

The Nodobo dataset (McDiarmid et al., 2013) is publicly available. This data was retrieved from mobile phones of 27 High School students over a period of 5 months using a software. It consists of 13035 call records, 83542 SMS records and 5.2 million proximity records. The part of the dataset used for this research is the call records. Based on the scope of this work, only the source and the target phone numbers and call durations were needed. These were extracted and duplicate edges (source-destination pairs) were merged. Since this work focuses on undirected graphs node pairs were considered as duplicates if they existed as a source-destination pair, however they were permutated, multiple times. Weights were calculated using call durations for each edge.

$$\textbf{Weight} = \frac{(\textbf{Call duration with respect to a given edge})}{(\textbf{Maximum call duration in the dataset})}$$

This pre-processing resulted in 575 nodes with 642 edges.
The result of the communities detected are shown as compared to that of the original Conductance-based Algorithm as shown in Table 1.

Table 1: Detected communities

Number of members	Detected by ECBA	Detected by CBA
> 5	23 communities	23 communities
>10	16 communities	18 communities
>20	11 communities	12 communities
>30	8 communities	9 communities
>40	6 communities	7 communities
>50	2 communities	3 communities
>60	1 community	1 community
>70	1 community	0 communities

From the results, it is clear our algorithm generally detected more small-sized communities than the existing algorithm. It showed more details of splits among members. Also the largest community detected by our algorithm had 105 nodes. The same community (with the same initial node pair), was detected with only 68 nodes using the CBA algorithm. Thus, with our algorithm, it was possible to identify

members of that community which were not detected by the CBA. Other metrics used for evaluation include:

a) Conductance

The performance of the Enhanced Conductance-based Algorithm (ECBA) was compared with the Conductance-based Algorithm (CBA) by plotting the graphs of their conductance versus community size. The result shows a scatter plot with Least Squares lines in Fig 1a. As seen from the least-square lines, the ECBA had lower conductance values as community sizes increased from about 20 members and above. By definition, the smaller the conductance, the tighter the connection (that is, the stronger the relationships) between members in the community. This means that the ECBA formed tighter clusters for communities with larger sizes (greater than 20) while the CBA formed tighter clusters for smaller community sizes. The last two scatter points to the right show that the community detected by the ECBA with 105 nodes is much tighter (conductance = 0.02456) than that of the same community which was detected as having only 68 nodes by the CBA (conductance = 0.2112). Hence, the hidden nodes ignored by the CBA were very important members of the community.

b) Average Distances

The distance between two nodes in a network is the length of the shortest path between them. For this network, the distances used were inverse of the weights as discussed earlier. The average distance is the sum of the shortest path distances between all pairs of nodes in the community divided by the community size. Fig 1b shows a plot of the average distances for each of the communities formed using the ECBA and the CBA against community sizes. The ECBA finds communities with lower average distances as the sizes of the communities are increasing, while the CBA shows lower average distances for smaller communities (<20). This can be interpreted as the ECBA clustering nodes with a stronger connection with one another (smaller shortest path lengths) than the CBA for larger communities.

c) Average Distances

The values of the scaled densities for each of the communities detected by both algorithms were plotted against community sizes in Fig 1c. The ECBA has a higher scaled density up to point 40 on the community size axis and CBA has a higher scaled density from point 80 upwards. However, the scatter points show only ECBA having a community at all up to 80 nodes in size. Hence we can best compare the two algorithms with earlier values than 80. This shows that the ECBA communities have averagely slightly higher density of clusters than the CBA.

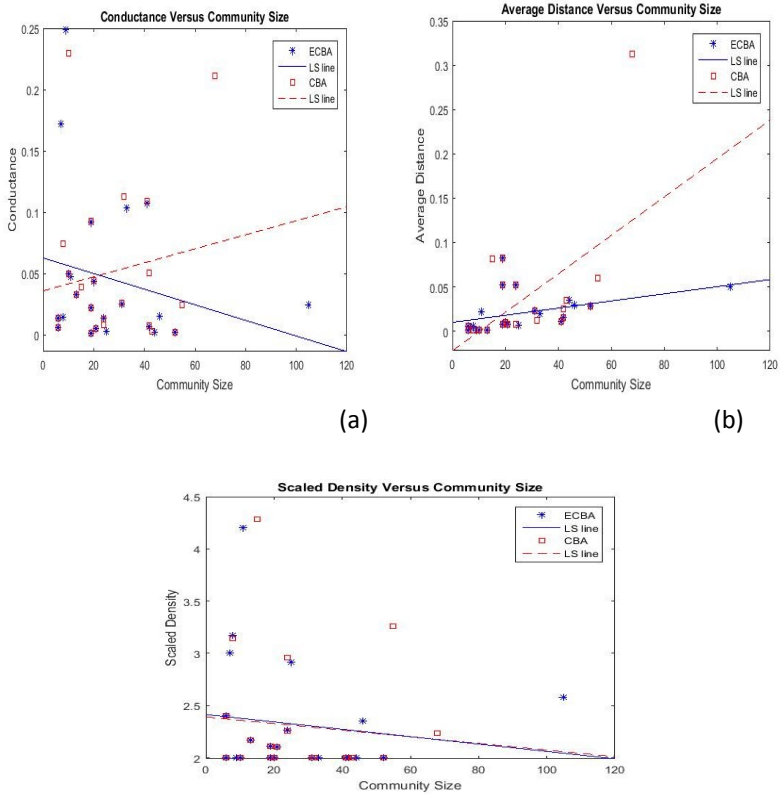


Fig 1: Graphs comparing performance of the ECBA and CBA

2) Test with Zachary Dataset

Our algorithm was also tested with the Zachary Karate club dataset. This data is from an already known community structure of 34 members of a Karate club observed for three years - from 1970 to 1972. After a conflict between the club's president and a part-time instructor over lesson fees, members of the club were split into two main groups (Zachary, 1977). In this work, a weighted version of the karate network was used to test the Enhanced Conductance-based Algorithm. The results of the detection were compared with the real life communities observed by Zachary. The original communities formed by Zachary are shown in Table 2.

Table 2: Community Structure as observed by Zachary

Community	Members
1	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22
2	9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34

The detection done with our algorithm shows four communities instead of the two observed by Zachary as seen in Table 3. However, it was observed that members of community 2 and community 4, except for node 9, 29 and 32, were both subsets of the community 1 in Zachary's original detection. Also, community 1 and community 3 in Table 2, except for node 2 and 3, are subsets of community 2 in Zachary's detection.

Table 3: Community Structure Observed by the ECBA

Community	Members
1	25, 26, 29, 32
2	1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22, 29
3	2, 3, 9, 10, 14, 15, 16, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34
4	6, 7, 17, 1, 9, 32, 5, 11

When the pairs of subsets are merged we have the communities as shown in Table 4. It can be seen from comparing Table 2 and Table 4 that every member of the community 1 in the Zachary dataset is also in the first community we have after merging the community pairs as shown in Table 4. Also, every single member of community 2 in the Zachary dataset is also placed in the second community in Table 4. The extra nodes in each of these communities, that is: 2, 3, 9, 14, 20 and 29, are overlapping nodes.

Table 4: Merged Communities

Community	Members
2 & 4	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 17, 18, 20, 22, 29, 32
1 & 3	2, 3, 9, 10, 14, 15, 16, 19, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34

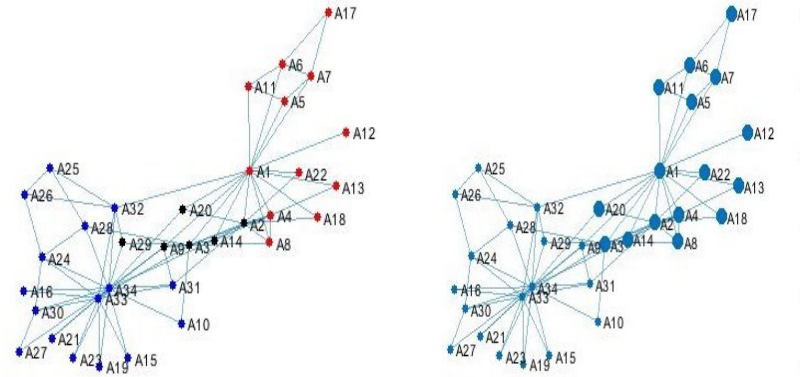


Fig 2: (a) Communities as observed Zachary (b) Communities as detected by ECBA

The result of this test shows that the proposed algorithm did not only detect communities but also sub-communities within these communities, showing more detailed clustering of nodes.

5. Discussion

The results from the detection and evaluation discussed in the previous section shows that more community members not seen by the CBA can be detected by the ECBA. The results point to the fact that using indirect relationship strength compensates for some of the closely connected nodes that have very little online communication and could hence be mistaken for weak ties. Apart from showing the true nature of such links, communities formed showed a great improvement in compactness. This also suggests that a higher amount of mutual information is shared across the links between members of communities formed via synthesising both direct and indirect paths. The test with Zachary karate club dataset also shows that the introduction of indirect relationships across the links gave rise to overlapping. Though every member was detected in its correct community, overlaps became very visible and sub-communities could be formed to show greater detail of how members related.

6. Conclusions

In this work, an Enhanced Conductance-based Algorithm (ECBA) for community detection in weighted networks with undirected graphs was presented. It was tested on a mobile phone dataset and a dataset from a social club with already known community structure. Results show that The Enhanced Conductance-based Algorithm outperforms the existing Conductance-based Algorithm in detecting communities of larger sizes (up to about 20 nodes or more). This work has also revealed that detecting communities with both direct and indirect relationship strengths

gives more details of node relationships than what is obtained by using only direct relationship strengths.

Correspondence
Bala A. Salihu
Department of Telecommunications Engineering
Federal University of Technology, Minna, Nigeria

References

Ahuja, M., Singh, J. & Neha, 2016. Practical Applications of Community Detection. *International Journal of Advanced Research in Computer Science and Software Engineering,* 6(4), pp. 412-415.

Blackburn, X.Z.J., Kourtellis, N., Skvoretz, J., Iamnitchi, A., 2014. The power of indirect ties in friend-to-friend storage systems, in: 14-Th IEEE International Conference on Peer-to-Peer Computing. IEEE, pp. 1–5.

Borgatti, S.P., 2009. 2-Mode concepts in social network analysis. Encyclopedia of complexity and system science 6.

Chen, D., Shang, M., Lv, Z., Fu, Y., 2010. Detecting overlapping communities of weighted networks via a local algorithm. Physica A: Statistical Mechanics and its Applications 389, 4177–4187. doi:10.1016/j.physa.2010.05.046
Christakis, N.A., Fowler, J.H., 2009. Connected: the surprising power of our social networks and how they shape our lives, 1st ed. ed. Little, Brown and Co, New York.

Friedkin, N. E., 1983. Horizons of observability and limits of informal control in organizations. *Social Forces,* 62(6), pp. 54-77.

Fortunato, S., 2010. Community detection in graphs. Physics Reports 486, 75–174. doi:10.1016/j.physrep.2009.11.002

Fortunato, S., Lancichinetti, A., 2009. Community detection algorithms: a comparative analysis: invited presentation, extended abstract, in: Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), p. 27.

Girvan, M., Newman, M.E., 2002. Community structure in social and biological networks. Proceedings of the national academy of sciences 99, 7821–7826.

Granovetter, M. S., 1973. The Strength of Weak Ties. *American Journal of Sociology,* 78(6), pp. 1360-1380.
Gregory, S., 2010. Finding overlapping communities in networks by label propagation. New Journal of Physics 12, 103018. doi:10.1088/1367-2630/12/10/103018.

Lu, Z., Wen, Y., Cao, G., 2013. Community detection in weighted networks: Algorithms and applications, in: Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference on. IEEE, pp. 179–184.

McDiarmid, A., Bell, S., Irvine, J., Banford, J., 2013. Nodobo: Detailed mobile phone usage dataset. Unpublished paper, accessed at http://nodobo. com/papers/ietel. pdf on 9–21.

Ovelgönne, M., Geyer-Schulz, A., Stein, M., 2010. Randomized greedy modularity optimization for group detection in huge social networks, in: Proceedings of the Fourth SNA-KDD Workshop, KDD 2010, July. pp. 1–9.

Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818. doi:10.1038/nature03607.

Qi, G.-J., Aggarwal, C.C., Huang, T., 2012. Community detection with edge content in social media networks, in: 2012 IEEE 28th International Conference on Data Engineering. IEEE, pp. 534–545.

Raghavan, U.N., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76. doi:10.1103/PhysRevE.76.036106.

Tiantian Zhang, Bin Wu, 2012. A Method for Local Community Detection by Finding Core Nodes. IEEE, pp. 1171–1176. doi:10.1109/ASONAM.2012.202.

XLin, X., Shang, T., Liu, J., 2014. An Estimation Method for Relationship Strength in Weighted Social Network Graphs. Journal of Computer and Communications 02, 82–89. doi:10.4236/jcc.2014.24012.

Zachary, W.W., 1977. An information flow model for conflict and fission in small groups. Journal of anthropological research 452–473.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.125-135

Good Governance As A Security Management Strategy: An Overview of Nigeria's Experience

PATRICK I. CHUKE
Ambrose Alli University, Ekpoma, Nigeria

CHIDIEBERE T. OKUTALUKWE
University of Ibadan, Nigeria

ABSTRACT The unabated authoritarian influence which impinges on leadership behaviors, policy and programmes in Nigeria's state, is a testament of antithesis of governance with consequential effects on national security. There is a causal relationship between governance and security, especially with the new thinking in security discourse that sees security from the perspective of "humanness". This paper attempts to anchor the numerous challenges of insecurity in Nigeria at the root of bad governance, which serves as a harbinger of deprivation, corruption, marginalization and discontent. The study made use of qualitative data as its main methodological guide. It is noted among others that transparency, accountability, and equitable distribution of national resources form a critical factor in security management. It also asserted that conceptualizing security from human security doctrine constitutes a fundamental strategy for improving governance, at the same time enhancing security management.

Keywords: Good Governance, security, Democracy, Human

Introduction

Governance has been at the centre stage of security discourse, and equally prominent on Nigeria's development agenda. Nigeria's dismal security management performance has been highly attributable to weak or ineffective governance (Imobighe, 2013). This paper tends to situate the crises of security in Nigeria at the root of mal-governance. Governance systems in each and every African state, cannot be divorced from issues of security and development. In fact, they are closely associated (Sessay and Ukeje, 2015). The two scholars, argued that both are Siamese twins, joined by heart, which cannot be successfully separated without inflicting irreparable damage on the other. The reason is not farfetched; experience has clearly shown that the conflict in Niger-Delta requires a coherent strategy for the provision and sustenance of soft infrastructures like schools, housing, industries, electricity, and roads to help in mitigating the rising wave of insecurity in the region. The sovereign agitation of south-easterners for a Biafran state demands an attention that is methodically rooted in inclusive and developmental governance. With disastrous results and consequence in Nigeria, that incoherent governance

breeds impunity, corruption, lack of transparency and accountability, disrespect for human right and insecurity, and profligacy, among others, which are the antithesis of sustainable security and development.

The elongated crises of security in Nigeria is a clear depiction of continuous desecration of governance that is anchored on the trinity of democracy, development and freedom that ensures national security. The "security-development nexus is brilliantly articulated in Mark Duffield's Global Governance and the New Wars. According to Duffield:

> The very notion of development has been radicalized in the process, and now require the direct transformation of Third World societies. This radicalization is closely associated with the redefinition of security because conflict is understood as stemming from a developmental malaise; under development itself is now seen as a source of instability (Duffield, 2001)

The poor quality of governance has contributed in no small measure to the escalating incidences of insecurity and conflict in Nigeria. The World Bank has defined good governance as a type of governance that is

> "Predictable, open and enlightened termed policy-making, a bureaucracy imbued with a professional Ethos acting furtherance of the public good, the rule of law, transparent process and a strong civil society participating in public affairs" (World Bank, 1994).

The political development across Nigeria is far from the norms of good governance as defined above. The social and economic upheaval that always accompany bad governance and the collapse of the developmental state have continued to undermine national solidarity and cohesion, threatening national and sub-federal security (Alli, 2010)

Security being a strategic object, connotes the systematic programme of government design to create harmonious relationship between governance- state and society. In this sense, governance denotes that government should be security accountable to its people. The idea of security is innate in human kind and a certain understanding of security plays a role in every aspect of life of the individual, groups, nation and the international community. The state is generally presented as a people organized on the basis of law and development in a given territory (Thomas, 2013). It is clear that the state requires the element of the people; law and order, territory and development encapsulated in sovereignty to operate. The state serves utilitarian purposes and not an end in its self. The primary purpose of the state is to ensure a secured people for development Imobighe (2003) adduced that a secured state is the one that is reasonably free from, or not exposed to external aggression and internal sabotage. The execution of security practices involves securitization (a process of articulating rationalizing how and why a factor should be treated as a security is sure), enactment of rules and institutional arrangement for rules enforcement and the administration of the system to ensure the realization of security objectives.

The security threats and challenges in Nigeria are enormously strategic and evident but had never been precarious and alarming as in this generation (Thomas, 2013). There is persisting spate of hostage taking and kidnapping, militant insurgencies, pipe line vandalism, ethnic insurrections, fatal inter-community and intra-community conflicts, inter- religious and intra-religious wars, illegal arms trafficking, armed banditry, political violence and gangsterism.

Consequently, the interplay of economic crises, social upheaval and political instability expose the inadequacies of Nigerian State in general and exacerbate the economic condition of the people to fall further into deprivation and desperation. As Nnoli (2006) argued, people who believe that the government no longer represent their interest seek by all means to overthrow it. Accordingly, it has been adduced that in these circumstances, it is wrong to continue to see security from the point of view as external enemies of the people also pose threat to peace and security. Thus, in addition to poor governance; rent seeking behavior, primitive accumulation and ostentatious living of the leaders, have combined to generate conflict and insurgent activities which threaten national security.

The ruling elites in Nigeria has embraced authoritarianism as an instrument of support and loyalty to the government. The state apparatus has been hijacked to unleash political, economic, and religious polarization, primitive accumulation and disempowerment. The nature of governance has been built around personal glorification, rather than concentrated on the basis of common wealth. The mismanagement of the commonwealth or the resources of Nigeria by the ruling political and economic elites through corruption and cronyism point to the fact that the irresponsible behavior of government connotes colossal security to Nigeria. The recent revelation of how national treasury was emptied for the purpose of 2015 general elections and the partisan nature of security agencies during the election underscores the great in ordinate quest for power even in democracies can negatively affect good governance.

This norm-less pattern of governance has created an environment of national betrayal which have left a huge mark on threat to security in the country. The continuous rise of ethno-religious inclined form of government constitutes a disturbance to national peace. It is factual to say that, in Nigeria key government policy and programs are viewed from the prism of ethnic and religious supremacy. The various sections of the country are at 'war' with each other. In the process, the country is orphaned by the people as it is denied of patriotic support. Every section of the country supports the country for whatever advantage it can get and disowns it whenever it perceives that its political advantage is in jeopardy. Support for the country is conditional and instrumental rather than principled or entrenched. Nigeria, therefore, carries a heavy burden of unresolved issues ranging from such basic issues as the nature of its federalism, allocation of powers, funding of government, administration of domestic security, political appointments and staffing of bureaucracy, regulation of religions. With this plethora of critical issues, some resort to faith and divine grace to say that Nigeria is well and kicking.

It is factual and evident to characterize the nature and system of governance in Nigeria as weak, ineffective and umpteen rudderless. At the crux of governance capacity is leadership, which the Nigerian state suffers in quantum deficiency.

Even the advent of democratization and democracy that serves as the locomotive of functional governance in other climes, have failed to re-clone itself in Nigeria. Effective democratic governance, its values and institutions constitute a factor of security management. However, managing conflict by democratic means and using its carefully engineered structures and set of institutions to equitably distribute the resources of the state to all segment of the society forms part of improving governance capacity to strategically address security challenges in Nigeria.

Methodology

This study applied qualitative research approach as its methodological insights. It also relied on secondary sources of data in its analysis, like archival materials, textbooks, internet or online materials, journals and other relevant publications. Conscious efforts were made in analyzing all secondary data objectively.

Conceptual Background

Governance

The world bank- defines governance as the manner in which power is exercised in the management of a country's economic and social resources for development (World Bank, 1989). Many scholars (Botchway, 2001; Nanda, 2006; Smith, 2007) have built on and expanded this world bank definition of governance and have come up with some parameters for determining what should constitute good governance (Imobighe, 2013). These include: democratic practice, transparency, accountability and bureaucratic efficiency.

Etymologically, both terms, good and governance, involve a conception of both ethic and aesthetic of politics. Ethics or moral conception refers to good politics as opposed to the bad. This paper sees good governance, effective governance, functional governance and reliable governance as semantics without changing the original meaning of the discourse governance. Aesthetic conception or artistic, in the sense that the words government and governance have Greek origin, meaning rudder of a ship or the steer of a ship. Thus government is a mechanism or political complex embracing the entire organs of the state for control of the people. While governance is the process of politics or the art of holding the steer (rudder); the art of piloting a ship bound for far way horizons so that it can arrive safely without mishap. In the context of the state, governance is the act of governing the society, the art to direct the state to ensure good functioning for political institute and of the state organs for a harmonious societal development (Sylla, 2001). Simply put, governance is the art of governing, or what you could call the management of public affairs. It can also be referred to as the work which government do. It also relates to the process of granting public power and the use to which such power is put, which ideally should be for service to the people (Imobighe 2013). To the conceptual understanding of this paper, governance connotes the expertise, skills and creativity to judiciously manage the resources of the state and direct its affairs with transparency, accountability and probity for the purpose of harmony, cohesion,

peace, are in a political organization or society. In consonance with this note, it is imperative to see selfless and purposeful leadership and political elites mobilization towards formulating and implementing people centered policies and programs as the bulwark of governance.

Security

Security, the freedom from danger, risk, care, intimidation apprehension, the feeling or assurance of safety, the peace of mind or absence of fear, constitutes the fundamental objectives and indeed the foremost responsibility of the state and every government (Thomas, 2013). A proper understanding of security is important for an adequate explanation of National security challenges. Security is a key concept in the social sciences that refers to framework and dimensions, applies to individuals, issue areas, social conventions, and changing historical condition and circumstances. There seems to be lack of definitional consensus on security. However, the perspectives that have been expounded by Imobighe and Buzan will aid in understanding the concept. It is necessary to note at length the definition offered by T.A Imobighe. He states that:

> Security has to do with free from danger, or with threat to a nation's ability to protect and develop itself, promote its cherished values and legitimate interest and enhance the well-being of its people. Thus, internal security could be seen as the freedom from or the absence of those tendencies which could undermine internal cohesion and the corporate existence of the nation and its ability to maintain its vital institutions for the promotion of its core values and socio-political and economic objectives, as well as meet the legitimate aspiration of the people (Imobighe, 1990).

The Elucidation of the concept of security that Buzan presents goes a little further in that he identifies three levels at which security can be analyzed. These are the individual level, the national level and the international level. Buzan concedes that at the level of the individual, it is difficult to provide an all-encompassing definition of security (Oche, 2001). Nevertheless, the value that individual attempt to secure tend to be the same and these include life, health, status, freedom and health. The threats which individuals try to achieve security from what he refers to as social threat which are of three main types. These are physical seizure or destruction of property; threats to rights such as denial of civil liberties and thereat to position or status such as demotion or public humiliation (Buzan, 1983). At the level of the nation-state, security assumes a slightly different perspective. It is a level reference is made to the concept of national security which involves essentially, the ability of a nation to protect its internal values from external threats. These internal values include the idea and conception of the state that is held by its citizen; the political economic and social institution within the state; and the geographical and territorial base of the state with all its endowment.

At the international level, security derives from the perceived need and requirement of nation-states to defend their core national values against the pursuits of

other states. The perception of threat from the international environment compels states to increase their level of security in the neighboring states.

Over the years, we have witnessed the systematic expansion of the frontiers of security as more and more people with different socio-economic, political, cultural and environmental backgrounds bring new perspective to bear on the subject. This has resulted in the emergence of a broader view of security that addresses all human vulnerabilities; be they economic, political, social, military, environmental threats to his physiological need, his values or identities. Adopting this broader view in this discourse, means our vision of security will not be limited to the conventional state- centric definition of the concept in which security relates to the elimination of threats to the territorial integrity or physical existence of the state. Rather this paper endeavors to construct security around human person and his needs. This piece of discourse emphasis is on human security.

Good Governance as a corner stone of human security

This study has been able to establish a theoretical linkage between governance and security. Logically, the aspect of security it covers is the "new thinking" in security discourse that is human security. Governance is a concept dedicated for solutions for human needs both individually and collectively for national solidarity and unity. Accordingly, security is increasingly understood to be about development, justice and fairness, all of which can only come through good governance (Alli 2010). Most of the threats to security in Nigeria are caused largely by the absence of governance. In the wider world most of the threats to peace and security also emanate from real or imagined injustice, oppression, exploitation, deprivation and marginalization.

The term "human security" may be of recent origin, the ideas that underpin the concept are far from new. For more than a century, since the founding of the international committee of the Red Cross in the 1860s- a doctrine based on the security of people has been gathering momentum (Hubert, 2007).

The specific phrase, "human security", is most common associated with the 1994 UNDP Human Development report, an attempt to capture the Post-Cold war peace dividend and redirect those resources towards the development agenda. The definition advance in the report was extremely ambitious, human security defined as the summation of seven dimension of security; economic, food, health, environmental, personal community and political (Hubert, 2001). By focusing on people and highlighting non-traditional threats the UNDP made an important contribution to post- cold war thinking about security. It went further to explain that:

> "for long security was narrowly interpreted as security of territory from external aggression or as protection of national interests in foreign policy or as global security from the threat of a nuclear holocaust. It has been related more to nation-state than to people, forgotten were the legitimate concerns of ordinary people who sought security in their daily lives. For many of them security symbolised protection from the threat of disease, hunger, unemployment crime, social conflict and environmental hazards" (UNDP 1994).

It is believed that security must go beyond the classical conventional theory and embrace a development component; meaning it must impact positively on the general well-being of all the people (Imobighe 2013). As the UN Secretary General, Ban Ki-Moon (2012) rightly remarked arms cannot address the grave security concerns rising from demographic trends, chronic poverty, economic inequality, environmental degradation, pandemic diseases, organized crime, repressive governance and other developments. Indeed, arms will eat up the resources required for the satisfaction of those tangible human needs.

Mo Ibrahim [Chair of the Board of Mo Ibrahim Foundation which sponsors *Ibrahim Prize for Achievement in African Leadership*] appreciates this linkage between governance and security and has decided to invest in the promotion of good governance. The Mo Ibrahim index is a new framework for assessing the equality of governance in Africa. Five criteria have been established for carrying out this assessment; safety and security, rule of law, transparency and corruption, participation and human rights and sustainable economic governance.

Human Security does not supplant national security. A human security perspective asserts that the security of the state is not an end in itself. Rather, it is a means of enduring security for its people. In this context, state security and human security are mutually supportive. Building an effective, democratic state that values its own people and protects minorities is a central strategy for promoting human security (Hubert, 2001). At the same time, improving the human security of its people strengthens the legitimacy, stability and security of a state. When state are externally aggressive, internally draconian, or too weak to govern effectively, they threaten the security of people. Where human security exist as a fact rather than an aspiration, these conditions can be attributed in large measure to the effective governance of states.

It is obvious that the purely militaristic state-centric and regime centered approach to Nigeria's security has failed to deliver peace and security in the country. Nigeria has failed to consciously improve its governance capacity, which we noted before, has to do with the manner the affairs of the country is conducted on a day to day basis especially the way the nation's resources is managed and the style the decisions affecting the people are taking. The critical choices the country's leaders make as to what resources to exploit, how the resources should be allocated and utilized and the manner of public participation in the decisions affecting them are all issues of governance which are critical to the stability of the nation (Imobighe, 2013) .

Altruistically, conforming to values of governance and democratizing leadership attitudes by championing the supremacy of peopled centered policy objectives constitutes a strategic management of security in the country. Until we get these issues right the country cannot have the security which everyone desires even if we field the strongest and most sophisticated armed forces and the best security agencies in Africa. At the moment, it does not appear that Nigerian leaders know that the manner they manage the country's affair on a daily basis has serious implications for the nation's security. Their ravenous attitude and behavior when it comes to the management of public funds is an indication that the Nigerian political elite

have not got those critical issues of governance that impinge on security right (Imobighe, 2013)

The emerging threat to Nigerian security

Youth unemployment is a cross-cutting challenge in Nigeria. It is mass youth unemployment and the hopeless plight of the teeming idle youth that serves as contributory factor to Nigeria militancy, kidnapping and political violence. It is ironical, that national security effort largely depends of strengthening the coercive apparatus of government rather than use policy instrument to eliminate the conditions that bred societal discontent and chaos. The Boko Haram insurrection is also rooted in the manipulation of the unemployed youth with tainted religious ideologies to torment crises against the state. Any country with a predominantly youth population unemployed and idle understood the potency of such threat to its national security. The situation is such that the government had to admit in its transformation agenda (2011-2015) that 'the Nigerian economy is experiencing growth without employment": adding that "the unemployed population is at present dominated by youth who are mostly, school leavers, with senior school qualifications and graduates from tertiary institutions". The government further remarked that "the national composite employment data shows that rate of unemployment surged from 11.9 percent in 2006 to 14.6 percent in 2007 and 21.1 percent in 2010. The ruling regime in Nigeria is not accountable to its citizen, and at the slightest suggestion of resistance to tyranny, it resort to brutish violence to silence it.

The agitation by ethnic nationalities clamoring for a sovereign state of their own ethnic nation is assuming a monumental security disturbance to the nation. The activities of the movement for the actualization of sovereign state of Biafra (MASSOB), indigenous people of Biafra (IPOB) Odua Peoples Congress (OPC) and other groups from Niger-delta signifies that the state lacks the capacity to initiate developmental agenda that will calm the wave of frustration discontent in the polity. Rather than look into the reason for their agitation, the main focus has been to deploy the full weight of the country's coercive power to suppress or eliminate those regarded as the agitators, the defiant elements or trouble makers within the society that is those who are dissatisfied with the existing state of affairs. Basically, no inclination has been shown to probe into the causes of discontent in order to apply appropriate measures to address them administratively.

The Nigerian state is replete with the obvious signs of institutional ability to respond to the challenges of governance. The absurdities of political class have weakened the institutional internal capability to regulate political excesses and interference. This trend constitutes a gigantic threats to governance and security in the country. When governance is answerable to individualistic dictation and crony innuendoes, it means that the state is dangling between failed or failing pendulum. The massive looting and corruption that is prevalent in Nigeria's security sectors is alarming. The current (since 2015) trials that is undergoing over the state of 2.2 billion dollars perpetrated (allegedly) by Dasuki, the former national security adviser, and some top military officers underlines the colossal deficiency in our security governance where transparency, due process and accountability in manage-

ment of national resources is elusive. The money meant for security was diverted for political activities. The unabated corruption in Nigeria's bureaucracy constitutes its greatest security threats.

Conclusion

This paper has attempted to draw an analogy of Nigeria's current security predicament in relations to the nature of governance in the country. It discussed how effective governance system can be tantamount to security management. It saw the problem of security challenges in Nigeria as the failure of governance to adequately address issues of discontent, corruption, deprivation, poverty and inequality. The main finding of the study is that government have invested enormous energy and resources to engender security from the classical point of view, rather than pursue security management from the human security perspective that governance synchronized with it. At the primacy of governance output is the resolving of the fundamental variables of conflicts triggers and catalysts by ensuring judicious, equitable, transparent, responsible etc. management of national resources for solidarity, harmony and peace among the people.

Correspondence
Patrick I. Chuke, PhD
Department of Political Science
Ambrose Alli University, Ekpoma, Nigeria
Email: Patrickchuke2014@gmai.com
Tel.: 08037751021

Chidiebere T. Okutalukwe
Doctoral Student, Department of Political Science
University of Ibadan, Nigeria
Email: Chidoo4Life@yahoo.com
Tel.: 08039156241

References

Alli,W.O (2010). Security Challenges of Governance in West Africa in Eze O.C Anigbo; C.A and Dokubo, C.Q. Nigeria's Security: Interest in Africa.

Botchway, F.N (2001) "Good Governance: The old New, the Principle and the Elements" in Florida Journal of International Law vol.3, no.1: p.159.

Buzan, B. (1983), People State and Fear: The National Security Problems in International Relations. Wheatsheaf Books. p. 18-20.

Duffield, M. (2001), Governance and New Wars: Merging of Development and Security. London Led Books.

Hubert, D. (2001) Human Security: Safety for People in a Changing World, in Akindele, R.A and Ate, B.E (Eds) Beyond Conflict Resolution Managing African Security in the 21st century. Vantage press, Ibadan.

Imobighe, T.A (1990), Doctrines for and Threats to Internal Security, Ekoko A.E and vost, M (eds), Nigerian Defence Policy. Issues and Problems. Lagos. Malt-house press.

Imobighe, T.A (2003) Nigerians Defense and National Security Linkage: A Frame-work of Analysis. Ibadan, Heinemann Educational Books

Imobighe, T.A (2013), Governance and Nigeria's national security. In Imobighe, T.A and Ebohon, S.I Themes and issues in Nigerian Governance and Politics. (nipps) Kuru Jos, Plateau State.

Ki-moon, B. (2012) The World Is Over-Armed and Peace Is Underfunded, New York: UNODA.

Nanda, V.P. (2006) "The Good Governance Concept Revisited" in Annals of American Academy of Political And Social Sciences, p.269.

Nnoli, O. (2006) National security in Africa: a radical view perspective, Enugu, PACREP.

Oche, O.C (2001) Democratization and the Management of African Security, In Akindele R.A and Ate,B.E (eds), Beyond Conflict Resolution: Managing African security in the 21st century. Vantage press, Ibadan. Print serve, Lagos.

Sesay, and Ukeje C, (2015) Africa in the 21th Century: A General Introduction: in sesay A and Ukeje, C. (CDS) Security and Developmental Challenges for Africa. In the 21st Century. Hamtul press, plateau state PP.1-18.

Sylla, L. (2001), "Good Governance and military question in west Africa", in Akhaine, S.O (Ed.). Path to Demilitarization and Democratic Consolidation in West Africa. (ENCOD).

Smith, B.C (2007) Good governance and development, Hound mills, united king-dom.

The Transformation Agenda (2015), Key Priority Policies, Programs and Prospects of the Federal Government of Nigeria.

Thomas, A.N (2013) National Security Management: the Nigerian Perspective in Imobighe, T.A and Ebohon, S.I (cds) Theme and Issues in Nigerian Governance and Politics. A Publication of National Institute for Policy and Strategic Studies (NIPPS).

UNDP, (1994) Human Development Report, UN, New York.

World Bank, (1958). Sub Saharan Africa: From Crisis to Sustainable Growth. The World Bank Washington.

World Bank, (1994), governance: The World Bank Is Experience, The World Bank, Washington.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.137-150

Design and Implementation of Autonomous Ground Control Station for Surveillance UAV

E.N. ONWUKA, A.J. ONUMANYI & YUSUF YUSUF FOLAWIYO
Federal University of Technology, Minna, Nigeria

ABSTRACT One of the world's challenges in the contemporary time is insecurity at both national and international level. Across the globe, both governments and individuals are devising various ways to improve the security of life and properties. It has been observed that different environments and areas require different ways to resolve its insecurity issues. Unmanned Arial Vehicle (UAV) has become one of the tools being used to address some national security problems. This is particularly necessary in hostile environments or large facilities where it is not cost effective to have enough security personnel to man all the place at all times – an example is a big campus. Currently, the quest in UAV research is autonomous control of its missions. In this paper, the design and implementation of autonomous ground control station (GCS) for a micro UAV for campus surveillance is presented. The aim of this paper is to create and implement an unmanned aerial vehicle ground control station for manual and autonomous control of an UAV. The Gidan kwano campus of the Federal University of Technology Minna, was used as a case study for the implementation. The work does not entail manufacturing a UAV from scratch, but it requires selection, and modification of pre-fabricated UAV frames and then to develop and install the electronic and software controls. Works have been done on autonomous UAVs, but we are yet to have a UAV control algorithm indigenous to our environment. The autonomous control system was implemented on a Java desktop application software that drives UAVs through waypoints to surveillance missions. The type of artificial intelligence employed in this work is the expert system that uses inference engine to make decisions based on information received from the UAV telemetry circuit. The driving algorithm works by arming the UAV at a reference point, the UAV is then rotated to the direction of the mission, and it is moved forward by the GCS until it reaches its destination. The results obtained from field trails were promising, the manual control of the GCS worked efficiently, the telemetry circuit was able to send information from the UAV to the GCS for analysis and proper planning of the UAV mission, and the autonomous control of the GCS was able to arm the UAV as well as control it through waypoints to its mission. The developed system can be used for military operations such as surveillance missions and aid delivery purposes, and commercial purposes such as goods delivery services.

Keywords: Unmanned Arial Vehicle, UAV, artificial intelligence, autonomous ground control station, GCS,

1. Introduction

As early as 1920, multicopter vehicles were being designed, built, and used for experimentation with aerial vehicle projects (Anton Nakazawa, 2013). Unmanned

Aerial Vehicle (UAV) systems are aerial vehicles without on-board human pilots, which involve widespread human participation to achieve effective flight manoeuvres (F.B. da Silva, 2007). A modern UAV requires a lot of people to monitor several operations on the UAV during flight missions. Therefore, there is a significant amount of human interactions with the UAV system during flight operations, this is probably why confidence on UAVs remains to grow in military and civilian tasks. UAVs are planned from the onset to accomplish a particular task. So, all UAV systems have features that distinguish them from other air vehicles and they are usually categorised by the abilities of the UAV such as the flight-altitude, flight range and size of the UAV (Austin, 2010). They come in a variety of sizes, designs and functionalities, for example, the target and decoy, exploration, combat types, research and development, civil, and commercial UAVS. Initially, UAVs were controlled remotely but autonomous control is now becoming prevalent (Austin, 2010). It is envisioned that autonomous UAVs will be very useful in the area of security surveillance due to their ability to carry high definition cameras and their ability to operate in stealth mode at a very high altitude.

One of the world's challenges in the contemporary time is insecurity at both national and international level. These are brought about by political unrest, religious intolerance all which has given rise to suicide bombings, kidnapping and the like. Across the globe, governments, groups, and individuals are devising various ways to improve the security of lives and properties. It has been observed that different environments and areas require different ways to resolve its insecurity issues. It is believed that refining security tools, such as UAVs, for better performance is necessary and should be a continuous exercise. Moreover, countries should design security tools that are customised to their peculiar security challenges. In this work, a ground control station for Unmanned Aerial Vehicles (UAVs) with autonomous and manual control is presented. This is designed to be used for surveillance operations in a campus environment. To achieve this the following subsystems were developed: (i) a telemetry circuit that can send control commands to a UAV and return data from the UAV's GCS, (ii) the GCS driver software for autonomous control of the UAV was developed on a java platform, (iii) an android App for manual control of the UAV, this is necessary because manual control of the UAV may be necessary sometimes. The Gidan kwano campus of the Federal University of Technology (FUT) Minna, was used as a case study for the implementation.

Gidan Kwano campus of the Federal University of Technology Minna is very big, covering about 10,000 hectares of land. There is yet no perimeter fence around this property, moreover, less than a quarter is currently occupied by the university. Certain security challenges like encroachment by villagers and cattle rearers, which among other ills, destroy agricultural research farms and sometimes attack students, are experienced. The encroachment of the University land, if not checked, will lead to eventual loss of such properties. It will be extremely expensive and inconvenient to employ and manage enough security operatives to patrol this expanse of land regularly, in order to check excesses. It is envisioned that the use of UAV will greatly reduce the cost of securing the campus, especially if this UAV is

locally enhanced. The enhancement done in this work is incorporation of the autonomous GCS with alternate manual control.

2. Related Works

Several works such as autonomous quadcopter docking system have been done on autonomous UAVs, but the algorithm employed in these UAVs are either inefficient or not expressed clearly, moreover, to the best of our knowledge, we are yet to have a UAV control algorithm indigenous to our environment. UAV applications have been changing from military applications into civilian uses such as aerial photography, environmental surveillance, warehouse delivery, and disaster relief operations. Most UAVs are often found to be expensive and complex. Moreover, researches are still ongoing in micro UAVs some of which are bio-stimulated projects (Matt Parker, 2011). These designs are modelled in the form of insects and birds, but just as the huge military UAVs are also costly. It is apparent that the micro-UAVs are too small to satisfy the essential technology (Matt Parker, 2011). Though modern-day technology is quickly changing and improving, UAVs developments began decades ago, even before the first manned airplane flight occurred in 1903. The initial efforts were made in France in the year 1782 by the Montgolfier brothers (Matt Parker, 2011).

Micro unmanned aerial vehicles are a fascinating point of study, (Matt Parker, 2011) made a research in this area and showed that there are several prototypes, some of which are motivated by animals such as the flapping wing model. Authors in (Andrew Gallagher, 2014) worked on surveillance UAV, the drone developed in the work was light and able to meet the weight requirement of 1.8 kg as well as transmitting video wirelessly. This work went beyond wireless transmission of video as it incorporated autonomous control. In (Matt Parker, 2011), the work was based on extending the transmission range of a UAV up to two miles radius, which was very effective. (Mitra, 2013) tried to explore the possibilities of an autonomous UAV, the designed model was able to detect red surfaces and land on them autonomously, but in this design, the UAV will also land on any red surface detected even the ones not meant for it to land on.

Though all drone models have several features that distinguish them from others, they are frequently characterised by their functionality or magnitude of the aerial vehicle that is needed for the execution of the mission (Anton Nakazawa, 2013). According to (Anton Nakazawa, 2013), the terms presently in use for classification of UAVs extend a range of models, from the high altitude long endurance (HALE), which is an airplane of 35 m or greater wing span, down to the Nano air vehicles (NAV) that could be just of 40 mm span. Classifications of UAVs according (Anton Nakazawa, 2013) are as follows:

i. *HALE – High altitude long endurance*. Over 15,000 m in altitude and more than 24hours fortitude. They can carry out enormously long-distance (trans-global) exploration and surveillance and can be equipped. They are frequently used by Air Forces from fixed air bases.

ii. *MALE – Medium altitude long endurance.* 5,000–15,000 m in altitude and 24 hours fortitude. The carry out missions that are like the HALE models but normally work at somewhat shorter distances, but still exceeds 500 km, and as well from fixed air bases.

iii. *TUAV – Medium Range or Tactical UAV* with range between 100 and 300 km, these UAV are smaller and operate inside simpler models than HALE or MALE and are functional also by land and naval forces.

iv. *MUAV or Mini UAV* – relates to UAV below a certain mass below 20 kg, but not as small as the MAV, these UAVs are proficient of being hand-launched and operate at distances that are up to about 30 km. These are, also, employed by mobile battle groups and predominantly for various civilian uses.

v. *Micro UAV or MAV.* The MAV was initially defined as a drone with a wing-span no greater than 150 mm. This has now been slightly relaxed but the MAV is primarily needed for missions in urban locations, predominantly inside structures. It is essential to fly gently, and preferably to hover and to 'perch' —i.e. to be able to stop and to sit on a wall or post. MAV are commonly anticipated to be thrown by hand and consequently winged models have very low wing loadings which ensure that they are very susceptible to atmospheric instability. This is the type of UAV employed in this work.

vi. *NAV – Nano Air Vehicles.* These are said to be of the size of sycamore seeds and used in groups for purposes such as radar misperception or imaginably, if camera, propulsion and control sub-systems can be made small enough, for ultra-short range observation.

vii. *RPH, remotely piloted helicopter or VTUAV, vertical take-off UAV.* These UAVs are proficient in vertical take-off. They are ordinarily proficient in vertical landing, and what can be sometimes are of even bigger operative importance, such as to hover flight through a task.

viii. *UCAV and UCAR.* These are UAVs which may fire weapons or even partake air-to-air battle. These are assumed the initials UCAV for drone battle air vehicle.

2. Design Methodology

2.1. System Description

The UAV used in this work comprise of: a QAV250 frame, four brushless D.C motors with two clockwise and two counter-clockwise 5030 propellers attached to them. The motors are controlled by ESCs (Electronic speed controllers) that vary

the brushless D.C motors' speed and direction of rotation. The UAV's primary parameters are controlled by a CC3D flight controller which is interfaced with an Arduino microcontroller. The arduino microcontroller is responsible for the communication between the UAV's artificial intelligence system and the ground control station. Two half-duplex transceivers of 400 MHz are used as full duplex and were used for transmitting data from the UAV to the GCS, as well as sending control commands to the UAV. At the ground control station, the RF transceivers are connected to the Java desktop application running on windows operating system via two serial ports. The Java desktop application is responsible for sending flight commands to the Arduino microcontroller residing on the UAV as well as retrieving the UAV information on the GCS. The design schematic is shown in figure 1, while the overall system flowchart is shown in figure 2.

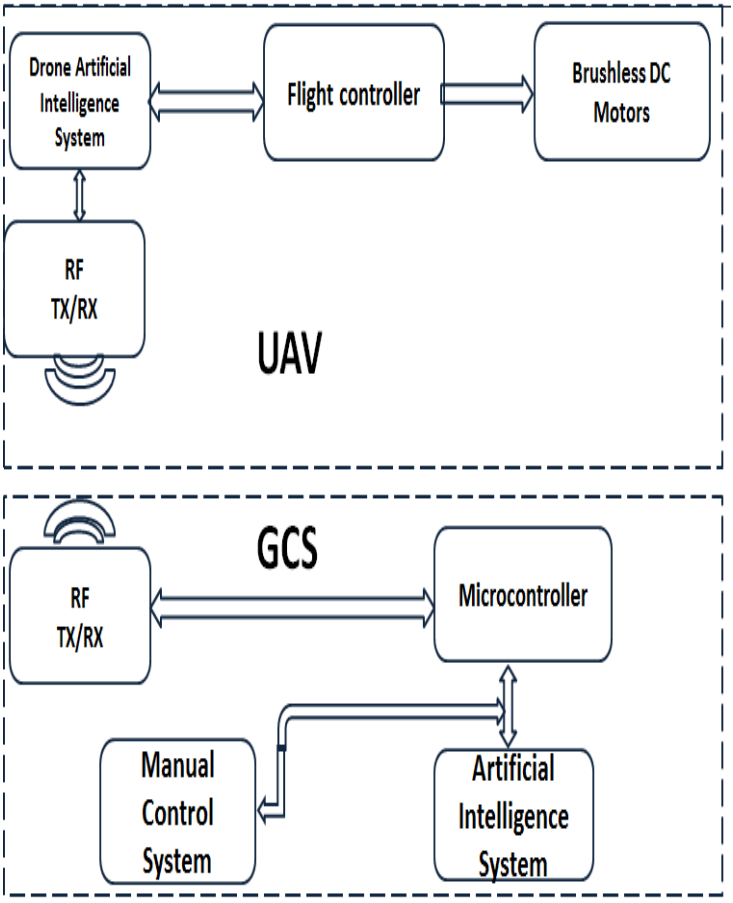


Figure 1: System design schematic

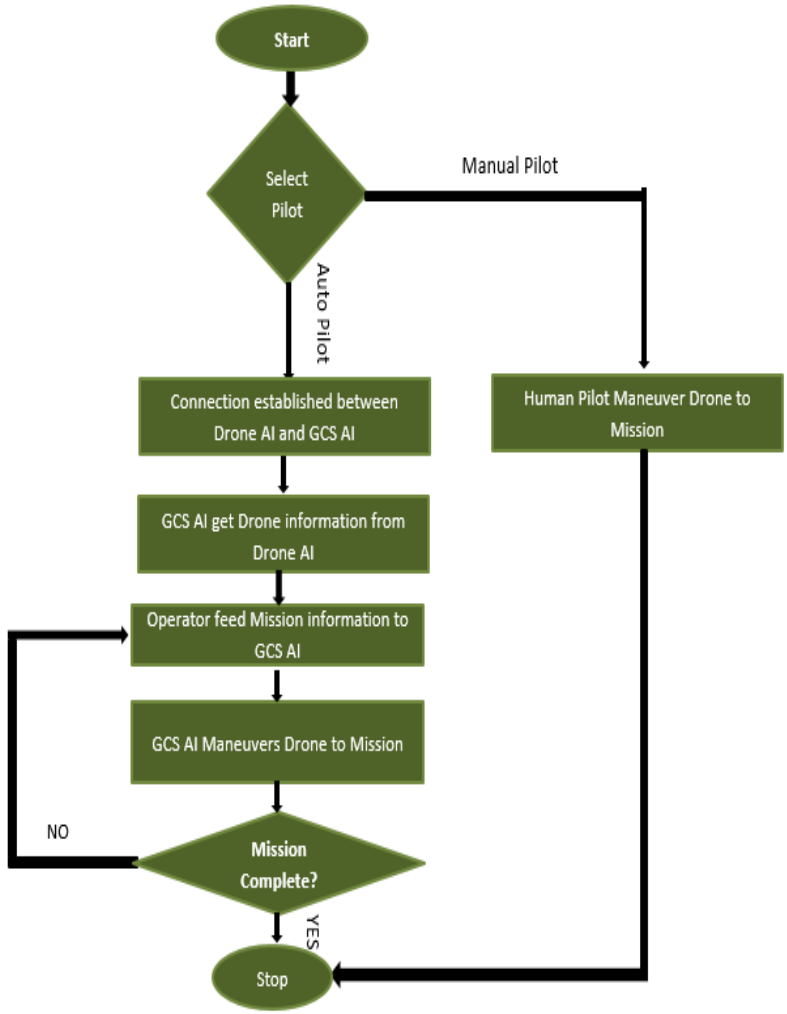


Figure 2: System flow chart.

3.2 Design Implementation

The implementation of the design was carried out in four phases which are: (i) the assemblage of the UAV chassis, (ii) the design of the UAV telemetry circuit, (iii) the design of the manual control system of the ground control station, and (iv) the design of the autonomous control phase of the ground control station.

(i) UAV Assemblage: the UAV chassis was selected after a thorough comparison of the available mini UAVs. The Lumenier QAV250 mini FPV Quadcopter Carbon Fibre edition chassis was the mini UAV chassis chosen, it is shown in Figure 3. Other parts of the UAV that made up the assemblage includes: Lipo Battery, Flight

controller, Brushless D.C. motors, Electronic speed controllers (ESCs), and NTSC FPV Camera.

(ii) UAV Telemetry Circuits: The telemetry circuit is the circuit responsible for sending data from the UAV to the ground control station, it sends data such as the GPS coordinates of the UAV, the number of satellites the drone's GPS is locked onto, the yaw, pitch, and roll values of the UAV. It also receives control commands from the UAV ground control station. The telemetry circuit was separated into two parts in this work. The first part is the downlink circuit that sends the UAV's information to the GCS, and the second part receives control messages from the GCS, both circuits transmit in the range of 400MHz – 500MHz.

Figure 3: QAV250 Chassis. *Source (www.lumenier.com)*

The downlink part of the telemetry circuit sends the following data to the GCS for proper planning and control of the drone when it is in autonomous mode: Number of satellites the UAV's GPS is locked onto, the Latitude position of the UAV, the Longitude position, the Yaw value, Roll value, and pitch values, the altitude of the UAV above sea level. The uplink part of the telemetry circuit receives the following control command messages from the GCS for the manoeuvring the UAV: Yaw, Pitch, Roll, and Throttle values. These control commands are received via the uplink RF receiver that operates at 410 MHz, which is a different from the downlink frequency in order as to avoid interference between the two signals. The message received by the RF receiver is decoded by the arduino Nano that it is connected to. The message is then converted to pulse position modulation signal which is sent to the flight controller for appropriate flight action on the UAV.

Pulse position modulation is a kind of signal modulation whereby Q message bits are encoded by transmitting a single pulse in one 2Q likely needed time shifts. This reoccurs every T seconds in a way that the transmitted bit rate is Q/T bits per second. It is suitable for fast and accurate transfer of data from the GCS to the flight controller. A complete PPM frame is about 22.5 ms. The signal low state is always 0.3 ms and it normally begins with a high state of about 2 ms with each

channel (8 channels) being encoded by the time of the high state which is given by the formula

PPM high state + 0.3 x (PPM low state = servo PWM pulse state (µs) (1)

(iii) GCS Manual Control Design: The Ground Control Station (GCS) manual control design implementation was done by interfacing an Ebyte RF transmitter with a Bluetooth module, which is connected to an android application that can be used to send control commands to the UAV. The Graphical User Interface (GUI) of the android application is shown in Figure 4. Once the GCS transmitter receives control message from the Bluetooth module, it forwards it to the receiver residing on the UAV and the command is effected on the UAV.

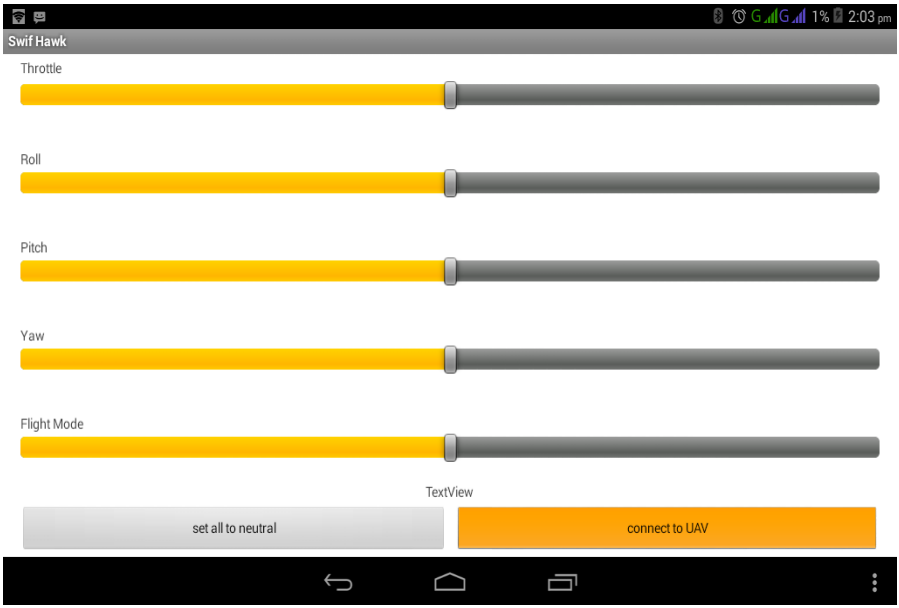


Figure 4. Android Application GUI design

(iv) ***GCS Autonomous Control Design:*** The driving algorithm works by arming the UAV at a reference point. The GPS coordinates of the drone are acquired from the UAV downlink telemetry circuit and the drone is now shown on the Map. The mission information is then fed into the GCS by the operator. The autonomous control starts by arming the UAV, rotating the UAV to the direction of the mission, and it is moved forward by the GCS until it reaches its destination. Any alteration in the path of the UAV is corrected by the flight controller and when the flight controller fails, it is corrected by the GCS autonomous control algorithm.

The GCS autonomous control was implemented with an expert system. An expert system is used to pass the knowledge of controlling the UAV to the UAV GCS java desktop application. Expert system was implemented by testing the UAV in manual mode in different locations. The latitude and longitude information received from the UAV through the downlink telemetry gives the GSC knowledge of where the drone is in space, and the distance to be travelled to execute the mission is known to the GCS. Any disorientation of the UAV is noticed by a difference in the yaw value of the UAV and it is then corrected by the GCS if the drone fails to correct it appropriately. The graphical user interface (GUI) of the Java desktop application is shown in figure 5.

3. 4. Results

Figure 5 shows a screen shot of the designed GUI of the developed Java desktop application. Figure 6 presents a picture of the manual flight test of the UAV, while figure 7 shows a picture of the drone as it is being armed for surveillance mission. Figures 8 – 11 show the responsiveness of the java desktop application during an autonomous missions.

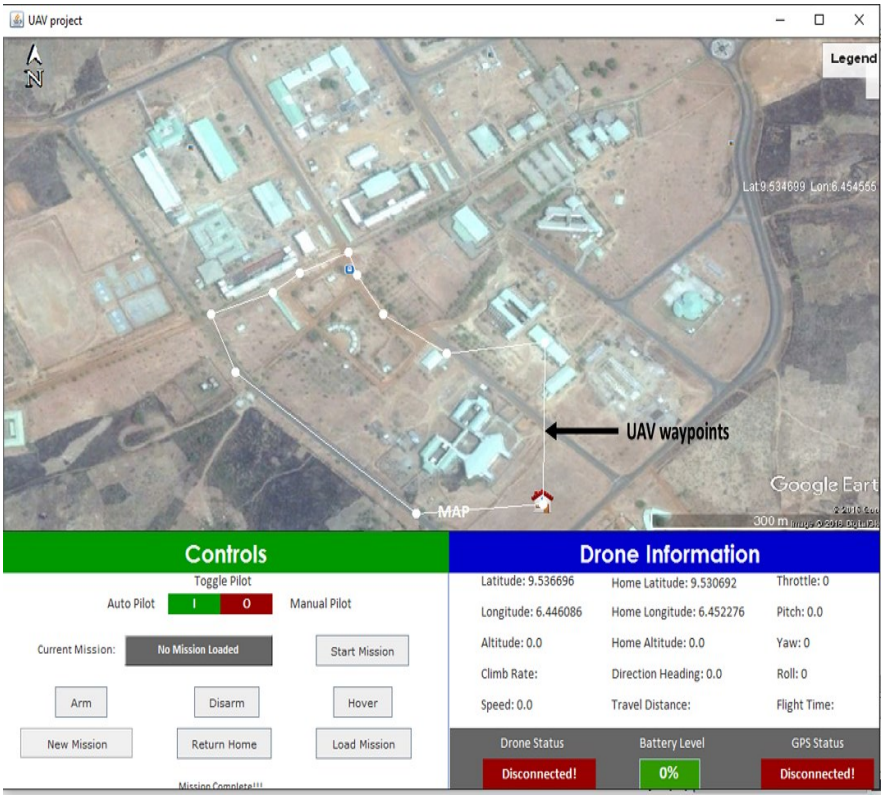


Figure 5: Java desktop application GUI

Figure 6: Manual Flight test (this test shows that the drone responds to the manual control)

Figure 7: Flight test

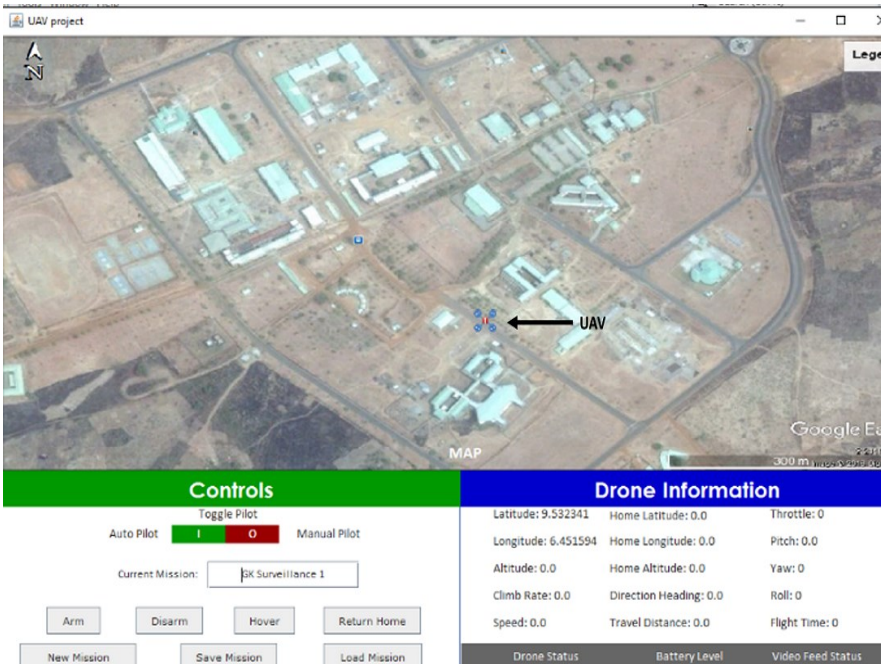


Figure 8: Mission Flight 1

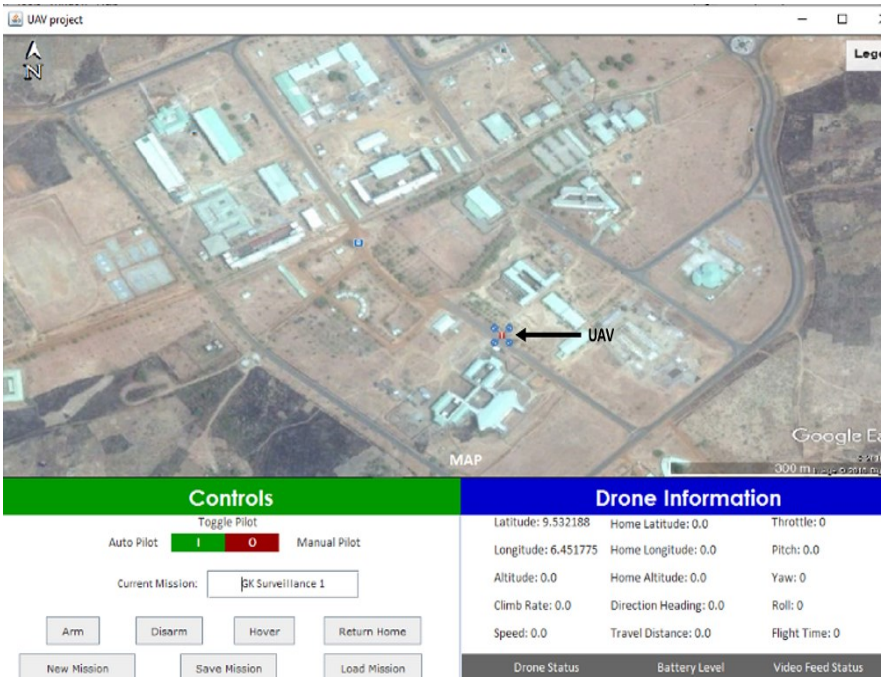


Figure 9: Mission Flight 2

Figure 10: Mission Flight 3

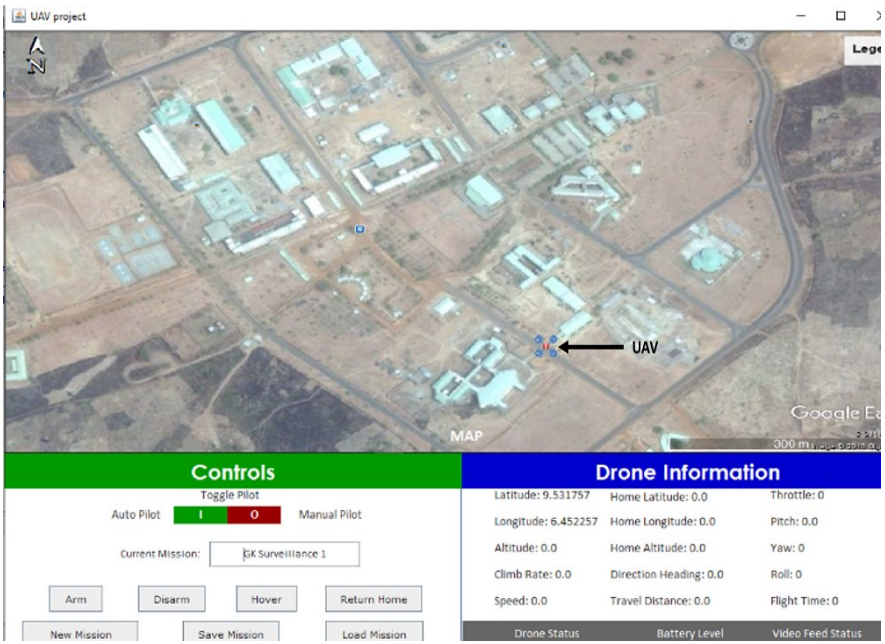


Figure 11: Mission Flight 4

5. Discussion

The figures 5- 11 show the results obtained during various tests of the designed system. The evaluation of the designed manual system was carried by testing the manual control of the GCS on the campus grounds. Figure 6 presents a picture of the manual flight test using the developed android application for manual controls, this means that the drone responds to the manual control system. Being on android platform, it means that with a smart phone, one could manually control the drone when necessary. The autonomous system was evaluated by using the developed Java application to carry out surveillance missions. Figure 5 shows a picture of the Java desktop application when a mission waypoints were being marked, the white dots joined by lines shows the different waypoints the UAV will be navigated to during the execution of an autonomous mission. Figure 8-11 shows different positions of the UAV on the map as an autonomous mission was being executed.

This designed model is a prototype that covered about 500 meters radius, range expansion is possible but will require an increase in the capacity of the battery for longer flight time to cover a greater distance. The range of the antenna used on the UAV will also be increased for better reception of signal to and from the GCS. The effective range of a UAV for surveillance is about 1km radius. For proper coverage of a large campus, multiple number of this micro UAVs will be deployed. Our future research will be focusing on communications between UAVs while on mission such that there could be transmission of control and mission information from one UAV to another to ensure coverage of a wider perimeter.

6. Conclusion

This paper presented the design and implementation of an unmanned aerial vehicle ground control station for campus surveillance with autonomous control, using the Federal University of Technology Minna as a case study. It explored the creation of an artificial intelligence system using an expert system approach. An alternate manual control with an android application was developed to manually control the UAV by interfacing it with a Bluetooth module. The android application could control the four basic parameters of an aerial vehicle; the roll, pitch, yaw and throttle. The Java desktop application was developed using Java Programming language. The autonomous control works by an operator marking out the waypoints the UAV will go through. Once the waypoints have been marked, the autonomous system will rotate the UAV to the direction of the waypoint and move it there, it will do this for all waypoints to the end of the mission. The system created, besides campus surveillance, can also be employed for military uses such as aid delivery purposes, surveillance, and it can also be used for commercial purposes such as goods delivery services.

Correspondence
Yusuf Y. Folawiyo
Department of Telecommunication Engineering
Federal University of Technology, Minna, Nigeria
Email: yusuffolawiyoyusuf@gmail.com

References

Andrew Gallagher, S. G. W. M. A. N. A. P. A. U., (2014). *Surveillance UAV,* Worcester: Worcester Polytechnic Institute,

Anon., (2009), *CC3D Boards.* [Online] Available at: www.cc3d.com/atom [Accessed April 2016].

Anon., (2015), *Arduino Forum.* [Online] Available at: www.ardunino.cc [Accessed March 2016].

Anton Nakazawa, B. X. J., n.d. *Quadcopter Video Surveillance UAV,* Victoria: University of Victoria.

Austin, R., (2010), *Unamanned Aircraft Systems.* West Sussex: John Wiley and Sons Ltd.

Chadly, J., (2015), *DIY Drones.* [Online]; available at: www.fpv.com [Accessed May 2016].

F.B. da Silva, S. S. M. C., (2007), *Design Methodology for Unmanned Aerial Vehicle(UAV) Team Coordination,* Cambridge: Massachusettts Institute of Technology.

Josh Bayliss, F. B. R. M., n.d. *Unmanned Aerial Vehicle,* Florida: Florida International University.

Matt Parker, C. R. G. B., (2011), *Quadcopter,* Colorado: Colorado State University.

Mitra, S., (2013), *Autonomous Quadcopter Docking System,* s.l.: Cornell University.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.152-163

Social Media Analytics: Indispensable Counterterrorism Framework to improve Intelligence Gathering towards Combating Terrorism in Nigeria

L. J. MUHAMMAD
Federal University, Kashere, Nigeria

I. A. MOHAMMED
Yobe State University, Damaturu, Yobe State, Nigeria

ABDULLAHI GARBA ALI
Bayero University, Kano, Nigeria

ABSTRACT Terrorism is one of the biggest threats to the peace and stability to Nigeria and has evidently threatens national security and socio-economic developments. The activities of the terrorists have created atmosphere of siege and desolation, which have impediment to trades, investments, peaceful co-existence and stability, as well as sustainable livelihood and development in Nigeria especially in North East Region. In order to combat this menace of terrorism in Nigeria, there is a need for proactive, reconnaissance, sustainable and invaluable intelligence gathering framework. This paper explain how Social Media Analytics which is one of the big data analytics can be harnessed an Indispensable Counterterrorism Framework to generate intelligence gathering of the activities of the terrorists or terrorist' groups such as membership, motivation, and operational modalities, foreclosing sources of inspiration, funding and recruitment through analysis of social media such as YouTube, twitter, istagram, Facebook among others. The paper recommended both open-source Social Media Analytics available such as Gephi, igraph, NetworkX, and that of commercial ones such as i2 Analyst, Sentinel Visualiser, among others, where all what the Nigeria Security Agents have to do is start entering data so as the analyze the activities of the terrorists activities on social media.

Keywords: Social Media Analytics, Big Data, Big Data Analytics, Social Media, Terrorist

1. Introduction

Terrorism is a sort of violence that uses one-sided violent approach against civilians. It also engages uneven violent confrontation against a stronger adversary, which could be a state or a group of states (Ekaterina, 2008). Terrorism is no doubt, it one of the biggest threats to global peace and stability in the contemporary times. Since the beginning of this millennium, the incidences of the terrorism have been on a steady rise worldwide. In Africa, indeed the manifestation of terrorism has been evident in some countries like Cameroon, Somalia, Mali, and Nigeria. In

Nigeria, the phenomenon has found expression in the emergence of Boko Haram insurgency in 2009. The Boko Haram appears to be the most visible security threat to the country and launched violent attacks on the Nigerian State, killing thousands of people and destroying public and private properties in different parts of the country (United States Embassy, Nigeria, 2014). Globally, the group is now labeled as one of the most deadly insurgent/terrorist groups in the world. Like most of the terrorist groups, it seems to defy several counter terrorism measures introduced by the Nigerian government (Ayuba, 2013).

Social media are computer-mediated tools or online platforms that allow individuals, companies and organizations to create, share, or exchange information, career interests, ideas, and pictures/videos in virtual communities and networks. (Buettner, 2016). The Social Media which have no doubt flooded every facet of human endeavor and are also being taken advantages by terrorists to propagate their extreme ideologides, recruit new members, train their members, communicate and conspire with international linkages, and raise funds among others. Arising from this background, this paper recommends that, the Social Media Analytics as an Indispensable Counterterrorism Framework to improve Intelligence Gathering towards Combating Terrorism in Nigeria through analysis of the Social Media platforms such as facebook, twitter. youtube among others so as to uncover the activities of Boko Haram such as membership, motivation, and operational modalities, foreclosing sources of inspiration, funding and recruitment.

1.1 Social media and Terrorism

Social media are web or mobile based interactive platforms through which individuals, communities and organizations to create, share, or exchange information, career interests, ideas, and pictures/videos. Social Media consists of websites such as YouTube (video-sharing), Twitter and Tumblr (micro-blogging), Facebook (social networking), StackOverflow (community based question and answering), Delicious (social bookmarking), online wikis, message boards and discussion forums. Social media platforms are highly participatory and collaborative in nature in which users can easily share content and post messages and comments (Kietzmann et al.,2011). Because of its cheap affordability, convenience and easy broad-reach of social media platforms, terrorists groups have increasingly use them to propagate their ideologies, spread their messages, recruit members, raise funds and gather intelligence. Today, almost 90 percent of terrorists' activities on the Internet take place via social media platforms (Gabriel, 2014). There are many instances where terrorists used social media for their activities, in 2008, Jose Pimentel was arrested for preparing bombs to use in attacking targets in New York City. Before his arrest, Mr. Pimentel had been active on-line. He ran a blog, held two YouTube accounts, and operated a Facebook profile, all dedicated to jihadi propaganda. In another instance, illustrates terrorist recruitment in the homeland via social networking, in December 2009 a group of five men in Washington, DC were arrested in Pakistan for attempting to join militants fighting along the border with Afghanistan. Later to become known as the Virginia Five, they were reportedly contacted by a Taliban recruiter through YouTube after one of the members of the group

praised an online video showing attacks on American troops (Department of Homeland Security, 2014)., Notorious terror group, Taliban has been active on Twitter since May 2011, tweeting under the handle @alemarahweb and has more than 7,000 followers. However, the account is currently suspended. In December 2011, it was discovered that the Somalia-based terror cell Al-Shabab was using a Twitter account under the name @HSMPress. Since opening on December 7, 2011, the account has amassed tens of thousands of followers and tweets frequently (Jenkins, 2016).

In Nigeria, however, terrorist groups particularly Boko Haram has increasingly turned to social media to communicate with and motivate their followers and supporters. The group has exploited social media, most notoriously Youtube and Twitter, to send its propaganda and messaging out to the world and to draw in people vulnerable to radicalization. An Arabic-language Twitter account purporting to be the official outlet for a Boko Haram media group called Al-Urwah al-Wuthqa was launched and immediately promoted by key pro-IS media operatives in 2015. Boko Haram is using exploitation skills to influence propaganda. For instance, the launch of the Boko Haram Twitter account has streamed several videos to show the public its success on the ground (BBC, 2015). According to a report by the US Department of Homeland Security, 2015 listed various terrorist uses of Facebook which include the following:

i. As a way to share operational and tactical information, such as bomb recipes, weapon maintenance and use, tactical shooting, etc.
ii. As a gateway to extremist sites and other online radical content by linking on Facebook group pages and in discussion forums.
iii. As a media outlet for terrorist propaganda and extremist ideological messaging.
iv. As a wealth of information for remote reconnaissance for targeting purposes.

2. Methodology

The methodology of the study is qualitative and exploratory in nature using only secondary data obtained through the review of related academic and non-academic publications. Qualitative research methodology involves studies that do not attempt to quantify their results through statistical summary or analysis and it typically involves interviews and observations without formal measurement. On other hand, exploratory research methodology is use to formulate a research problem for more precise or an in-depth investigation or for using hypothesis from an operational aspect. The methodology aids the researchers to clarify their understanding to a problem and asses the phenomennon in a new light. Moreover, exploratory research is used when problems are in a preliminary stage, when the topic or issue is new and when data is difficult to collect and it is effective in laying the groundwork that will lead to future studies like this research. (Saunders et. al., 2003).

The paper draws insights from scholarly exegesis and empirical historical evidences that dangerous terrorist group in Nigeria is Boko Haram, and it is a

greatest threat to Nigeria's national security. The group had been using the Social Media to publish a stream of propaganda, including several videos, even though there has been some disruption to its media activities following the suspension of the original account by Twitter in 2015. The paper discussed how big data analytics called Social Media Analytics such as Gephi, igraph could be used to generate intelligence gathering towards combating terrorism in Nigeria.

3.1 Big Data

Big data can be defined as data that exceeds the processing capacity of conventional database systems. Therefore, it is too big, moves too fast and doesn't fit the architecture of the structure of database (Edd, 2012). Data becomes big data when its volume, velocity, or variety exceeds the ability of typically IT systems to ingest, store, analyze, and process it. Many organizations have equipment and expertise to handle large quantities of structured data—but with the increasing volume and faster flows of data, they lack the ability to "mine" it and derive actionable intelligence in a timely way. Not only is the volume of this data growing too fast for traditional analytics, but the speed with which it arrives and the variety of data types necessitates new types of data processing and analytic solutions (Oracle, 2013).

There are six characteristic of big data which include the following:-

i. Volume: the quantity of generated and stored data for every second. This determines the value and potential insight- and whether it can actually be considered big data or not
ii. Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous.
iii. Variety: describes different formats of data that do not lend themselves to storage in structured relational database systems. Thus, it refers to the many sources and types of data both structured and unstructured
iv. Veracity: What is the provenance of the data? Does it come from a reliable source? It is accurate and by extension, complete. Thus, veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed.
v. Volatility refers to how long is data valid and how long should it be stored. In this world of real time data you need to determine at what point is data no longer relevant to the current analysis
vi. Validity means is the data correct and accurate for the intended use.

However, IBM data scientists break big data into four dimensions: volume, variety, velocity and veracity. Figure 1.1 info graphic explains and gives examples of each.

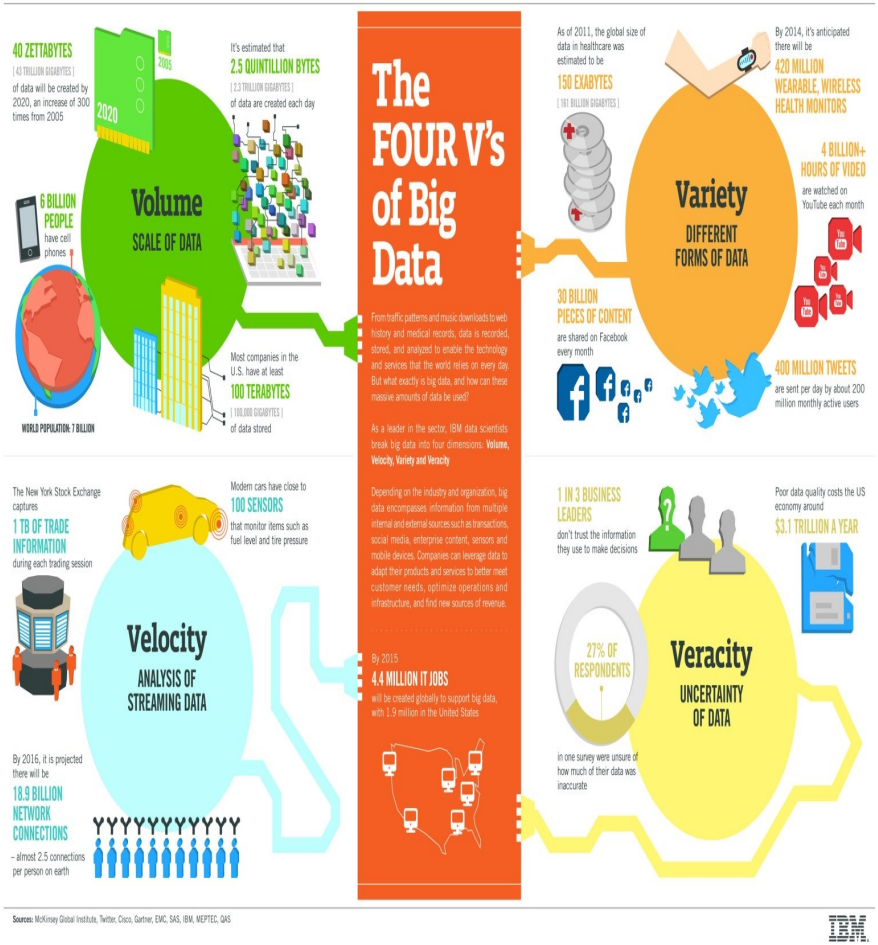


Figure 1.1: IBM Four V's of Big Data

The ways in which Big Data is revolutionizing business are much appreciated because it has transformed the global business landscape. Organization are analyzing hug volumes of diverse, fast changing data gain new insight that help run their business better and get advantage over the competitors. However, the big data may also be leveraged to transform security intelligence gathering landscape. Since the activities of terrorist in Nigeria are in an unstructured form then, the application of Big Data would help to deepen proper understanding of these data for better decision making (Nwanga et. al. 2014).

With Big data, Nigeria may have the ability to achieve superior value from analytics on data at higher volumes, velocities, varieties or veracities. With higher data volumes, Nigeria can take a more holistic view of its security intelligence gathering of the activities terrorists in the past, present and likely future. At higher data velocities, Nigeria can ground its counter terrorism decisions in continuously

updated, real-time data. With broader varieties of data, Nigeria can have a more nuanced view of the matter at hand. And as data veracity improves, Nigeria can be confident that it is working with the truest, cleanest, most consistent data.

3.2 Big Data Analytics

Big Data Analytics is the use of advanced analytic techniques against very large, diverse data sets that include different types such as structured/unstructured and streaming/batch, and different sizes from terabytes to zettabytes thus big data. (IBM, 2016) Big Data Analytics are big data platforms that provide a scalable, robust, and low-cost option to process large and diverse data sets; however, the key is not in organizing and managing large data sets but to generate insights from the data. This is where specialists such as data scientists come into the picture, interpreting and converting the data and relationships into insights. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analyzing shoppers' transactions and social and geographical data. Being able to process every item of data in reasonable time removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports. (Zeng et. al. 2010)

Nigeria Security Agencies can use big data analytics to uncover patterns in a wide variety of unstructured data generated from social media and associates the patterns with security intelligence gathering outcomes. It can be harnessed to detect unusual, interesting, previously unknown, or new patterns of terrorists' activities particularly boko haram in data generated from social media.

3.3 Social Media Analytics

Social media analytics is a term used to describe the process of gathering and consolidating unstructured raw data from social media channels like Facebook and Twitter and analyzing it to support planning and decision making. It can also be defined as an evaluating big data informatics tool and framework to collect, monitor, analyze, summarize, and visualize social media data so as to extract useful patterns and intelligence (Zenge et. al. 2010). Unstructured data is everything from social media posts and sensor data to email, images and web logs and it is growing at an unprecedented pace and increasing exponentially.

Social Media analytics is used to process, measure, analyze and interpret the unstructured data generated from interactions and associations among people, topics and ideas discussed on social media and converting them into insights, which help may some businesses in identifying areas of their customer satisfaction or any customer grievance for a product. Likewise, Social Media Analytics can be used to tap and analyze activities of the terrorists (Boko Haram) on social media platforms, convert them into insight, which may assist Nigeria Security Agencies to identify terrorist' locations and generate early warnings and real-time alerts to improve situational awareness. Therefore, it may to also provide Nigeria Security Agencies with huge amounts of critical, relevant information, including specific events and

occurrences, relationships between terrorists (Boko Haram) and their respective affiliated organizations and even recruitment tactics.

4. Social Media Analytics: Indispensable Counterterrorism Framework

Today, there are almost 2.3 billion active social media users in the world; Social media giant Facebook has nearing 1.65 billion users, thus almost a quarter of the world's population and Twitter alone has reported to have 100 million people log in to its site every day. And these numbers are set to increase as social media outlets compete for more users by improving their service. Twitter recently announced new ways for people to share richer content by rolling out a series of changes designed to make the service easier and more intuitive. However, Social media are proving to be valuable tools for terrorists or terrorist groups' internal communication and for reaching larger audiences around the world. In Nigeria, the notorious terror group known as Boko Haram increasingly relies on social media to communicate its message, and builds links with other Islamist organizations. . In January, 2015, the group launched its official Twitter feed—a sign that it is learning from the likes of Al Qaeda and the Islamic State (ISIS) and has repeatedly for over 3 years embarked upon massive propaganda using social media to demoralize the Nigeria Army.

Accessing and analyzing voluminous and unstructured raw data generated from various social media' platforms would be a potential source of intelligence gathering towards improving national security in Nigeria. Therefore, Nigerian Security Agencies may deploy and harness Social Media Analytics to analyze the information that is produced and exists through the social interaction of Nigerians especially among suspected terrorists or terrorist groups on various social media networks so as to improve their intelligence gathering. However, understanding the nature, relationships and content of social media is essential for effective situational awareness, intelligence gathering, counter terrorism and insider threat scenarios. Social Media Analytics can be an indispensable counterterrorism framework that would enable proactive monitoring, analysis, and engagement through extremist social media networks and their associated digital properties such as website. This may help to have a treasure-trove for information on terrorists or terrorist' groups either directly posted or indirectly referred to.

There are many open-source Social Media Analytics available such as Gephi, igraph, NetworkX, SNAP among others which require a fair amount of software development and data analysis and that of commercial ones such as i2 Analyst, Sentinel Visualiser, SilentRunner Sentinel among others, where all what the Nigeria Security Agents have to do is start entering data to analyze the connections between a suspect terrorist and individuals in its relationship network quantitatively and qualitatively, either through numerical or visual representation. The networks can consist of anything from families (immediate and extended); professional links (office colleagues or the suspect's business-cards folder); membership on networking sites such as Facebook, LinkedIn, and Twitter; social circle; mobile phone records; and various others.

Nigeria Security Agencies would be able to analyze social media by harnessing Social Media Analytics for intelligence gathering towards combating terrorism in Nigeria need to, among other things:

- Detect specific, credible terrorist' threats or monitor adversarial situations.
- Geospatially locate terrorists or terrorist groups and analyze their movements, vulnerabilities, limitations, and possible adverse actions.
- Predict likely developments in the situation or future actions taken by terrorists or terrorist groups (by conducting trend, pattern, association, and timeline analysis).
- Detect instances of deception in intent or action by terrorists or terrorist groups for the explicit purpose of misleading law enforcement.
- Develop domain assessments for the area of interest (more so for routine scenarios and special events).

5. Implications for Nigerian Security Policy and Decision Making

For effective harnessing of the Social Media Analytics to improve Intelligence Gathering towards Combating Terrorism in Nigeria the following challenges need to be addressed:-

i. A big data department in every security agencies, need to be established and be saddled with the responsibility of harnessing the social media analytics to uncover the activities of terrorists or terrorist' groups.
ii. A comprehensive and total synergy relational ship among those departments for collaborative and effective intelligence gathering need to be established, maintained and monitored by the Office National Security Adviser.
iii. Highly Qualified Personnel need to be recruited or trained
iv. Uninterrupted power supply and internet service need to be provided at each department concerned
v. All the necessary relevant laws, regulations and policies need to be amended or formulated for effective startup of the usage of the analytics toward improving national security.

6. Discussion

Terrorism is of the greatest threat to national security and peaceful co-existence and it spreads fast like wild fire ready to consume and object close to it. Terrorists or Terrorist' Group had since used the various social media to communicate with and motivate their followers and supporters. Therefore, with big data, Nigeria may have the ability to achieve superior value from Social Media Analytics on unstructured data at higher volumes, velocities, varieties or veracities generated from various social media towards improving national security. The paper expatiates how Social Media Analytics can be used as an Indispensable Counterterrorism Framework to collect, monitor, analyze, summarize, and visualize unstructured data generated from interactions and associations among terrorists or terrorist' groups so as

to improve intelligence gathering towards combating Terrorism in Nigeria. Harnessing the Social Media Analytics for intelligence gathering by Security Agencies in Nigeria towards combating terrorism may help to achieve the following:

i. Capture terrorists' data from various social media to understand their attitudes, opinions and trends and so as to prepare and plan for counter actions.
ii. Predict terrorists' behavior and improve counter actions by anticipating terrorists plans so as to recommend the best counter actions
iii. Create customized campaigns against terrorism that resonate with social media participants.
iv. Identify the primary terrorist recruiters within specific social network channels and target them with unique counter-terrorist action

7. Conclusion

Social media offer terrorists or terrorist' groups enormous reach of audience that knows no borders or nationalities. With the help of the social media, terrorists or terrorist' groups especially Boko Haram get unlimited access to impressionable young people sharing similar ideas. In order to combat the terrorism in Nigeria, there is need for proactive, reconnaissance, sustainable and invaluable intelligence gathering via analysis of social media for effective counter-terrorism. However, the paper illustrate how Social Media Analytics can be used an Indispensable Counter-terrorism Framework to improve Intelligence Gathering and a complementary source of intelligence gathering to Security Agencies in Nigeria towards combating terrorism.

8. Limitation

Nigeria, like many countries in the world, faces terrorism challenge, which becomes one of the greatest threats to its national security. However, the study only explores how big data analytics called social media analytics can be harnessed by Nigeria Security Agencies to generate intelligence gathering of the activities of notorious terrorist group known as Boko Haram on various social media towards improving national security. The study only also recommends both commercial and open source social media analytics that can be harnessed.

Correspondence
L. J. Muhammad
Mathematics and Computer Science Department
Federal University, Kashere
Email: lawan.jibril@fukashere.edu.ng

I .A. Mohammed
Department of Computer Science
Yobe State University, Damaturu, Yobe State, Nigeria
Email: ibrahimsallau@gmail.com

Abdullahi Garba Ali
Computer Science Department
Bayero University, Kano
Email: jgewel@yahoo.com

References

Ahmed, A,, Ravichandran, M., Kamarulnizam A., Sity, D. (2015) Managing Terrorism and Insurgency through African Traditional Institutions: The Role of Kano Emirate Council –Nigeria, Mediterranean Journal of Social Sciences MCSER Publishing, Rome-Italy Vol 6 No 4 S2 July 2015 ISSN 2039-2117 (online) ISSN 2039-9340 (print) pp. 126-136.

Ayuba, I. (2013). Terrorism: A New Challenge to Nigeria`S Stability in the 21st Century. International Affairs and Global Strategy ISSN 2224-574X (Paper) ISSN 2224-8951 (Online) Vol.12, 2013.
Awareness (2011) Actionable Social Analytics: From Social Media to Business Insights; available online: http://www.socialmediopolis.com/resources/white-papers/437-actionable-social-analytics-from-social-media-metrics-to-business-insights.

BBC (2015) Is Islamic State Shaping Boko Haram Media? BBC News, 4 Mar. 2015. available online: http://www.bbc.co.uk/news/world-africa-31522469.

Buettner, R. (2016). Getting a Job via Career-oriented Social Networking Sites: The Weakness of Ties. 49th Annual Hawaii International Conference on System Sciences. Kauai, Hawaii: IEEE. doi:10.13140/RG.2.1.3249.2241.

Department of Homeland Security (2015) "Terrorist Use of Social Networking Facebook Case Study," Public Intelligence, December 5, 2010. http:// publicintelligence. net/ufouoles-dhs-terrorist-use-of-social-networking-facebook-case-study. Retrieved February 10, 2016.

Dark, C. (2011). "Social Media and Social Menacing". Foreign Policy Association. Retrieved April 5, 2016.
Edd, D. (2012)Big Data Now,O'Reilly Media, Inc. Printed in the United States of America. IBM (2016) What is Big Data Analytics? https://www-01.ibm.com/ software/data/infosphere/hadoop/what-is-big-data-analytics.html Accessed date 19th June, 2016.

Ekaterina, A. S. (2008). Terrorism in Asymmetrical Conflict: Ideological and Structural Aspects. New York, NY: Oxford University Press.

Gabriel W. (2014) New Terrorism and New Media. One Woodrow Wilson Plaza 1300 Pennsylvania Avenue, N.W. Washington, DC, USA 20004-3027 202-691-4000.

Jenkins, B. (2016) “Is Al Qaeda’s Internet Strategy Working?”. Retrieved JUly 5, 2016.

Kietzmann, J., Kristopher, H., Ian P, M. & Bruno, S. S. (2011). "Social media? Get serious! Understanding the functional building blocks of social media". Business Horizons. 54: 241–251. doi:10.1016/j.bushor.2011.01.005.

Nwanga, M. E., Onwuka, E. N., Aibinu, A. M. & Ubadike, O. C. (2014). Leveraging Big Data in Enhancing National Security in Nigeria. International Journal of Knowledge, Innovation and Entrepreneurship Volume 2 No. 2, 2014, pp. 66—80.

Okoli, A. C. & Iortyer, P. (2014) Terrorism and Humanitarian Crisis in Nigeria: Insights from Boko Haram Insurgency, Global Journal of HUMAN-SOCIAL SCIENCE F POLITICAL SCEINCE, Volume 14 Issue 1 Version 1.0 Year 2014, Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc. (USA), Online ISSN: 2249-460x & Print ISSN: 0975-587X.

Oracle (2013) Big Data Analytics, Advanced Analytics in Oracle Database. An Oracle White Paper March 2013

Oremus, W. (2011). "Twitter of Terror". Slate Magazine. Retrieved April 7, 2016.

Paul, C. Zikopoulos, C. E., Dirk, D. Thomas, D. & George, L. (2012) Understanding Big Data: Analytics For Enterprise Class Hadoop and Streaming Data; available online: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=IML14296USEN.

Prerna, C. (2011) Telecom White Paper Social Analytics: Tata Consultancy Services (TCS) http://www.ibmbigdatahub.com/infographic/four-vs-big-data accessed date 19th June, 2016.

Sam, C. (2013). Big Data Fuels Intelligence Driven Security, Rapid Growth in Security Information Create New capabilities to Defend against the Unknown.

Saunders, M., Lewis, P., Thornhill, A., (2003) Research Methods For Business Students, Second Edition, UK, Financial Times, Prentice Hall.

United States Embassy in Nigeria (2014). Boko Haram and US Counterterrorism Assistance to Nigeria. www.nigeria.usembassy.gov/mobile/factsheet_05142014.html. Retrieved on August 18, 2016 Chomsky, N. (2001). 9-11. New York: Seven Stories Press.

Zeng, D., Chen, H., Lusch, R., & Li, S-H. (2010). Social media analytics and intelligence. IEEE Intelligent Systems, 25(6). http://www.internetlivestats.com/internet-users-by-country/ Accessed date 24h July, 2016

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.165-175

Predicting the Time Lag between Primary and Secondary Waves for Earthquakes using Artificial Neural Network

OGBOLE C. I.
Federal University of Technology Minna, Nigeria

MUHAMMED S.
Eastern Mediterranean University, Famagusta, North Cyprus

MUHAMMAD E.B., FOLORUNSO T. A. & NUHU B.K.
Federal University of Technology Minna, Nigeria

ABSTRACT Seismic waves experienced prior to earthquake are the primary and the secondary waves. This paper investigates the time lag after the primary wave before the occurrence of the secondary (destructive) wave. The aim is to allow for necessary warning signals and safety steps to be taken prior to the impending disaster. Seismometer records from previous earthquakes were used in this investigation, putting into consideration the time lag between the primary and secondary waves. Other parameters considered include the magnitude, the epicenter distance, the seismic station's distance and the direction (in azimuths). Consequently, a prediction model was developed from the derived data using Artificial Neural Network (ANN). Data obtained from earthquakes of magnitude 6.0 to 7.0, based on Richter's scale, was used to train the ANN. The results therein showed high performance, with regression values greater than 0.9 and root mean squared errors of 0.1003-0.1148 for the most satisfactory architecture. The final results showed that the developed ANN model achieved a high performance, hence, adequate for this type of application.

Keywords: Earthquake, seismic waves, fault lines, hypocenter, epicenter, neural networks

1. Introduction

Earthquakes are the result of plate tectonics, and it occurs when energy is released in the earth crust resulting in seismic waves. Earthquake occurrence varies spatially and its prediction has been a goal of mankind for millennia (Wiemer 2000). The basics in earthquake prediction begins with measuring the changes in distance (geodetic), also a creep-meter can be used; this is a device to measure movement across a fault line. In (Liu et al. 2013), a measure of the change in slope on earth's surface using a tilt-meter is considered, this inclinometer measures small changes on the ground and on physical structures. Other changes in the properties of physical structures can also be measured; solid rocks are highly resistive but under ex-

cessive strain, they develop cracks and shatter, thus allowing water to percolate through, resulting to increase in its conductivity (Furen 2010). Seismologist uses various tools for earthquake prediction analysis, the most common of this is the seismograph machine (or seismometer) which detects and records seismic waves. For a region's seismicity factors considered includes: the air ionization around rock surface which increases prior to earthquakes (Freund et al. 2009), the geology of the area, location of faults, the earthquake history of the area, the previous earthquake intensities and evidence for recent fault movement.
Further useful works give reviews on various animal behavioral anomaly (Bhargava et al. 2009), the possibility of prediction from physical climatic elements using neural networks (Maitha et al. 2011), evidence on the relationship between seismic electric signal (SES) with earthquake focal mechanisms (Varotos & Alexopoulos 1982), and the use of wireless sensor networks in earthquake monitoring (Azzam et al. 2011).

In addition, (Kirschvink 2000; Adi & Schnytzer 2011; Grant & Halliday 2010) considered behavioral activities (seismic-escape response) put up by some animals in response to the precursor to be helpful in earthquake prediction. These animals through natural selection are forced to develop anticipatory mechanism for predicting possible natural disasters. Although, the issue with the belief that certain animal do anticipate earthquakes is that it is poorly supported by evidence (Grant & Halliday 2010).

In another related approach, called the VAN method, coined from the initials of the three Greek scientist (Varotsos, Alexopoulos and Nomicos). They found out that seismic electric signals which results to variations in the earth's electric field occurs prior to an earthquake. Depending on this SES's types, the earthquake can be predicted to occur within days to weeks (Varotsos et al. 2006). This has attracted a high level of debate, which has been majorly on how to distinguish between similar electric signals from natural occurrence like thunderstorms and other man-made disturbances (Moustra et al. 2011). In (Kai Tan & Xiushan Cai 2010), data from earthquakes of magnitude 3.5 and greater collected from 1970 to 2008 in Yunnan region (22-28^{o}N, 98-104^{o}E) were used to predict earthquakes in 1999-2008 and verified using the Support Vector Machine (SVM).

The mitigation to earthquakes is in its prediction which could be long term, medium or short term. In short-term prediction, specific information of the earthquakes time and location is given within minutes, weeks, or months and are therefore very useful for public safety and evacuation (Uyeda et al. 2009). This method of prediction has attracted extensive research lately. Most earthquake studies in the past were basically on understanding the basics; its occurrence and extent of damage. Prediction study on this area only started in the 1980's (Zuniga & Wyss 1995).

Seismograph is used to detect and record seismic waves and the seismic measurements are the basis for short-term prediction (Uyeda et al. 2009). There are two basic types of seismic waves; the primary wave (P-wave) and secondary wave (S-wave). Though a third wave exists that is called the surface wave (this is the resulting wave formed when the P & S-waves combines at the surface). This research work presents a novel approach which focuses on the prediction of the arrival time of the S-wave after a P-wave has been detected. It further shows how the time lag

between these two wave forms can be computed using Artificial Neural Network (ANN).

This work uses a supervised learning strategy in ANN. The supervised learning can also be termed 'learning with a teacher'. Illustration for this kind of learning uses a teacher. The teacher is believed to have full knowledge of the system, this knowledge is given in a set of input-output mapping, but the neural network does not know this. Thus, the knowledge of the system is transferred from the teacher to the neural network to a certain degree measured with statistical tools. While in contrast, unsupervised learning looks at how systems learn to represent input pattern in ways to reflect the structure of the entire collection of the inputs. For this method of learning, there are no explicit target outputs associated with the inputs.
The remainder of this paper is organized as follows: Dataset description in section II, section III shows the ANN Model Development. The result and its discussion is presented in section IV. Finally, in section V the conclusion and limitations are presented.

2. Dataset Description

A number of factors are known to influence the occurrence of earthquake and these factors have varying effect on the strength as well as impact of the quake. The sampled seismic data set used for this study was obtained from the World Data Center for Seismology in China, measured from January, 2012 to August, 2014. The dataset contains varying values of five key parameters namely; the distance, the azimuth, the measured magnitude, the depth and the time lag between the primary and the secondary waves. The range of values for the parameters are as depicted in Table 1.

Table 1: Dataset description and their value ranges

Parameters	Value range
Distance (degrees)	8.8-164.6
Azimuth (degrees)	0-358
Measured Magnitude (Richter)	6.0-7.0
Depth (meters)	5,000-80,000

Characteristically, the distance (D) measured in degrees, is the representation of the distance from the earthquake's source and the seismological station (point of observation); the Azimuth (Az) is a clockwise measurement referenced from the earth's true north, also in degrees; the Magnitude (M) is the measure for the earthquake's primary wave as recorded by the seismograph at the station; the Depth (D) is the distance from the earthquake's hypocenter (wave origin) to the epicenter, and finally, the time lag (Ts-Tp) is the time difference between the arrival of the first primary wave and the first secondary wave signals.

A total of 86 earthquake cases were sampled from 1,478 stations, and they were all of the magnitude range of 6.0 to 7.0 (Strong earthquakes) on the magnitude scale. The choice of this magnitude range is because of the frequency of occurrence amongst the earthquake classes that poses serious threats to lives and properties. Table 2 shows the classes of earthquakes base on the magnitude and their effects.

Table 2: Earthquake magnitude, effect and annual frequency (source UPSeis)

Magnitude	Earthquake Effect	Annual Frequency
8.0 or more	Can totally destroy communities near the epicenter	One in 5-10 years
7.0 – 7.9	Causes serious damage	20
6.1 – 6.9	May cause a lot of damage in very populated areas	100
5.5 – 6.0	Slight damage to buildings and other structures	500
2.5 – 5.4	Often felt, but only causes minor damage	30,000
2.5 or less	Usually not felt, but can be recorded by a seismograph	900,000

3. Artificial Neural Network (ANN) Model Development

In this work, the design of an ANN model that will give a significantly high level of generalization (prediction) for the time lag between the primary and secondary earthquake waves is presented.

The structure of the neutral network is as shown in Figure 1, this depicts the perceptron process, with ,.. representing the input parameters.

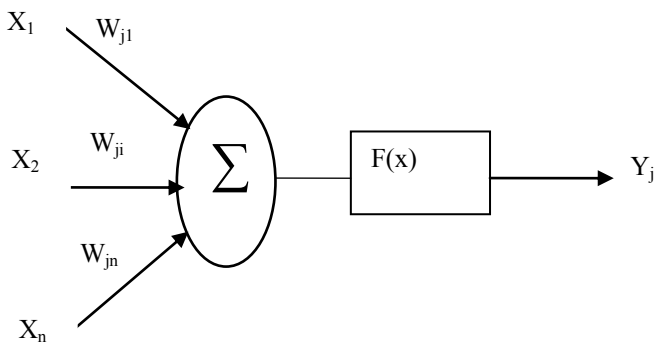


Figure 1: A detailed perceptron process

The input neurons buffers the inputs $\times_i$ ($\times_1$, $\times_2$,..$\times_n$) to the neurons in the hidden layer. Summation of inputs is done in each neuron j of the hidden layer, where these inputs are weighted with the inter-neuron connection weights w_{ji} and the output y_{ji} is computed as a threshold function of the sum using equation (1).

$$y_{ji} = f\left(\sum_{i=1}^{n} w_{ji}\, x_i\right) \tag{1}$$

In the Multilayer Perceptron (MLP) structure, the threshold function is a continuous derivative. The goal is to minimize the error function, which is achieved by finding the squared error of the network. Equation 2 gives how the training weights are adapted:

$$\Delta w_{ji} = \eta \delta_j x_i \tag{2}$$

Where η is the η learning rate; it determines the level of modification to the link weights (w) and node biases base on the change rate and direction.
A "momentum" term (μ) is added to help the network skip over the local minima and successfully reach the global minimum, while still maintaining the change rate and direction. This is adopted into the weight update equation as shown in Equation (3), while the change rate for both the output and hidden neurons are computed as given in Equations (4) and (5) respectively:

$$\boldsymbol{\Delta w_{ji}(t+1) = \eta \delta_j x_i + \mu \Delta w_{ji}(t)} \tag{3}$$

For the output neurons,

$$\delta_o = \left(\frac{\partial f}{\partial net j}\right)\left(y_j^{(t)} - y_j\right) \tag{4}$$

For the hidden neurons,

$$\delta_{\downarrow} h = (\partial f / \partial net j \;)(\; \Sigma_{\downarrow} q \cong w_{\downarrow} ji \;\; \partial_{\downarrow} j) \tag{5}$$

And training continues until the error function reaches a target minimum value.

The parameters considered for this prediction work are; the distance (D), the azimuths (Az), the measured magnitude (M), the depth (Ep), the measured time lag (Ts-Tp).

The design for a neural network on MATLAB adopts certain systematic procedures. In general, these five basic steps are followed; importing the data, preprocessing data, building the network, training the network, testing the network and evaluating the system's performance. The data is first grouped in two sets; the training set and the testing set. In the preprocessing stage, normalization of the data set is applied. This is necessary considering the range of values of the parameters which largely varies. The design program for this work follows the flow chart presented in Figure 2.

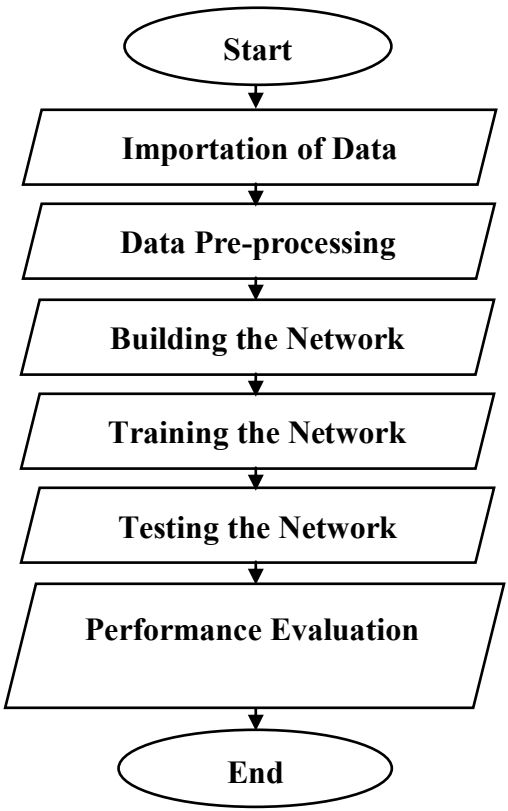


Figure 2. Flowchart for developing MLP using MATLAB

In the training, the network is taught how to generalize for the presented data set. These data sets consist of input-output pairs. The neural network learns from the input and updates its weight; this is why it is termed a supervised learning, since the neural network is taught what the output should be from the input set intro-

duced to it. Figure 3 shows the neural network MLP architecture, presenting the network flow from input to the output for the design.

The architecture used has four (4) inputs neuron, one (1) hidden layer with hidden number of neurons varied from 3-7, 10 & 20. Each of this architecture was trained and tested with a learning rate (η) of 0.1 to 0.9. The network was observed while varying the number of neurons in the hidden layer, the momentum constant and also the learning rate for the 9 different structures, and the best performing structures are selected. The training is stopped whenever any of the network's performance parameter is met.

After training is completed, the network is tested with unseen data and the output compared with the target (measured result). This is to check how well the network can generalize (predict output from the unseen inputs). Checking the performance is carried out using statistical measures on the obtained results: the root mean square error (RMSE), the mean absolute error (MAE) and the mean bias error (MBE) are computed with the formulas as given in Equations (6), (7), and (8) respectively:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(t-O)^2} \tag{6}$$

Where, n is the number of samples. t is the target output (measured value), and O is the network output (predicted value).

$$\mathbf{MAE} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}|\mathbf{t}-\mathbf{O}| \tag{7}$$

If MAE=RMSE, it means all the errors in the sample are of the same magnitude.

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(t-O) \tag{8}$$

For every simulation, the computed values are recorded in excel and used to evaluate the system's performance. The performance of the trained network on testing data is the focus; it is most important measure of the training success.

4. Results and Discussion

Several network variations were investigated. The computed data for the network architecture with satisfactory performance are given in Tables 3 and 4 for this section. Table 3 gives the best result for the different architectures while varying both momentum constant and the learning rate for the eight different architectures.

Architec-ture	*Test error statistics at μ=0.01*			*Test error statistics at μ=0.001*		
	RMSE	*MAE*	*MBE*	*RMSE*	*MAE*	*MBE*
4-2-1	*0.1241*	*0.0955*	*0.1241*	*0.2139*	*0.154*	*-0.0247*
4-3-1	*0.1126*	*0.0925*	*-0.0243*	*0.2009*	*0.1474*	*-0.0024*
4-4-1	*0.1072*	*0.0937*	*-0.0079*	*0.2132*	*0.1526*	*-0.0236*
4-5-1	*0.1003*	*0.0867*	*-0.0661*	*0.2173*	*0.1555*	*-0.0222*
4-6-1	*0.1065*	*0.0899*	*-0.0054*	*0.2193*	*0.1468*	*-0.0054*
4-7-1	*0.1049*	*0.0915*	*-0.0553*	*0.2175*	*0.1387*	*-0.0318*
4-10-1	*0.1188*	*0.0998*	*-0.0439*	*0.2142*	*0.1497*	*-0.0252*
4-20-1	*0.1132*	*0.0971*	*-0.0414*	*0.224*	*0.1576*	*-0.0047*

Table 3 Statistical performance evaluation for the different ANN architectures

In Table 3, the notation 4-2-1 implies 4 input neurons, 2 hidden neurons and 1 output. The RMSE and MAE values are found to be better using a momentum constant of 0.01 (μ=0.01). Also from the table, we can deduce that the best overall RMSE value was at five (5) hidden neurons with a momentum constant of 0.01, while the test using the three (3) hidden neurons gave the best RMSE with the momentum constant of 0.001.

Another observation from Table 3 is that the RMSE and MAE values were considerably consistent from N=2 to N=7. This indicates that optimal performance of the network was within this range and then much increase from N=10. The performance plot of the training also indicates a good training. The performance plots selected for N=5 for μ=0.01 is seen in Figure 3.

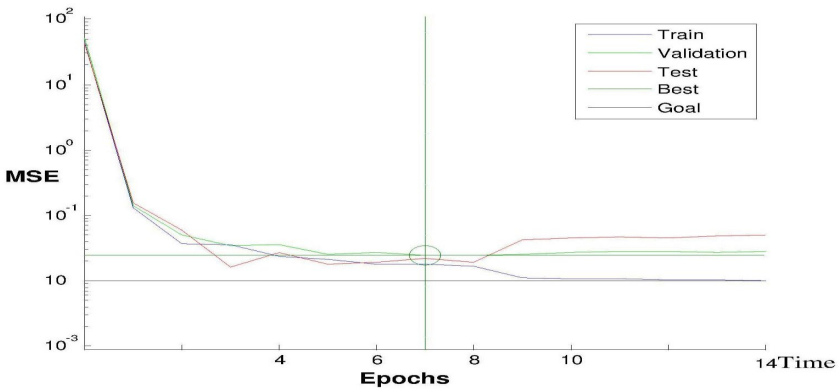
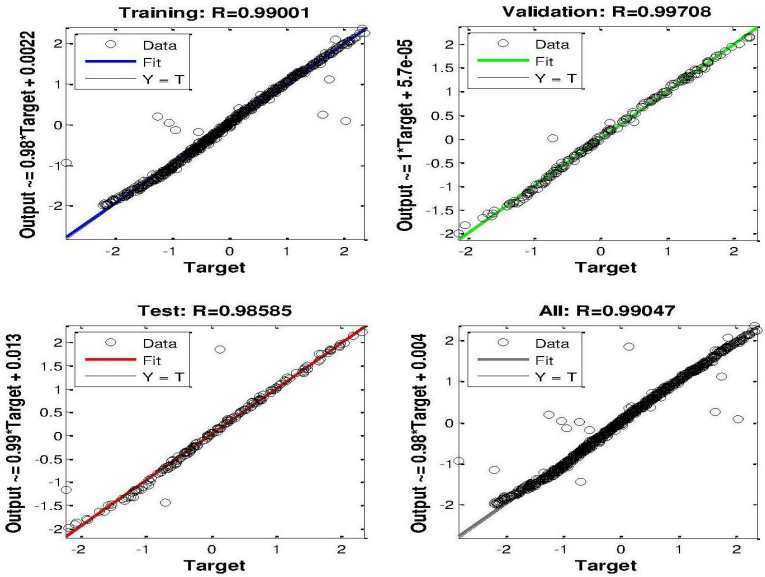


Figure 3. The performance plot for N=5 for μ=0.01



The plot as seen in Figure 3 shows that the best validation was reached at epoch seven (7), even though the training still proceeded for 7 more epochs and with a mean square error of 0.024839. Also the test plot is noticed to be similar with that of the validation. Figure 4 gives the linear regression between the output and the target for the training, validation, test and all three results combined.

Figure 4. Schematic of the regression plot at N=5 for μ=0.01

The R-values shown in Figure 4 are all greater than 0.9. This is an indication of a very good fit for the training data and it shows how very close the output of the network and the target (measured) values are. The training stops whenever any of the performance goal is met. Also in Figure 5, it is observed that the network gave a better result (i.e. a lower root mean square value), for the testing when a momentum constant (μ) of 0.01 than when using 0.001. For the different training using varying number of hidden neurons, it is observed that this trend is maintained.

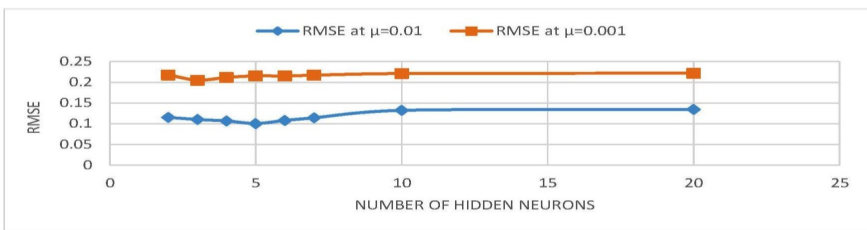


Figure 5. Plot of obtained RMSE values for the different number of hidden neurons.

5. Conclusion

From the result, this research work does not only present the possibility of earthquake prediction using the primary and secondary earthquake waves (P & S-waves), its performance evaluation also shows a high confidence for its adoption as a prediction model. The neural network was trained to a level that it was able to achieve a very good generalization from the input parameters (distance, azimuth, depth of the source and the magnitude of the received primary wave) & the output (time-lag between the P & S-waves). A limitation to this proposed model is the high amount of data sampling required for effective training. This will require a lot of computational resources. As such, for an effective and efficient implementation of this model on an industrial scale, it is recommended that large computing and processing machines be used.

Correspondence
Ogbole C. I.
Department of Computer Engineering
Federal University of Technology Minna, Nigeria
Email: ogbole.inalegwu@futminna.edu.ng

References

Adi, S. & Schnytzer, Y., 2011. *Animal modelling of earthquake and prediction markets*, Israel.

Azzam, R. et al., 2011. Monitoring of landslides and infrastructures with wireless sensor networks in an earthquake environment. *Memorias de la conferencia 5th International Conference on Earthquake Geotechnical Engineering*, pp.10–13.

Bhargava, N. et al., 2009. Earthquake prediction through animal behavior: A review. *Indian J. Biomech*, 78.

Freund, F.T. et al., 2009. Air ionization at rock surfaces and pre-earthquake signals. *Journal of Atmospheric and Solar-Terrestrial Physics*, 71(17–18), pp.1824–1834.

Furen, X., 2010. *Rock stress and earthquakes*, London.

Grant, R.A. & Halliday, T., 2010. Predicting the unpredictable; evidence of pre-seismic anticipatory behaviour in the common toad. *Journal of Zoology*, 281(4), pp.263–271.

Kai Tan & Xiushan Cai, 2010. Prediction of earthquake in Yunnan region based on the AHC over sampling. In *2010 Chinese Con-*

trol and Decision Conference. IEEE, pp. 2449–2452.

Kirschvink, J.L., 2000. Earthquake prediction by animals: evolution and sensory perception. *Bulletin of seismological society of America*, 90, pp.312–323.

Liu, G. et al., 2013. Volcanic earthquake timing using wireless sensor networks. In *Proceedings of the 12th international conference on Information processing in sensor networks - IPSN '13*. New York, New York, USA: ACM Press, p. 91.

Maitha, H., Ali, H. & Hassan, A., 2011. *Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in Al Ain city-UAE*,

Moustra, M., Avraamides, M. & Christodoulou, C., 2011. Artificial neural networks for earthquake prediction using time series magnitude data or Seismic Electric Signals. *Expert Systems with Applications*, 38(12), pp.15032–15039.

Uyeda, S., Nagao, T. & Kamogawa, M., 2009. Short-term earthquake prediction: Current status of seismo-electromagnetics. *Tectonophysics*, 470(3–4), pp.205–213.

Varotos, P. & Alexopoulos, K., 1982. Physical properties of the variation of electric field of the earth preceeding earthquakes. *Technophyics*, 110(1), pp.73–98.

Varotsos, P. et al., 2006. Additional evidence on some relationship between Seismic Electric Signals (SES) and earthquake focal mechanism. *Tectonophysics*, 412(3–4), pp.279–288.

Wiemer, S., 2000. *Earthquake statistics and earthquake prediction research*,

Zuniga, F.R. & Wyss, M., 1995. Inadvertent changes in magnitude reported in earthquake catalogues: Influence on b-value estimate. *Bulletin of seismological society of America*, 85(6), pp.1858–1866.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.177-190

Improved Influence Factor Scheme for Detecting Influential Nodes in Mobile Phone Network

ELIZABETH N. ONWUKA, BALA A. SALIHU & SHERIFF MURTALA
Federal University of Technology, Minna, Nigeria

ABSTRACT The number of mobile phone users is increasing tremendously. Network of users are formed using the call (or social) interactions between these mobile phone users. Such networks could be represented using social network graphs where the nodes represent persons and the edges are the communications between them. In such networks, communities of nodes with certain commonalities could be identified using community detection techniques. It should be noted that in every community there are usually nodes that have high influence, referred to as influential nodes. Knowledge of such nodes helps to understand the communities better and to relate with the community members. For example, removal of influential nodes from a criminal community will collapse the community and probably also the network they belong to. Also, influential nodes could be used to feed information to an entire network. Therefore, it is important to accurately identify nodes that are prominent in a network. For these reasons, work on techniques for identifying influential nodes in communities is currently receiving attention in the research arena. One such technique in literature is the influence factor scheme, which indicates how important an individual node can be in a network. The scheme integrates betweenness centrality, closeness centrality and eigenvector centrality. However, the use of eigenvector centrality in the scheme strongly affects the measure of influence across the network by limiting the detection of influential nodes to the neighbouring nodes around the most influential nodes within the largest component (community) of the network. It neglects the fact that there could be an influential node in other smaller components in the network. This can be misleading, especially in a massive social network like the mobile phone network that contains several neighbourhoods with hundreds or thousands of nodes and edges. This is because it is not necessarily true that every node that is connected to the most important nodes is truly important. This limitation makes it difficult to detect the real influential nodes in large social networks. Principal component centrality is a variant of eigenvector centrality that considers every component (community) in a network when searching for influential nodes across a network graph. In this research, we present an improved influence factor scheme that incorporates closeness centrality, betweenness centrality and principal component centrality to identify nodes that are truly influential in a mobile phone network. The improved scheme has better accuracy, precision and specificity. Furthermore, in terms of accessibility, the improved scheme outperforms the existing scheme because information through the detected influential nodes reached all members of the communities in the network.

Keywords: social network, mobile phone network, influential nodes detection, centrality measures

1. Introduction

Since the invention of mobile communication and other services attached to it, many people find it better and cheaper to communicate using the medium than wired communication thereby attracting more subscribers to use mobile communication network. A survey carried out by international telecommunication union (ITU) shows that the population of mobile phone subscribers increased from 738 million in the year 2000 to 7 billion in 2015 and within this same time the proportion of population covered by a 2G mobile cellular network rose from 58% to 95% with more remote areas captured (Reserved, 2016). In developing countries, at least one member of every household communicates using a mobile phone. Each subscriber enjoys making calls and receiving calls from other users, and also enjoys the same for short message and Internet services. Telecommunication networks have really made the world a global village in the sense that peoples' social reach has expanded even across borders. The log of activities of each user is stored on the user's phone and also recorded with the Mobile Network Operators (MNOs). The information collected by the MNOs is referred to as Call Detail Record (CDR).

CDR contains metadata that describe a specific instance of a telecommunication transaction (calls, messages and Internet services) but does not include the content of that transaction, for example, CDR for a particular call contains both the caller and receiver's number, the time stamp (date and time), the duration of that call and other relevant information. CDR may capture thousands or millions of users within a specific time and place and it can be used to create a network of mobile phone subscribers. CDR is a huge repository of human behavioral data and it belongs to the group of data being currently described as Big Data. Inter-relationship network between humans at various spheres, generally called social networks, can be reconstructed with CDR. A mobile phone network is a social structure that represents the interconnection of mobile phone subscribers based on call detail record (CDR). The idea of forming a social interaction between mobile phone users support researchers in different area of studies like personal mobility prediction, fraud detection in telecommunication (Pinheiro, 2012), urban planning and development, geographical partitioning (Blondel *et al*., 2015) and intelligence gathering for national security (Farley, 2003). Considering the benefits of knowing the most influential nodes in a group or clusters, it is important to develop a technique for identifying the most influential nodes in any given group. This is because identification of such nodes gives a good insight into that group. The major problem in this area is how to accurately determine the genuine influential nodes (individuals) in a social network.

Related work

2.1 Background

There is a rapidly growing literature on influential nodes discovery in social networks, which indicates that a lot of study had been carried out in this field (Borgatti,

2006) (Probst, 2013) (Zhang *et al*., 2013) (Ilyas and Radha, 2010) (Ilyas and Radha, 2011) (Sathik and Rasheed, 2009) (Ahsan, *et al*., 2015) (Singh *et al*., 2013). However, due to the challenges of getting mobile phone data, little studies have been carried out on discovering communities and important mobile subscribers in mobile phone network. A mobile phone network is treated like any other social network that has a tree network structure. Social network is usually modelled as a graph, $G= (V, E)$ where V is a set containing all nodes (actors) in the network and E is also a set containing all edges (links) between two elements (pairs) of set V. If the direction of the edges is considered the graph is said to be directed and undirected otherwise. Also, when the weight of the edges is considered, the graph is said to be weighted and binary (unweighted) otherwise.

Exploring social network data requires basic concepts of graph representation, analysis and visualization (Abraham, 2012). These concepts include centrality measures, shortest path problems, clustering techniques and network density. This is necessary when interpreting result in order to have a good understanding of the social interactions between nodes in a network. Due to the rich resources in social network analysis, it serves as a tool for analyzing and visualizing big data (Lieberman, 2014). Some major areas of study in social network analysis are community structure, detection of cliques and discovery of key nodes and neighbours. Recently, more attention has been given to detection of influential nodes in social network. This is added to the fact that researchers and investigators have taken full advantage of social network analysis to unravel the operation of terrorists and criminals (Farley, 2003). This crime investigation application becomes more necessary now that communication networks has changed the way people live and transact business. It is intuitively believed that criminals rely on this network for planning criminal activities of all sorts. In this study, we focused on identifying important and interesting nodes in a mobile phone network.

2.2 Influential Nodes

Influential nodes are set of nodes whose roles are very important in the spread of influence across the network. These nodes have the tendency to influence other nodes either constructively or destructively. Influential nodes and "key nodes" seem to be the same. Recently, (Probst, 2013) (Singh *et al*., 2013) presented an overview of existing techniques of finding important and influential nodes in social networks. In this subsection, we discuss some of the previous studies that had been done in this area of research. For clarity, we classify the methods of influential nodes detection into two categories: centrality measures and non-centrality approach.

Centrality measures

In graph theory and network analysis, the most important tool is centrality measure. Centrality measures are considered as structural measures of influence that indicate a node's position in a social network. Degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality are the four widely used centrality measures in determining the relative importance of a node within a network. Alt-

hough these measures have limitations, they have been proven to be the basis of other methods for identifying key nodes within a social network (Landherr *et al.*, 2010).

i) Degree Centrality: Degree centrality is defined as the number of edges incident upon a node. In other words, this measure indicates how many nodes can be directly reached by a particular node. The degree centrality of a user, *v* is given by

$$DC(v) = \deg(v) \tag{1}$$

$$\deg(v, G) = |\{u \in V : (u, v) \in E\}| \tag{2}$$

Nodes with high degree centrality scores might be considered important. But one major flaw of this centrality measure is that it relies on direct connections between nodes. Using this individual centrality alone to determine the key nodes will result in the selection of nodes that have high number of direct connections.

ii) Closeness Centrality

Bavelas defined closeness centrality of a user as the reciprocal of the sum of its distances from all other nodes (Bavelas, 1950). This measure is effective in describing the hierarchy among members within a group and can also be used to indicate how fast a node can reach every other node in the network. The weakness of closeness centrality is that it is unsuitable for disconnected graphs.

iii) Betweenness Centrality: This expresses the number of times a user acts as a bridge along the shortest path between two other nodes. Nodes with high betweenness are responsible for controlling the spread of information across the graph. However, they might not be responsible for causing maximum disconnection (fragment) within the network (Borgatti, 2006).

iv) Eigenvector Centrality: Eigenvector centrality (also called eigencentrality) is a measure of how well a particular node is connected to other influential nodes. This is one of the oldest centrality measures developed to assist social analyst to recognize the behavior of people (Seeley, 1949). To determine eigenvector centrality, it is imperative to first find the adjacency matrix, A of the graph, G. with $A = a(v,u)$; and $a(v,u) = 1$ if there exist a link between nodes "v" and "u" and $a(v,u) = 0$ if otherwise for a binary network. The eigenvector centrality of a node, *v* is expressed mathematically as

$$x_v = \frac{1}{\lambda} \sum_{u \in M(v)} x_u \tag{3}$$

where M(v) is the set of neighbours of node, *v*. In matrix representation, eigenvector centrality is given as

$$x_v = \frac{1}{\lambda} A x_u \tag{4}$$

where λ is the eigenvalue (constant) and x_v is the corresponding eigenvector.

v) Other Centrality-Based Approaches: The number of centrality measures extend beyond the four metrics discussed earlier. It is quite interesting that most of the new measures were related in one way or the other to the four most popular centrality measures with a little modification. Ilyas and Radha introduced a new centrality called principal component centrality (PCC), a variant of eigenvector centrality (Ilyas and Radha, 2010). PCC is based on principal component analysis (PCA) and karhunen loeve transform (KLT) which handles graph adjacency matrix as a covariance matrix. Contrary to Eigenvector centrality, PCC provides more features for centrality computation. Moreover, an investigation was carried out to detect influential nodes in two separate datasets using eigenvector centrality and principal component centrality (Ilyas and Radha, 2011). Eigenvector centrality usually considers the most influential user within the largest community in a network and consequently ranks the neighbours of the influential node and ignores other nodes in the remaining small communities that have low eigenvector scores. In the case of PCC, it considers both the nodes in the largest community and other nodes with zero eigenvalues in the remaining small communities.

Despite the introduction of these new centrality measures. The fact still remains that an individual centrality measure might not be the most appropriate for a given network application. A centrality measure is applied depending on specific purpose and the position of a user in a network. For instance, nodes that are most spreaders of virus act as regulators in the network. Another different purpose is identifying nodes that can maximally disrupt the social network. This has opened up more fascinating research fields on group and improved centrality measures that can be universal in identifying the most influential nodes (Everett and Borgatti, 1999). Some studies also considered combining two or more centralities measures in getting a general set of influential nodes. Sathik and Rasheed proposed an algorithm to identify sets of key players based on centrality measures (Sathik and Rasheed, 2009). The authors addressed the key player problems (Borgatti, 2006), using closeness centrality, degree centrality and betweenness centrality.

Lately, in order to adequately discover real influential nodes. Ahsan *et al*. described a scheme that combines closeness centrality, betweenness centrality and eigenvector centrality to determine the influence factor of actors in an online social network obtained from Facebook (Ahsan, *et al*., 2015). The study shows that these three centrality measures are important in measuring the influence of each user and as well as the influence of the entire social network.

Non Centrality Approach

In this subsection, we would be looking at previous studies that employed other techniques different from centrality approach in detecting influential nodes.

$$IF(v) = \frac{2CC_{norm}(v)BC_{norm}(v)PCC_{norm}(v)}{CC_{norm}(v) + BC_{norm}(v)} \quad ; \quad 0 < IF(v) < 1 \tag{9}$$

where $IF(v)$ is the influence factor of node v, $CC_{norm}(v)$ is the normalized closeness centrality of node v, $BC_{norm}(v)$ is the normalized betweenness centrality of node v and $PCC_{norm}(v)$ is the normalized principal component centrality of node v. The normalization for each centrality measure is done using the expression

$$Z_{norm}(v) = \frac{Z(v) - Z_{\min}}{Z_{\max} - Z_{\min}} \tag{10}$$

A node will have an influence factor value between 0 and 1. Where 0 describes node as insignificant and 1 defines node as highly influential.

Results

The data processing was done using the Data laboratory tab in Gephi (Bastian *et al*., 2009), while other analysis was carried out using Python 2.7 on a Dell Latitude computer system with Intel(R) Core(TM) i5 CPU M 540@ 2.53GHz processor and 4GB RAM. NodeXL, a free open-source template for Microsoft® Excel® was used in evaluating the performance of the improved scheme.

4.1 Processed Data

Nodobo dataset contains unsuccessful calls which make up about 30% of the total number of calls. These calls are either calls missed by the call receiver or outgoing calls that failed to connect due to low airtime or weak service signal of network operators. The details are presented in Table 1.

Table 1. Details of the call records

Call Status	Outgoing calls	Incoming calls	Missed calls	Total
Successful calls	5,976	2,998	Nil	8,974
Unsuccessful calls	2,068	169	1,824	4,061

Thus, 691 distinct links with 577 distinct nodes were discovered. As mentioned earlier, privacy of mobile phone users is critical in this analysis. In order to achieve this, each phone number is represented with a new identity number. The new identity number starts with letter "V" and a number ranging from 1 to 577 is attached. To keep track of the 27 seed nodes they are represented using V1, V2, V3……,

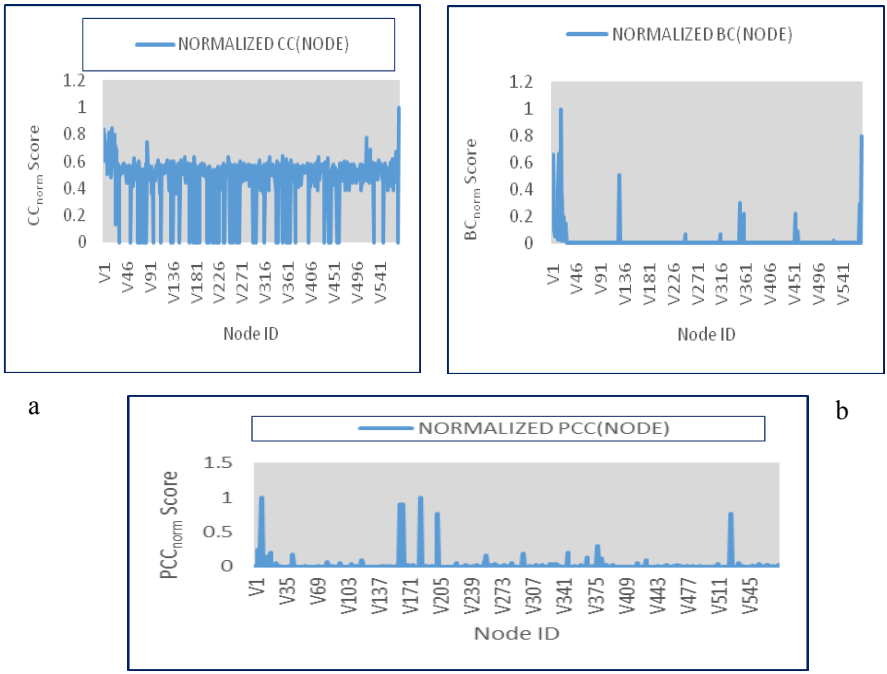
V27. Next, the communication links (edges) and their corresponding duration (weight) were labelled.

4.2 Extracted Features

The centrality measures for each individual node were determined. Their minimum, average and maximum scores are listed in Table 2. The individual score is normalized and the distributions of the normalized scores among the nodes are illustrated in Figure 1. The top ten node with high closeness centrality, betweenness centrality and principal component centrality scores are presented in Table 3.

Table 2. Minimum, average and maximum scores of the centrality measures

CENTRALITY MEASURES	MINIMUM SCORE	AVERAGE SCORE	MAXIMUM SCORE
CC	0.0464450	0.208766	0.3818100
BC	0	0.00740253	0.529191
PCC	0.00000829	0.018398	1.0025730



c Fig. 1: Distribution of normalized centrality scores across the network (a) normalized closeness centrality scores (b) normalized betweenness centrality scores (c) normalized principal component centrality scores

Table 3: Top ten nodes with high closeness centrality scores, high betweenness centrality scores and high principal component centrality scores

Rank	1	2	3	4	5	6	7	8	9	10
CC	V577	V18	V1	V12	V3	V22	V514	V13	V14	V86
BC	V18	V577	V13	V3	V126	V12	V20	V2	V350	V573
PCC	V8	V183	V163	V160	V525	V202	V7	V378	V4	V17

4.3 Evaluation of the Improved Influence Factor Scheme

The improved scheme was evaluated by using it to detect the influential node in the constructed graph. Based on the individual influence factor, thirty-nine nodes were detected while the remaining five hundred and thirty-eight nodes have zero influence. The influential nodes detected are shown in Fig. 2. The minimum and maximum influence factor for the mobile phone network is 0.057797486 and 1.85216E-06 respectively.

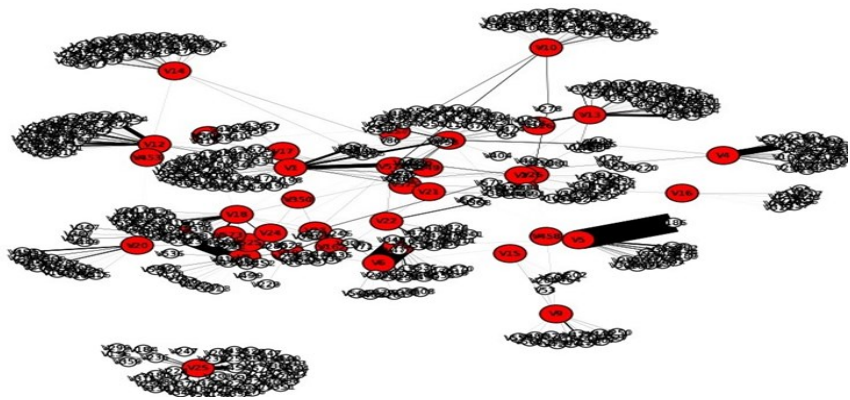


Fig. 2: Location of the identified influential nodes (in red) based on improved Influence factor scheme.

To gain insight into the performance of the improved scheme, a scatter plot was used to show the influence factor of each node for both schemes. This is depicted in Figure 3. According to the existing scheme, sixty nodes were detected as being influential nodes. The two set of influential nodes identified by both scheme is shown in Figure 4. Table 4 summarises the observation from the investigation of the two schemes and Table 5 presents the statistical measures used in comparing the two schemes.

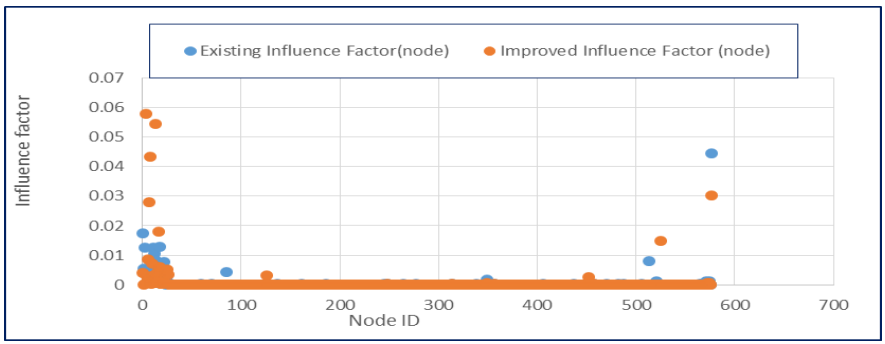


Fig. 3: Comparison between the existing and improved scheme.

Fig 4. Location of the two set of influential nodes detected using both scheme.

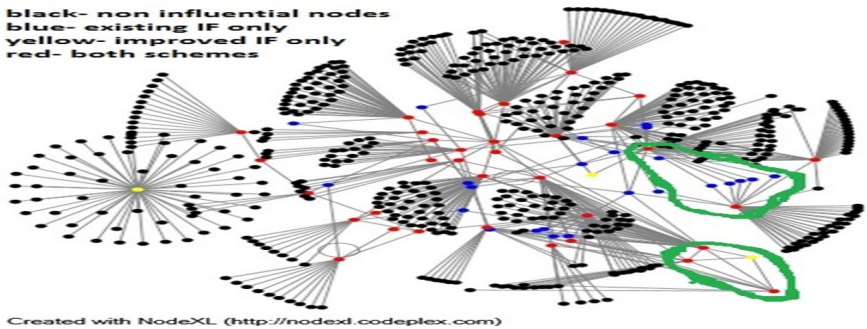
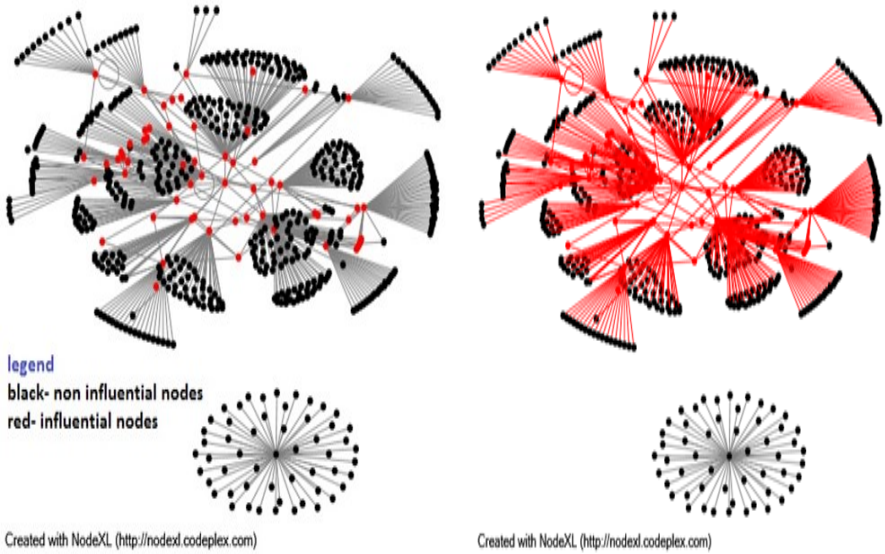


Table 4. Investigation of both schemes

Scheme	True Positive	True Negative	False Positive	False Negative
Existing Scheme	44	2	16	1
Improved Scheme	37	16	2	8

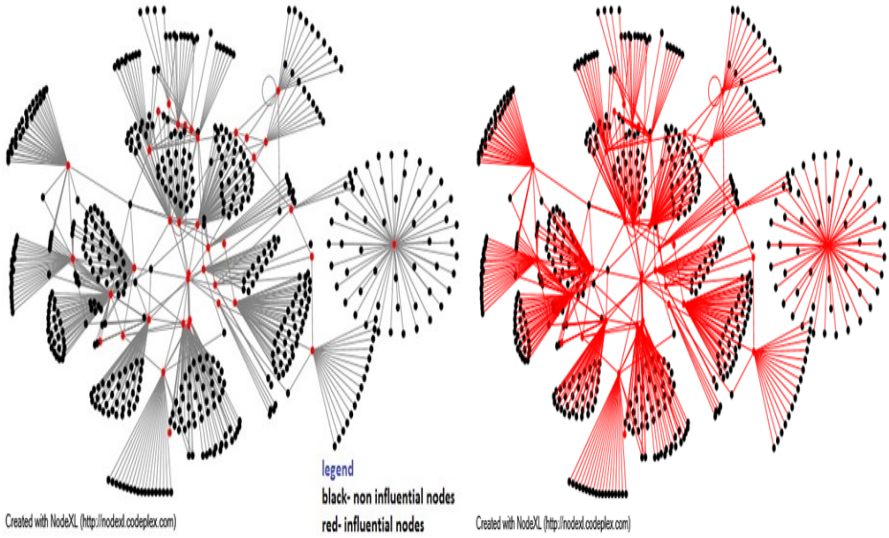
Table 5. Statistical measure of the two schemes

	EXISTING IF SCHEME	**IMPROVED IF SCHEME**
ACCURACY	73.01	84.12
PRECISION	73.33	94.87
SENSITIVITY (Prob. of Detection)	97.78	82.22
F1 MEASURE	83.81	88.10
SPECIFICITY(TNR)	11.11	88.89



a b

Fig. 5 (a): Location of the 60 influential nodes identified by the existing influence factor scheme and (b) nodes accessible by the influential nodes.



a b

Fig.6 (a): Location of the 39 influential nodes detected by the improved influence factor scheme and **(b)** nodes accessible by the influential nodes.

5.0 Discussion

This section discusses the result of the performance of the improved influence factor scheme on a mobile phone network created from the call record dataset. We discovered that identifying high number of influential nodes does not matter especially when these nodes are detected because they are linked to the top influential nodes. The normalized betweenness centrality score distribution discloses that a few selection of nodes are responsible for transferring information from one node to another. This is in contrast with closeness centrality, a node can be surrounded by other nodes but does not necessarily mean that it allows the flow of information. Such node is said to have a betweenness centrality score of 0 and this explains why large number of nodes in the mobile phone networks have 0 betweenness value. The principal component centrality with tuning parameter 'p' equal to 300 (approximately 52% of the 577 largest eigenvalues), nodes with 0 eigenvalues now have a significant principal component centrality value.

The improved scheme concentrated more on nodes that are highly important and reduced the detection of nodes that are not necessary important (with low IF score) by ignoring them. Also, the improved scheme gives a high influence factor than the existing scheme, though it has a less count of detected influential nodes when compared to the existing scheme. It is important to note that most of the nodes that are not discovered by the improved scheme have low IF scores in the existing scheme. These nodes are seen to be influential only because they are connected to top influential nodes and not that they are necessarily important, this can be confirmed from their positions in the network graph. Ignoring these nodes will not affect the network in any way.

Thorough investigation of the two set of influential nodes detected by both schemes reveals that some influential nodes detected only by the existing scheme as shown in Figure 4 (in blue) are connected to the same pair of influential nodes and they are not connected to any other nodes, this is attributable to the fact that the existing scheme incorporated eigenvector which has the weakness of considering only the influential nodes and neighborhoods. The improved scheme outperforms the existing scheme in terms of accuracy, precision, F1 measure and specificity. However, it underperforms in sensitivity (probability of detection).

To further compare the improved scheme, the percentage of nodes that are reachable through the influential nodes identified using each scheme is determined. The nodes detected according to the existing influence factor were able to reach 90.64% of the nodes while the improved scheme was able to reach 100% of the nodes. The existing scheme ignored the smaller component of the graph; therefore no influential node was detected in it which means that there is no way of accessing that component. But the improved scheme detected an influential node in the smaller component through which the nodes in that component can be accessible. Hence in a large graph with many components, the improved scheme will detect all influential nodes in every component.

6.0 Conclusion

The influential nodes detection in a mobile phone network is a difficult job as huge amount of mobiles subscribers (nodes) are connected to the mobile phone network every seconds. In this paper, a method to identify the most influential nodes based on influence factor measure is developed. The basic components of the proposed approach are closeness centrality, betweenness centrality and principal component centrality. The proposed scheme integrates these three centrality measures to improve the detection of influential nodes across the mobile phone network. The results obtained from the experimental analysis and comparison with the existing influence factor scheme showed that the improved scheme is more accurate and precise in identifying influential nodes that can maximally spread influence across the entire mobile phone network, however the probability of detection is slightly lower. The specific recommendations for further studies based on limitations in this research is the collection of dataset which captures more components, more nodes (users) and more connection links (edges). More so, finding what can be done to correct the sensitivity of the improved scheme.

Correspondence
Sheriff Murtala
Department of Telecommunications Engineering
Federal University of Technology, Minna, Nigeria

References

Abraham, A. (ed.) (2012) Computational social networks: Mining and visualization. London: Springer London.

Agarwal, N., Liu, H., Tang, L. and Yu, P.S. (2008) 'Identifying the influential bloggers in a community' In Proceedings of the 2008 international conference on web search and data mining (pp. 207-218). ACM.

Ahsan, M., Singh, T. and Kumari, M. (2015) 'Influential node detection in social network during community detection', In Cognitive Computing and Information Processing (CCIP), 2015 International Conference on (pp. 1-6). IEEE.

Bastian, M., Heymann, S. and Jacomy, M. (2009) Gephi: an open source software for exploring and manipulating networks. ICWSM, 8, pp.361-362.

Bavelas, A. (1950) 'Communication patterns in Task-Oriented groups', The Journal of the Acoustical Society of America, 22(6), pp. 725–730. doi: 10.1121/1.1906679.

Blondel, V.D., Decuyper, A. and Krings, G. (2015) 'A survey of results on mobile phone datasets analysis', EPJ Data Science, 4(1). doi: 10.1140/epjds/s13688-015-0046-0.

Borgatti, S.P. (2006) 'Identifying sets of key players in a social network', Computational and Mathematical Organization Theory, 12(1), pp. 21–34. doi: 10.1007/s10588-006-7084-x.

Brandes, U. (2001) 'A faster algorithm for betweenness centrality*', The Journal of Mathematical Sociology, 25(2), pp. 163–177. doi: 10.1080/0022250x.2001.9990249.

Canali, C. and Lancellotti, R. (2012) 'A quantitative methodology based on component analysis to identify key users in social networks' , International Journal of Social Network Mining, 1(1), pp.27-50.

Catanese, S., Ferrara, E. and Fiumara, G. (2012) 'Forensic analysis of phone call networks', Social Network Analysis and Mining, 3(1), pp. 15–33. doi: 10.1007/s13278-012-0060-1.

Eirinaki, M., Monga, S.P.S. and Sundaram, S. (2012) 'Identification of influential social networkers', International Journal of Web Based Communities, 8(2), pp.136-158.

Erlandsson, F., Bródka, P., Borg, A. and Johnson, H. (2016) 'Finding Influential Users in Social Media Using Association Rule Learning', Entropy, 18(5), p.164.

Everett, M.G. and Borgatti, S.P. (1999) 'The centrality of groups and classes', The Journal of Mathematical Sociology, 23(3), pp. 181–201. doi: 10.1080/0022250x.1999.9990219.

Farley, J.D. (2003) 'Breaking Al Qaeda cells: A mathematical analysis of counterterrorism operations (A guide for risk assessment and decision making)', Studies in Conflict & Terrorism, 26(6), pp. 399–411. doi: 10.1080/10576100390242857.

Ferrara, E., De Meo, P., Catanese, S. and Fiumara, G. (2014) 'Detecting criminal organizations in mobile phone networks', Expert Systems with Applications, 41(13), pp. 5733–5750. doi: 10.1016/j.eswa.2014.03.024.

Freeman, L.C. (1977) A set of measures of centrality based on betweenness. Sociometry, pp.35-41.

Goldenberg, J., Han, S., Lehmann, D.R. and Hong, J.W. (2009) 'The role of hubs in the adoption process'. Journal of Marketing, 73(2), pp.1-13.

Han, B., Li, J. and Srinivasan, A. (2014) 'Your friends have more friends than you do: Identifying influential mobile users through random-walk sampling', IEEE/ACM Transactions on Networking, 22(5), pp.1389-1400.

Heidemann, J., Klier, M. and Probst, F. (2010) 'Identifying key users in online social networks: A pagerank based approach.

Hinz, O., Skiera, B., Barrot, C. and Becker, J.U. (2011) 'Seeding strategies for viral marketing: An empirical comparison', Journal of Marketing, 75(6), pp. 55–71. doi: 10.1509/jm.10.0088.

Ilyas, M.U. and Radha, H. (2010) 'A KLT-inspired node centrality for identifying influential neighborhoods in graphs'. In Information Sciences and Systems (CISS), 2010 44th Annual Conference on (pp. 1-7). IEEE.

Ilyas, M.U. and Radha, H. (2011) 'Identifying influential nodes in online social networks using principal component centrality'. In 2011 IEEE International Conference on Communications (ICC) (pp. 1-5). IEEE.

Landherr, A., Friedl, B. and Heidemann, J. (2010) 'A critical review of centrality measures in social networks', Business & Information Systems Engineering, 2(6), pp. 371–385. doi: 10.1007/s12599-010-0127-3.

Lieberman, M. (2014) Visualizing big data: Social network analysis. In Digital research conference.

McDiarmid, A., Bell, S., Irvine, J. and Banford, J. (2013) 'Nodobo: Detailed mobile phone usage dataset'. Unpublished paper, accessed at http://nodobo. com/papers/iet-el. pdf on, pp.9-21.

Narayanam, R. and Narahari, Y. (2011) 'A shapley value-based approach to discover influential nodes in social networks', IEEE Transactions on Automation Science and Engineering, 8(1), pp.130-147.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Menezes, M.A. de, Kaski, K., Barabási, A.-L. and Kertész, J. (2007) 'Analysis of a large-scale weighted network of one-to-one human communication', New Journal of Physics, 9(6), pp. 179–179. doi: 10.1088/1367-2630/9/6/179.

Ortiz-Arroyo, D. and Hussain, D.A. (2008) 'An information theory approach to identify sets of key players'. In Intelligence and Security Informatics (pp. 15-26). Springer Berlin Heidelberg.

Perkins III, F.C., Convergys CMG Utah Inc., 2002. System and method for processing call detail records. U.S. Patent 6,396,913.

Pinheiro, C.A.R. (2012) Community detection to identify fraud events in telecommunications networks. SAS SUGI Proceedings: Customer Intelligence.

Probst, F. (2013) Customer Relationship Management in a Digitally Connected World (Doctoral dissertation).

Reserved, I.A.R. (2016) ITU: Committed to connecting the world. Available at: http://www.itu.int (Accessed: 16 October 2016).

Sathik, M.M. and Rasheed, A.A. (2009) A centrality approach to identify sets of key players in an online weblog. International Journal of Recent Trends in Engineering, 2.

Schult, D.A. and Swart, P. (2008) Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conferences (SciPy 2008) (Vol. 2008, pp. 11-16).

Seeley, J.R. (1949) 'The net of reciprocal influence; a problem in treating sociometric data', Canadian Journal of Psychology Revue Canadienne de Psychologie, 3(4), pp. 234–240. doi: 10.1037/h0084096.

Shetty, J. and Adibi, J. (2005). 'Discovering important nodes through graph entropy the case of enron email database' ,In Proceedings of the 3rd international workshop on Link discovery (pp. 74-81). ACM.

Proceedings on Big Data Analytics
& Innovation (Peer-Reviewed),
Volume 1, 2016, pp.192-198

A Survey of Range-Based Techniques for Localizing Primary User Emulators in Cognitive Radio Network

S.A. ADEBO, E.N. ONWUKA, A.J. ONUMANYI & A.S. USMAN
Federal University of Technology, Minna, Nigeria

ABSTRACT A problem currently facing the wireless communication industry is spectrum underutilization, wherein licensed Primary Users (PUs) only marginally use their bands in time and space. Consequently, Opportunistic Spectrum Sharing (OSS) has been proposed as a solution to this problem with Cognitive Radio (CR) being the enabling technology. CR describes the capability of a radio to dynamically and opportunistically access licensed spectrum without causing interference to the PU. While CR is being developed, one main security problem it faces is Primary User Emulation (PUE) attack. A PUE attack occurs when mischievous Secondary Users (SU), that is CRs, begin to emulate PU signals either to deny other SUs from gaining access to free bands or to simply acquire bandwidth for selfish purposes. In this regard, localization is considered as a solution, wherein the location of the PUE attacker is determined and compared to the location of the true PU transmitter and any significant deviation in their positions is indicative of a PUE attacker. Thus, this paper is a short note describing the different state-of-the-art localization techniques, elucidating their advantages and disadvantages. The approach adopted included identification of the various localization techniques for PUE identification, with subsequent analysis under the following metrics: accuracy, precision, computational complexity, cost, energy efficiency, and size of hardware. It was concluded that the Angle of Arrival and Received Signal Strength Indicator techniques do not require the cooperation of PUE for its detection, hence making these techniques most suitable for use. It is believed that this mini-review will be beneficial particularly for budding researchers seeking enlightenment in this research area.

Keywords: Cognitive Radio, Primary User, Localization, Angle of Arrival, Range

1. Introduction

Localization refers to the accurate determination of the position of a node in physical space. It is typically classified into two broad categories namely: Range-free and Range- Based Localization techniques. Range-Free localization techniques do not depend on the use of distance and angle for localization, while Range-based techniques use the distance and angle for localization (Alrajeh et al., 2013). In essence, the range-based algorithms compute the distance between nodes and use the principle of geometry to calculate the location of a focal node. These algorithms are used to compute range metrics such as Angle of Arrival (AOA), Time of Arrival (TOA), Time Difference of Arrival (TDOA), and Received Signal Strength

Indicator (RSSI) (Chen and Park, 2006). When implementing any localization technique, certain metrics such as accuracy, precision, computational complexity, cost, energy efficiency, and size of the hardware is considered because they easily elucidate the similarities and differences between these approaches, (Amundson and Koutsoukos, 2009). The different types of Range-Based techniques available in literature are the Time of Arrival (TOA), Time Difference of Arrival (TDOA), Angle of Arrival (AoA), and Received Signal Strength Indicator (RSSI). These are discussed in the following sections.

2. Time of Arrival

The velocity and time of arrival of the radio signals are used to calculate the location of the unlocalized transmitter in this localization method (Singh and Sharma, 2014). When a signal is transmitted from all nodes to their neighbours with the same predefined velocity, w, each node sends the received signal back to the transmitter. A node 'a' estimates its distance from its neighbour 'b' by using this formula:

$$Dis_{ab} = 2^{-1}\left(t^{a}_{rec} - t^{a}_{tra}\right) - \left(t^{b}_{rec} - t^{b}_{tra}\right) \quad (1)$$

where,

Dis_{ab} = distance between a and b t^{a}_{rec} = time of receiving signal for node 'a', t^{a}_{tra} = time of transmission of signal for node 'a', t^{b}_{rec} = time of receiving signal at node 'b', t^{b}_{tra} = time of transmission of signal at node 'b', t^{b}_{rec} = time of receiving of signal for node 'b', t^{b}_{tra} = time of transmission of signal for node 'b'. Time of Arrival (TOA) has two main limitations to its effectiveness (Srbinovska et al., 2011). These limitations are: a.) Synchronization at microsecond level between all sensor nodes: Roundtrip time was applied to surmount the problem of synchronization. Roundtrip propagation time measurements compute the difference between the time when a signal is sent by a sensor and the time when a signal returned by a second sensor is received at the original sensor. Even with the application of roundtrip propagation time, the problem of synchronization was not solved (Guvenc and Chong, 2009). b.) Internal delay required for handling the signal in the receiver sensor affects this method, which leads to inaccurate determination of the distance travelled by the signal (León et al., 2012). In a mobile-based approach, it is required that the transmitting node to be located collaborates with other nodes. This technique is suitable in wireless sensor networks where all nodes are friendly nodes. In a cognitive radio network (CRN) scenario where the objective is to locate attacker, this will not work because the attacker will not cooperate with other nodes to reveal its position (León et al., 2012).

3.Time difference of Arrival

Time Difference of arrival (TDOA) which is an upgrade of TOA estimates the location of an unlocalized transmitting node. In this method, the position of the sensor node is estimated by using the time difference of arrival of the radio and ultrasound signals at different sensor nodes. In this scenario, each node will have a microphone and a speaker, an anchor node is required to coordinate the measurements. When the anchor node transmits signal to other nodes in the network, it waits for some time lapse, after which the receiving node generates short high sound with the help of speaker (this marks the time the signal was received by it). The microphone saves the time it identifies the chirps. The unlocalized node uses this time information to determine the distance between the anchor node and itself. If there are two references, 'a' and 'b' the TDOA measurement can be transformed into a distance by the following formula:

$$Dis_{ab} = dis_a - dis_b = c(t_a - t_b) = c.t_{ab} \quad (2)$$

This technique requires two types of senders and receivers on each node and its location estimation is computationally intensive (Singh et al., 2012, Leelavathy and Sophia, 2014, Jin et al., 2009). Both TOA and TDOA have difficulty with measuring time resulting from synchronization of the devices involved. Hence the problem of synchronization by TOA was not also solved by TDOA.

4. Angle of arrival

Angle of Arrival (AOA) is the angle between some reference direction (orientation) and propagation direction of an incident wave. As a localization scheme, AOA is used to know the location of unlocalised sensor node by computing absolute or relative angles between neighbours. The measurement is in degree taken in clockwise direction from the north. For absolute AOA, the orientation should be 0^0 and it should point to the north; otherwise, it is a relative AOA (Kułakowski et al., 2010). The direction of neighbours is found by use antenna array and some anchor nodes equipped with compass and Global Positioning System (GPS). Unlocalised nodes update the location information broadcast by the anchor nodes along the way. Location is computed after getting the information from at least three anchor nodes. This approach is expensive to use in wide CRN since it requires additional hardware for transmitting and receiving location information (Heng and Gao, 2013).

5.received signal strength indicator

RSS-based localization technique arises from the fact that there exists a strong connection between the distance of a wireless link and the RSSI. If the signal strength travels a distance, d , its strength is inversely proportional to the distance travelled as given in equation (3) (Leelavathy and Sophia, 2014).

$$RSSI\, a\, \frac{1}{d^2} \qquad (3)$$

Table 2: Suitability of Range-Based Methods for Detecting PUE

S/N	Technique	Suitability	Reason
1.	GPS	Unsuitable	Attacker has to reveal itself
2.	TOA	Unsuitable	It requires cooperation of the attacker and suffers synchronization
3.	TDOA	Unsuitable	It cannot handle tight synchronization among the participating nodes
4.	RSSI	Suitable	Accurate with or without cooperation
5.	AOA	Suitable	Does not need the cooperation of other nodes

It assumes those transmission power and path loss models are known and uses them to calculate the distance between the PU and the reference node (Chen and Park, 2006). Because of the dynamics of the indoor/outdoor environments, it is expected to have high error rate. This method is unsuitable for long distance network links. Specifically, given a transmitter-receiver pair, Received Signal Strength (RSS) can be modeled as a function of transmitted power and transmitted-receiver distance. Therefore, if a correct model is used and there are multiple observers taking RSS measurements, a transmitter location can be estimated using the model.

RSS-based technique is very effective in estimating the distance of the PUE from a reference node, but, there are some two issues that may affect its operation: Possible manipulation by malicious or multiple transmitters, and inaccuracy of the RSS measurement. These issues can be addressed if many RSS measurements are taken properly and the measured data are properly processed. Typically, as the distance between transmitter and receiver increases, RSS value decreases.(Pu et al., 2011).

To be able to use the approach above, two problems need to be addressed. a.) Path fade change over time and a PUE attacker may change its location or vary its transmission power frequently to stall localization, thereby causing drastic fluctuation in the RSS measurements within a short period of time. Averaging the measurement taken at different times cannot mitigate this problem since different RSS values abound for different measurements taken at different times. The panacea to this challenge is to take a "snapshot" of the RSS distribution in a given network. That requires the receiving nodes of a cognitive radio network (CRN) to take synchronized RSS measurements in a given band (Boukerche et al., 2007).

Short variation of distance causes high variation in the magnitude of the RSS (30dB to 40dB) (Liu et al., 2010). This poses challenge to deciding the location of PUs by just reading the raw data in a snapshot of RSS distribution. Since snapshot data is not free from uncertainty, the required data integrity is attainable using appropriate data smoothing technique so that the snapshot data can be used to solve localization problem. Data capturing techniques aims at eliminating noise from the snapshot data while capturing important patterns in raw data. Identifying the RSS measurements is possible once the variance in the raw RSS measurements is decreased through data smoothing methods such as Fourier filters, Loess fitting, Local averaging (Chen et al., 2008). A Summary of range-Based localization techniques is presented in Table 1while that of suitability of each range method for localizing the PUE is presented in Table 2.

Table 1: Comparison of Range-Based Localization Techniques

Technique	Accuracy	Precision	Computational Complexity	Cost	Energy Efficiency	Size of Hardware
GPS	High	High	Low	High	Low	Small
TOA	Medium	High	High	High	Low	Large
TDOA	High	Medium	High	High	High	Large
AOA	Low	Medium	High	High	Medium	Large
RSSI	High	Low	Low	Low	High	small

6. Discussion

Since cooperation of the participating nodes is needed for TOA localization technique, and an attacker will not cooperate with other nodes in detecting it, this localization technique is not suitable for detection of PUE. Although TDOA technique bypasses cooperation, it requires tight synchronization among the cooperating nodes, but it cannot handle such tight synchronization. GPS requires the attacker to reveal itself, thus, it is unsuitable for detecting PUE (D. Gumey, et al., 2008). Whereas AOA does not require the cooperation of other nodes as it measures the direction of signal at different nodes and computes its location by using simple triangulation method, RSSI is accurate within a short range (Guvenc et al., 2009).

7. Conclusion

Radio spectrum is a very valuable resource for wireless communication. CR system leads to spectrum usage efficiency and alleviation of spectrum scarcity. For CR to perform optimally, PUE's activities must be prevented. In this paper, we have discussed the four basic Range-Based PUE localization methods for detecting PUE. Based on the comparison of the four Range-Based localization techniques, it is evident that AOA and RSSI techniques do not require the cooperation of PUE to detect it although each of them has some merits over the other.

Correspondence

S.A. Adebo
Department of Telecommunications Engineering
Federal University of Technology
Minna, Nigeria
Email: adebosamuel@yahoo.com

References

ALRAJEH, N. A., BASHIR, M. & SHAMS, B. 2013. Localization techniques in wireless sensor networks. *International Journal of Distributed Sensor Networks,* 2013.

AMUNDSON, I. & KOUTSOUKOS, X. D. 2009. A survey on localization for mobile wireless sensor networks. *Mobile Entity Localization and Tracking in GPS-less Environments.* Springer.

BOUKERCHE, A., OLIVEIRA, H. A., NAKAMURA, E. F. & LOUREIRO, A. A. 2007. Localization systems for wireless sensor networks. *IEEE wireless Communications,* 14**,** 6-12.

CHEN, R. & PARK, J.-M. Year. Ensuring trustworthy spectrum sensing in cognitive radio networks. *In:* Networking Technologies for Software Defined Radio Networks, 2006. SDR'06.1 st IEEE Workshop on, 2006. IEEE, 110-119.

CHEN, R., PARK, J.-M. & REED, J. H. 2008. Defense against primary user emulation attacks in cognitive radio networks. *IEEE Journal on selected areas in communications,* 26**,** 25-37.

GUVENC, I. & CHONG, C.-C. 2009. A survey on TOA based wireless localization and NLOS mitigation techniques. *IEEE Communications Surveys & Tutorials,* 11**,** 107-124.

HENG, L. & GAO, G. X. 2013. Accuracy of Range-Based Cooperative Localization in Wireless Sensor Networks: A Lower Bound Analysis. *arXiv preprint arXiv:1305.7272*.

JIN, Z., ANAND, S. & SUBBALAKSHMI, K. 2009. Mitigating primary user emulation attacks in dynamic spectrum access networks using hypothesis testing. *ACM SIGMOBILE Mobile Computing and Communications Review,* 13**,** 74-85.

KUŁAKOWSKI, P., VALES-ALONSO, J., EGEA-LÓPEZ, E., LUDWIN, W. & GARCÍA-HARO, J. 2010. Angle-of-arrival localization based on antenna arrays for wireless sensor networks. *Computers & Electrical Engineering,* 36**,** 1181-1186.

LEELAVATHY, S. & SOPHIA, S. 2014. Providing Localization using Triangulation Method in Wireless Sensor Networks. *International journal of innovative technology and exploring engineering,* 4**,** 47-49.

LEÓN, O., HERNÁNDEZ-SERRANO, J. & SORIANO, M. 2012. Cooperative detection of primary user emulation attacks in CRNs. *Computer Networks,* 56**,** 3374 -3384.

LIU, Y., NING, P. & DAI, H. Year. Authenticating primary users' signals in cognitive radio networks via integrated cryptographic and wireless link signatures. *In:* 2010 IEEE Symposium on Security and Privacy, 2010. IEEE, 286-301.

PU, C.-C., PU, C.-H. & LEE, H.-J. 2011. *Indoor location tracking using received signal strength indicator*, INTECH Open Access Publisher.

SINGH, A. & SHARMA, A. 2014. A Survey of Various Defense Techniques to Detect Primary User Emulation Attacks. *International Journal of Current Engineering and Technology,* 4.

SINGH, B., SAHOO, S. K. & PRADHAN, S. R. 2012. Performance evaluation of anchor-based range-based localization systems in wireless sensor networks. *International Journal of Computer Applications,* 52.

SRBINOVSKA, M., DIMCEV, V., GAVROVSKI, C. & KOKOLANSKI, Z. 2011. Localization Techniques in Wireless Sensor Networks using Measurement of Received Signal Strength Indicator. *FACULTY OF ELECTRICAL ENGINEERING UNIVERSITY OF BANJA LUKA***,** 67.

SRBINOVSKA, M, DIMCEV, V, GAVROVSKI, C, & KOKOLANSKI, Z. (2011). Localization Techniques in Wireless Sensor Networks using Measurement of Received Signal Strength Indicator. *FACULTY OF ELECTRICAL ENGINEERING UNIVERSITY OF BANJA LUKA*, 67.

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.200-211

Performance Evaluation of the Interconnect Clearing Houses in the Nigerian Telecommunications Industry

ELIZABETH N. ONWUKA, CAROLINE ALENOGHENA & E.S. DIKKO
Federal University of Technology, Minna, Nigeria

ABSTRACT A telecommunications Clearinghouse (CH) refers to a central exchange where calls from different mobile network operators (MNOs) are interconnected for the purpose of independently measuring how much traffic from one MNO was carried by another MNO. This enables interconnection traffic and billing to be correctly determined without dispute between concerned MNOs. There are ongoing contentions between the MNOs and the Clearing Houses in Nigeria, as to whether the Clearing houses have adequate infrastructural capacity to carry the off-net traffic routed through them by the MNOs. This paper presents the performance analysis of two out of the five licensed interconnect clearing houses in Nigeria. This was done by analyzing data obtained from the facilities. Basic parameters from this data, which was collected at Time Consistent Busy Hours (TCBH) are the carried calls (i.e. total incoming traffic), the Circuit Seizures or the CCS figures, and the congestion figures. These parameters are then used to calculate some critical exchange performance parameters such as: exchange Grade of Service (GOS), route utilization, route congestion percentage, answer-to-seizure ratio (ASR), traffic intensity (in Erlangs), as well as the number of channels the infrastructure requires to conveniently carry the measured Busy Hour traffic without terminating or delaying other lines. The data is analyzed using traffic engineering schemes. Results Obtained show that in compliance with the regulator on the 10% minimum threshold of interconnect traffic that must be routed through the interconnect exchange operators, the two CHs studied were found to have adequate capacity to conveniently carry the amount of traffic being routed through them, with a GOS of 1.38% and 0.2% respectively. Hence, the authors conclude that, at the moment, the mobile operators can conveniently route 10% of their interconnect traffic through the Clearinghouses.

Keywords: Clearinghouse, Grade of Service, Traffic, Erlang-B, Off-net

Introduction

The Interconnect Clearing House model was established and licensed by the Nigerian Communications Commissions (NCC) in 2004 to serve as a third party to Mobile Network Operators. They were also licensed to work as transit operators for receiving and distributing calls from different operators as well as to ensure appropriate revenue sharing and to ensure central monitoring of nationwide off-net calls of different operators. Part of their function is to provide systems for policing fraud and ensuring revenue/tax collection. Currently, active clearing houses in the

country include: (i) Interconnect Clearinghouse Limited, (ii) Medallion Communications Limited, (iii) NicoonX Communications Limited, (iv) Solid interconnectivity Services Limited, and (v) Breeze Interconnect Limited. The figure 1 shows a clearinghouse interconnecting MNOs.

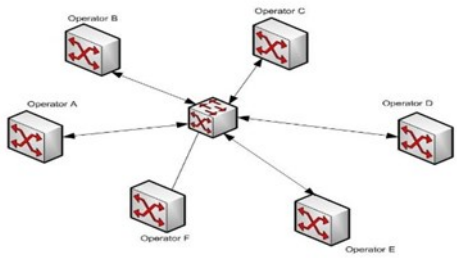


Fig.1: Network Architecture of an interconnected clearing house model (Medallioncom.com)

2. Background of Study

Since 2001, the era of the telecom industry deregulation in Nigeria, the country, inter alia, has been plagued with the issue of interconnection indebtedness. Operators are indebted to one another due to non-payment of the tariff for their traffic that passed through fellow operators' network. This is threatening the growth of the industry as the debts have been on a steady rise till date. Our research has unravelled one of the major causes of the indebtedness, inter alia, as the differentials in billing records amongst MNOs that are connected on the peer-to-peer basis. When the amount of interconnection fees payable is in dispute amongst these MNOs, the financial obligation of remittance of such fees are neglected, thus the debt piles (communications week, 2012). However, in the event that a clearinghouse (which is a third party that routes traffic that originated from one MNO and terminates on another) is involved, there is, therefore, an independent record stating "how much" traffic was exchanged between both concerned operators. Hence, the resulting interconnection fee can unanimously be agreed upon (The communicator, 2016). But in the wake of time, mobile operators have preferred to exchange traffic directly amongst themselves than through the CHs, on the grounds that the CHs, inter alia, does not have enough capacity to route 10% of the combined capacities of the MNOs as mandated by the regulator. Moreover, as at first quota of 2016, an MNO with 23% market share was the only mobile operator that did not route any percentage of its traffic through any CH (NCC, 2016). This has therefore created a miasma in the telecommunications sector in Nigeria.

Thus, there is a need to carry out a performance evaluation on the licensed clearing houses, by carrying out tele-traffic engineering analysis on them based on the current traffic they carry, in order to ascertain their current infrastructural

capacity. This will enable proper determination of the capacity gap they need to fill to be able to route the requisite amount of traffic.

The rationale for the establishment of ICHs is unique and excellent (The communicator, 2016). This indicates that if properly implemented, the clearing houses in the telecommunications' scene, has the potency of propelling the sector to the next phase. However, if improperly, partially or not implemented at all, it could seriously impair the Nigerian telecommunications sector. The clearing houses are licensed to perform the following functions:

i. Accurate billing, reconciliation, and better service provision
ii. Zero debt tolerance
iii. Remove Hazards posed by peer-peer system
iv. Faster off-net calls connection
v. Eliminate congestion posed by peer-to-peer connection
vi. Provision of mobile number portability services
vii. Co-location services and single point of connectivity services to VAS providers
viii. Central monitoring of traffic for tax and security purposes
ix. To reduce money and time spent in resolving interconnection disputes amongst operators (The communicator, 2016).

Research Methodology

The method used in this work is tele-traffic engineering analysis. This is usually employed in telecommunication network dimensioning. Network dimensioning refers to processes involved in determining the minimum capacity requirements that will still allow the Grade of Service (GoS) requirements to be met. This is often achieved with traffic measured during the busy hour. To effectively dimension a telecommunications network, traffic engineering techniques must be employed (Chettinad College, n.d.). Traffic is defined as the occupancy of the server. The basic purpose of traffic engineering is to ensure that while resources are economically utilized, conditions for adequate service provision to subscribers are determined. That is, without compromising Quality of Service (QOS). Traffic engineering offers the foundation for scrutinizing the design of telecommunications networks or model, as well as the dimensioning capacity for the design of telecommunication networks (Bharat, 2007). It provides means to determine the quantity of the basic equipment needed to provide a particular level of service with respect to the given traffic pattern and volume.

For a loss system, the GOS is the probability of loss. With Traffic Engineering, it is possible to determine the ability of a network to carry a given traffic at a particular loss probability(GOS). This is achieved by balancing the following factors based on a given amount of traffic: (i) Grade of Service (GOS) and (ii) The requisite resources or Infrastructure. The following parameters are used in network dimensioning:

(i) Busy Hour: This is the 60-minute interval that has the highest call traffic in a day. Busy Hour may depend on several factors such as stock market, weather, international events and holidays. In addition to these variations, there are unpredictable peaks which could be caused by money activity, weather, sporting events, etc. Hence, it is expedient that such fluctuations be taken into account when designing switching networks. There are, therefore, three kinds of busy hours (Iversen V.B., 2015):

(ii) Average Daily Peak Hour (ADPH): In this kind, the busiest hour is determined separately for each day, with each day having different timing and then averages over a number of days.

(iii) Time Consistent Busy Hour (TCBH): This is a one-hour period, same for each day, which produces the highest traffic over specified number of days.

(Vi) Fixed Daily Measurement Hour (FDMH): This is a predetermined, fixed measurement hour, say between 19:00 - 20:00 hours; with the measured traffic is averaged over the days under considerations.

Erlang(E): It is an international dimensionless unit of telephone traffic as a measure of offered load or carried load. For calls carried, the value followed by "erlangs" is a representation of how many concurrent calls were carried by the circuits in a specified period (often one hour). One erlang traffic refers to a single channel that is in continuous use. While 1.5E, refers to two channels in which one is used to 100 percent capacity and the other, 50 percent capacity (Introduction to switching systems).

Traffic ***Intensity(TI):*** This is the measure of the traffic in terms of occupancy of the servers on the network within a specified period of time, usually a busy hour. It is calculated as:

$A = CTh \quad Erlangs \qquad (1)$

From the above, it is obvious that the Traffic Intensity is a Call-Time product, therefore, two parameters are involved: *Th* is the average holding time of a call while C is the average calls arrival rate. Traffic Intensity is sometimes also referred to as "Subscriber traffic" or "Trunk traffic". These, therefore, introduces to traffic viewed in two dimensions. While one is based on traffic generated by subscribers, the other focuses on the observation of busy servers in the network. There are instances when the traffic generated by subscribers overgrows the capacity of the network. When this situation occurs, either of the following happens:

i. The excess traffic is either gets discarded without being serviced or
ii. The excess traffic gets queued and wait till the busy channels are free.

These two kinds of systems listed above are the Loss Systems and Delay Systems respectively. Clearing houses, which have conventional automatic switches behave like loss systems.

A loss system is made up of two basic performance parameters, which are the Grade of Service and blocking probability.

Grade of Service (GOS) is a measure of the probability that a percentage of the offered traffic will be blocked or delayed. It is often expressed as a fraction of calls that fail to receive immediate service and ends up eventually blocked or the fraction that is forced to wait longer than a given time.

$$GOS = \frac{LostTraffic}{OfferedTraffic} = \frac{BlockedBusyHourCalls}{OfferedBusyHourCalls}$$

$$GOS = \frac{A - Ao}{A} \quad (2)$$

where A_0 = traffic carried, A = Traffic and $(A - A_0)$ = Lost Traffic

A general standard for GOS telecom networks is 0.02 (2%). That is, 2 out of every 100 offered calls may be blocked. The smaller the GOS value, the better the QOS. The blocking probability B is defined as the probability that all the servers in a system are busy. It is a very important parameter to be considered in the dimensioning of any telephone network. However, there is a difference between GOS and blocking probability. While the GOS is from the subscribers' point of view, the blocking probability is from the network/switching point of view (Okechukwu U., 2013). Blocking probability can be calculated using techniques such as Lee graphs and Jacobaeus methods. The probability that traffic exceeds some set threshold, constitutes the basic idea of congestion theory. During which no new calls are granted access into the system. Congestion could be grouped into either Time or Call congestion. While time proportion is the percentage of the TIME that all channels on the network are busy, Call congestion is the ratio of the CALLS that arrive when all servers are busy. We can therefore also call GOS, Call Congestion and the Blocking probability, Time Congestion.

An ideal switching system is one which has the number of subscribers equal to the number of channels. If such a system exists, traffic analysis would be useless. But this is practically impossible and uneconomical. thus, the need for traffic analysis. So, in the situation where an incoming call finds all channels busy, the call would be blocked. There are two kinds of systems with different reactions to call block: (Teletraffic engineering, 2005) Loss Systems and Waiting Systems.

In these Loss systems, call-block is handled by simply refusing to service calls that arrive when the lines are all occupied. Delay system: In this system, the blocked call waits in the system until when resources are available. Conventional telephone systems are Loss-like. In this research, attention is focused on loss blocking. while for a delay system, the probability of waiting. The Erlang formula is used to determine the GOS of loss systems (that is, the blocking probability) having N channels, with offered traffic A.

The erlang formula is also useful for the determination of any of the three parameters when the other two are known. The erlang formula is based on the following assumptions:

i. Occurrence of calls are independent
ii. That calls that arrive at a switching center have a poison distribution
iii. Calls are served in the order of arrival.

Based on the assumptions above stated, there are three models of loss systems, they are:
i. Lost calls cleared (LCC)
ii. Lost calls returned (LCR)
iii. Lost calls held (LCH)

There are times we dial a number and we receive the "Network Busy" message. When this happens, it means all the lines are presently in use and there is no available line momentarily. Your call is therefore terminated until you reinitiate the process again, which is then assumed to be a new call. This is how an LCC system operates.

The first person to account fully and accurately for the effect of cleared calls in the calculation of blocking probabilities was A.K. Erlang in 1917 (Teletraffic engineering, 2005). In an Erlang loss system, three variables are involved and are constantly manipulated. When two variables are known, the third can then be seamlessly computed.
i. The number of Channels
ii. The GOS
iii. The Traffic Intensity

According to erlang, the probability of loss is given by:

$$P(N) = B(N, A) = \frac{A^N}{N! \sum_{k=1}^{n} \frac{A^K}{k!}} \quad (3)$$

Equation 3 above is referred to as the Erlangs formula of the first kind, the Erlang-B formula or Erlangs loss formula (Ally j,) In telecommunications network design and dimensioning, it is necessary to find the number of trunk lines needed for a given offered traffic and a specified grade of service.
Utilization or trunk occupancy, on the other hand, is the carried load per channel. This, in the data, is expressed as:

$$Utilization = \frac{MaximumBusyChannel}{TotalChannel} \quad (4)$$

Hence, we shall be using the Erlang-B calculator to dimension two Interconnect clearing houses in Nigeria, in order to determine the capacity (number of channels) needed to carry the busy-hour traffic with insignificant or no blocking, while keeping the QOS to standard (Teletraffic Engineering, 2005).

4. Traffic Measurement and Data Collection

Data collected from two licensed interconnect operators, named X and Y (for anonymity) were processed and computed. Traffic engineering schemes were employed to perform data and load analysis in order to determine the interconnection traffic at peak period for a service provider and the infrastructure or bandwidth needed to maintain it without terminating other lines and affecting QoS.

Data from Interconnect Clearinghouse X was taken during the busy hour period of 18:00 - 19:00 hours, from the 1st to the 28th February 2016. The summarized data, which is used for our analysis excludes the identity of all the operators involved. Only the total traffic routed into and out of the facility at the defined busy hour is here represented, for anonymity reasons, as seen in table 1 below:

Table 1: Data from clearinghouse X

CLEARING HOUSE X

TIME: 18:00 Hrs - 19:00 Hrs

DATE	CC:I	CSS:I	CC:O	CSS:O	TOTAL CHANNEL	CONGESTION	CONGESTION % (GOS)	UTILIZATION %	TRAFFIC INTENSITY	ASR	
										INC %	OUTG %
2/1/2016	53761	22297	32536	13669	1522	81	0.25	30.36	813.40	41.47	42.01
2/2/2016	52251	21231	31789	13081	1522	91	0.29	25.47	794.73	40.63	41.15
2/3/2016	44837	18812	32844	13869	1522	65	0.20	28.32	821.10	41.96	42.23
2/4/2016	45204	18749	34738	14473	1522	100	0.29	28.42	868.45	41.47	41.66
2/5/2016	44921	18905	38102	15942	1522	77	0.20	29.84	952.55	42.08	41.84
2/6/2016	44332	18176	44307	18176	1522	120	0.27	32.6	1107.68	41.00	41.02
2/7/2016	39990	15897	39968	15897	1522	64	0.16	29.65	999.20	39.75	39.77
2/8/2016	43688	18536	43665	18536	1522	76	0.17	29.55	1091.63	42.43	42.45
2/9/2016	44422	18847	22227	9266	1522	49	0.22	24.86	555.68	42.43	41.69
2/10/2016	43151	18262	22321	9294	1522	43	0.19	22.95	558.03	42.32	41.64
2/11/2016	55858	23951	28674	12150	1522	51	0.18	29.49	716.85	42.88	42.37
2/12/2016	56333	24016	29011	12203	1522	58	0.20	29.02	725.28	42.63	42.06
2/13/2016	54879	21345	28998	11696	1522	54	0.19	24.77	724.95	38.89	40.33
2/14/2016	54066	21306	28049	10901	1522	31	0.11	29.23	701.23	39.41	38.86
2/15/2016	65910	27397	41467	16993	1522	75	0.18	40.6	1036.68	41.57	40.98
2/16/2016	78045	31071	51692	20729	1522	104	0.20	46.01	1292.30	39.81	40.10
2/17/2016	45447	17482	41208	15954	1522	73	0.18	39.12	1030.20	38.47	38.72
2/18/2016	52848	20281	52622	20199	1522	52	0.10	39.65	1315.55	38.38	38.39
2/19/2016	60786	23710	60565	23628	1522	48	0.08	38.97	1514.13	39.01	39.01
2/20/2016	47351	18680	47167	18608	1522	39	0.08	28.95	1179.18	39.45	39.45
2/21/2016	41441	15764	41284	15708	1522	27	0.07	27.8	1032.10	38.04	38.05
2/22/2016	46351	18490	46129	18421	1522	50	0.11	28.24	1153.23	39.92	39.93
2/23/2016	77441	28065	77335	28030	1522	387	0.50	47.74	1933.38	36.24	36.24
2/24/2016	111360	39262	107165	39260	1522	540	0.50	56.1	2679.13	35.26	36.64
2/25/2016	114536	41009	114518	41009	1522	221	0.19	55.15	2862.95	35.80	35.81
2/26/2016	109077	38165	109075	38165	1522	230	0.21	49.33	2726.88	34.99	34.99
2/27/2016	117750	41094	117746	41094	1522	142	0.12	73.31	2943.65	34.90	34.90
2/28/2016	129494	45128	129467	45128	1522	51	0.04	65.82	3236.68	34.85	34.86
SUMMARY	63411.79	24497.43	53381.04	20431.39	1522.00	107.11	0.20	36.83	1334.53	39.50	39.54

$$TrafficIntensity = \frac{VolumeofCalls}{CallPeriod} = \frac{C \times Th}{T} \quad (5)$$

For this analysis, call holding time is taken to be 90 seconds = 1.5minutes

Therefore, $Th = 90\sec s \equiv 1.5\min s$

$$C = 53381 \quad , \quad T = 60\min s \equiv 1hour$$

$$TrafficIntensity = \frac{53381 \times 1.5}{60}$$

$$= \frac{80071.5}{60}$$

$$TrafficIntensity = 1334.525Erlangs$$

$$GradeOfService = \frac{CallsOffered - CallsCarried}{CallsOffered}$$

.

$$GOS = \frac{LostTraffic}{OfferedTraffic} = \frac{BlockedBusyHourCalls}{OfferedBusyHourCalls} = \frac{A - Ao}{A}$$

$$AverageGOS = 0.2\%$$

Table 2: Data from Clearinghouse Y

			CLEARING HOUSE Y							
			DAILY PEAK HOUR TRAFFIC ANALYSIS		TIME: 19:00 Hrs - 20:00 Hrs					
DATE	CC:I	CSS:I	CC:O	CSS:O	TOTAL CHANNEL	CONGESTION	CONGESTION % (GOS)	UTILIZATION %	ASR	
									INC %	OUT %
4/1/2016	13856.83	5109.25	11417.38	4114.17	523.00	169.00	1.48	34.18	36.87	36.03
4/2/2016	9938.00	3606.79	10496.13	3902.92	523.00	101.00	0.96	22.70	36.29	37.18
4/3/2016	11381.21	4315.00	11448.13	4259.21	523.00	99.00	0.86	28.63	37.91	37.20
4/4/2016	10224.00	3634.58	21329.83	8241.38	523.00	132.00	0.62	22.70	35.55	38.64
4/5/2016	11905.54	3055.63	12511.33	3298.04	523.00	198.00	1.58	34.18	25.67	26.36
4/6/2016	12081.08	1720.17	13685.79	2234.08	523.00	119.00	0.87	22.70	14.24	16.32
4/7/2016	10617.71	2609.17	2163.46	1006.54	523.00	71.00	3.28	22.70	24.57	46.52
Summary	**11429.20**	**3435.80**	**11864.58**	**3865.19**	**523.00**	**127.00**	**1.38**	**26.83**	**30.16**	**34.04**

5. Results

The Erlangs-B calculator is utilized to compute the infrastructure (number of channels) needed to maintain the measured-offered traffic without termination other lines and affecting QOS. After computation, we obtain the following results for the ICN clearing house:

- In order that the facility may support an average Busy Hour Traffic Intensity of 1,334.525E
- Given a standard GOS of 2%,
- The Number of Lines/Channels conveniently required for this are 1403 lines.

Plots for Clearinghouse X

Figures 1 below shows the graphical representation of total, carried and answered calls, while Fig.2 shows the system utilization with respect to the threshold mark of 70%, set by the regulator,

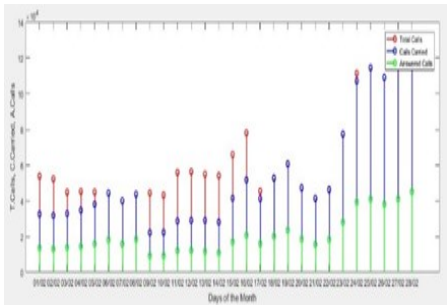


Fig.1: Total, Carried & Answered Calls

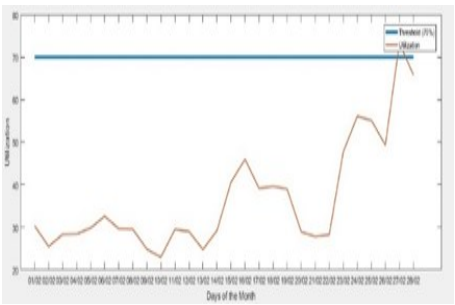
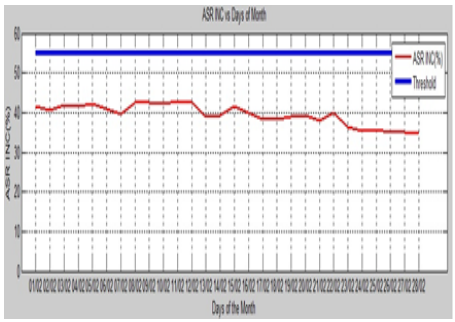


Fig.2: System Utilization

Figures 3 & 4. below, are the graphs for Answer-to-seizure ratios (ASR) for traffic coming in and out of the exchanges, with respect to the 55% (INC threshold) and 65% (OUTG threshold) values set by the regulator.



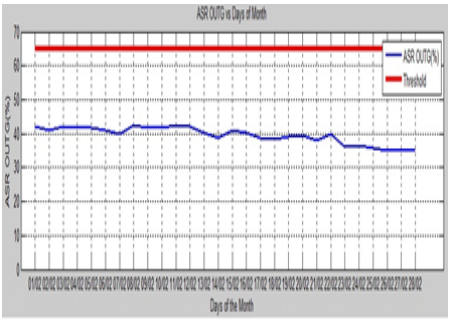


Fig.3: Answer-to-Seizure Ratio (INC) Fig.4: Answer-To-Seizure Ratio (OUTG)

Figure 5. below shows the plot for GOS obtained from the CH, with respect to the 2% standard threshold.

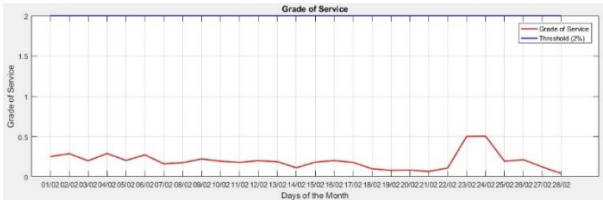


Fig.5: Grade of Service

Plots for the Clearinghouse Y

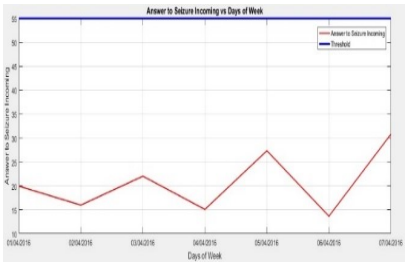


Fig.6: Answer-to-Seizure ratio (INC)

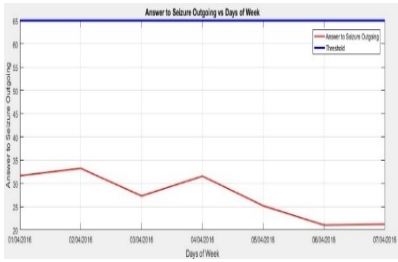


Fig.7: Answer-to-Seizure Ratio (OUT)

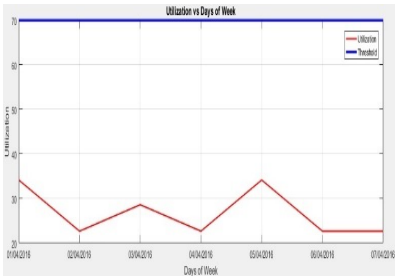


Fig.8: System Utilization

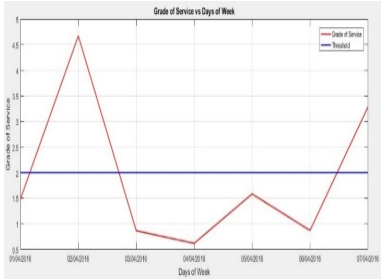


Fig.9: Grade of Service

Discussion

As can be read from the data for clearinghouse X, the average number of channels available to route the busy hour traffic for the month of February 2016 was 1,522 channels, having obtained an average of 2% GOS. On using Erlangs calculator, the ideal value needed to route the equivalent amount of traffic was obtained to be 1,403 Channels. Thus, the difference in the gap to be filled by this CH is 1,522 – 1,403= 119 channels. This means the facility has an excess of 119 unused channels that were not used to route the traffic that was passed through the facility in the whole month. To maintain the stated amount of traffic without termination other lines and affecting QOS, the facility requires 1,403 channels or more.

Similarly, for clearinghouse Y, the average number of channels available to route the busy hour traffic was 523 channels, with the GOS of the facility obtained to be 1.38%. On using Erlangs calculator, the ideal number of channels needed to route the equivalent amount of traffic was obtained to be 316 Channels. Thus, the difference in the gap to be filled by this CH is 523 – 316= 207 channels. This means the facility has an excess of 207 unused channels that were not used to route the traffic that was passed through the facility at the period data was collected. To maintain the stated amount of traffic without termination other lines and affecting QOS, the facility requires 316 channels.

6. Conclusion

Nigeria has achieved so much growth and advancement in the telecoms sector especially in the past decade, however, more effort need to be put in to ensure that as forerunners and leaders in the growth and development in this sector in Africa, are not inhibited and overtaken by other countries due to lack of vision political undertones. The regulations guiding the CHs be reviewed and the facilities constantly monitored, to enable them handle interconnect are non-discriminatory and independent manner, while carrying the confidence of both regulator and operators especially in matters relating to money. Finally, the policy implication of this study suggests that; for the CH to be fully effective, the Federal government, through the regulator should ensure that all operators are persuaded or coerced to route all their off-net traffics through these facilities (as obtained in Ghana), although this would require an equivalent increase in the capacities of the CHs.

Correspondence
E.S. Dikko
Department of Telecommunications Engineering
Federal University of Technology, Minna, Nigeria

References

Ally J, Switching Systems, Dar es Salaam institute of Technology (DIT), Traffic Engineering and Network Planning. [Online] Available: http://www.slideshare.net/JumanneAlly/switching-systems-lecture3

Bharat Sanchar Nigam Limited, National Centre for Electronic Switching, (2007); *Maintenance support (AMC for hardware and software) for EWSD New Technology Switches: Improvement in CCR (call completion ratio)*.

Chettinad College of Engineering and Technology, Department of Information Technology Communication Switching Techniques, Unit 4, Traffic Engineering. [Online] Available: 27-1286-samridhi.pdf

Circuit switching: Traffic Engineering, [Online] Available: www.eie.polyu.edu.hk/ ~em/dtss05pdf/Traffic%20Engineering.pdf

Communications Week, NCC Blames Exchanges, Others for High Indebtedness. (14 December, 2012)

Etedgenews, (2013, July 12), Nigeria's interconnect debt trap, Retrieved from: http://www.etedgenews.com/index.php/k2/internet/item/2500-nigeria%E2%80% 99s-interconnect-debt-trap/2500-nigeria%E2%80%99s- interconnect-debt-trap? start=110/

InterComms - The International Communications Project, different models for running your interconnect accounting solution. [Online] Available:

http://www.intercomms.net/AUG03/content/azurebrowne.php

Introduction to switching systems, [Online] Available: http:// www.newagepublishers.com/samplechapter/000969.pdf

Iversen, V. B. (2015). Teletraffic engineering and network planning, [Online] Available: orbit.dtu.dk/files/118473571/Teletraffic_342_V_B_Iversen_2015.pdf

Medallioncom.com

NCC, (2016), Summary of The Commission's Compliance Monitoring and En- forcement Activities for Quarter One. [Online]Available: http://www.ncc.gov.ng/ index.php?option=com_content&view=article&id=1259&Itemi=214

Okechukwu C. Ugweje, (2013); Lecture notes, CME621 Random and Stochastic Processes Teletraffic Engineering http://www.slideshare.net/mazlina1202/lecture5- teletraffic ITU–D, Handbook, Teletraffic Engineering. January 2005, [Online] Available: http://www.slideshare.net/deepakksinghagra/ teletrafficengineeringhandbook?qid=e3480e7d-0519-4494-9f22- 7a7122055995&v=&b=&from_search=2

The communicator, In Favour of the Interconnect Clearinghouse Model, June 2016, http://www.ncc.gov.ng/thecommunicator/index.php? option=com_content&view=article&id=807&Itemid=67
Tutorials point on Telecommunications interconnection. [Online] Available: *http://www.tutorialspoint.com/telecom-billing/*

Wikipedia, Network planning and design, [Online] Available: https:// en.wikipedia.org/wiki/Network_planning_and_design

Proceedings on Big Data Analytics & Innovation (Peer-Reviewed), Volume 1, 2016, pp.212-221

A Grouped Half-Life Variable Quantum Time Round Robin Scheduling (GHLVQTRR) Algorithm for CUP Process

SALISU IBRAHIM YUSUF
Baze University Abuja, Nigeria

S. E. ABDULLAHI
Nile University of Nigeria

ABSTRACT The computing requirement for big data analytics has been reduced tremendously through the use of parallel and distributed computing in enhancing multiprogramming in operating systems, various scheduling algorithms are used to manage jobs in the system, and however, the optimal scheduling algorithm is required. The half-life variable quantum time round robin (HLVQTRR) scheduling algorithm uses half the burst time of individual job as quantum time (QT) for execution. For each job, its QT has to be computed increasing the computation of the entire system with a context switch of the number of jobs. Moreover, the waiting time for very small jobs will be relatively large affecting the overall throughput. In this research, HLVQTRR has been modified to iteratively group jobs based on their burst times less than the average of the set of ungrouped processes. Thereby execute very small process without breaking thus reducing context switching and improving the throughput. Hence, a grouped HLVQTRR (GHLVQTRR) has been proposed improving the resource management in operating systems.

Keyword: Operating Systems, Process Scheduling, Round Robin Scheduling Algorithm, Quantum Time, context switching, throughput

1. Introduction

Operating systems are the intermediary between computer components and users, it allocates resources to user processes, CPU is the most important and limited resource in a computer system as all program instances are executed by the processor. To effectively use the CPU a form of scheduling is required. The basic scheduling algorithms are: Shortest Job First (SJF), Longest Job First (LJF), First Come First Serve (FCFS), priority based and round robin (RR). The round robin (RR) allocates quantum time (QT) or time slice for processes in the ready queue in a turn -by turn manner. This approach has been proven to be more effective, the challenge over time has been the choice determining the QT of a set of processes. The larger the QT the more the system behaves like the first come first serve (FCFS) approach, the smaller the QT the higher the number of context switching. Varia-

tions of RR algorithms proposed have been proven to better than other scheduling algorithms such as the SJF, FCFS, LJF and their variations in a real time processing environment based on the average turnaround time, response time or CPU Utilization among others (Abbas, et al., 2011) (Dawood, 2012) (Ashiru, et al., 2014). In an exponentially distribution of process burst times (Abur, et al., 2011) simulation shows that the RR resulted in a minimal average waiting time. The average waiting time, the average turnaround time and context switching are the most considered factors for comparing RR algorithms.

1.1 Preliminaries

Programs are executed as series of processes which are instances of programs in execution. Process much like living things has a life cycle *from cradle to grave*, processes are created sometimes by other parent processes and may create a child process and eventually terminate or die. Through the life time of a process it could be in the following states

- New: birth of the process
- Running: process is being executed
- Waiting: process at hold for an event
- Ready: waiting for processor assignment.
- Terminated: processes has been executed or killed

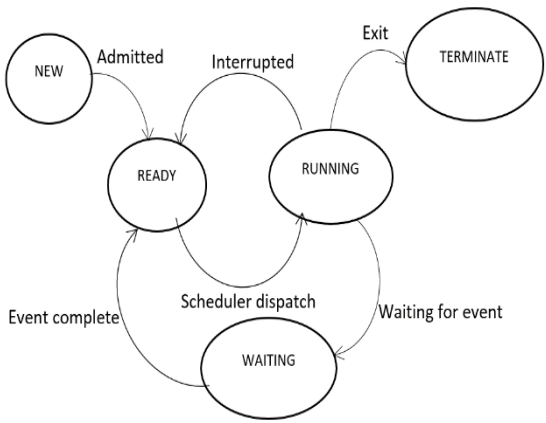


Figure 1: Process Life Cycle

Process after creation is at the new state, it is admitted in the ready queue for CPU allocation. While running state is when the process is being executed by the CPU which may be interrupted for a number of reasons such as the input or output events or context switching, at this instance the process is said to be in a waiting state upon completion of the event the process is registered back into the ready queue, furthermore, process could execute completely transiting from running to termination or killed for some reasons.

1.2 Scheduling Criteria

- CPU Utilisation: The degree of usage of the CPU, a good scheduling algorithm is to keep the CPU 100% busy with no idle times
- Throughput: number of process executed at a given period of time.
- Turnaround time: total time spent by the process for entering the queue to its final execution.
- Waiting time: total time spent in the system minus the execution time(s). That is time spent in the waiting state.
- Response time: time taken until the process gets its first CPU allocation.
- Number of context switching: number of time a process is paused for another process to execute.

2. Motivation

Deciding the value of the QT of time slice is the major challenge in the RR algorithm. Having very small time slice will lead to many context switching and less process executions deteriorating the utilisation of the processor. As processes have varying burst times having a static QT with larger burst times will experience many context switching before termination while others might need controlled pre-emption as their burst time is more than the QT. The use of dynamic QT seems to be reasonable but given individual process its QT will add to the cost of context switching. By grouping process and assigning a QT for each will reduce context switching and improve the turnaround time – thus, yielding a better result.

3. Related Work

The RR algorithm has been proven over time to be better than other basic scheduling algorithms, these make the algorithm enjoy the attention of researchers for optimisation among these are: Dynamic Round Robin with controlled pre-emption (DRRCP) here the remaining time required for a process to complete execution is considered before context switching, if the process requires a little time it is allowed to execute fully (Ashiru, et al., 2014).

The Half Life Variable Quantum Time Round Robin (HLVQTRR) take half the burst time of each process as its QT therefore having number of context switching equal to the number of processes (Ashiru, et al., 2014); Priority based Round Robin CPU scheduling algorithms for real time system, combines the RR with priority based scheduling (Singh & Deepa, 2012).

Variable Quantum Time (VQT) algorithm which is based on averaging technique to allocated a variable quantum time (QT) to each process in the ready queue (Yashasvini, 2013)

A New Proposed Dynamic Quantum Time with Re-Adjusted RR Scheduling Algorithms and its performance analysis proposed the use of the median after first round of execution the QT is recalculated (Abbas, et al., 2011). In Even Odd Round Robin (EORR), there are two QT (QT1 and QT2). QT1 is the average of processes that are in odd position in the ready queue, while QT2 is the average of

processes in even position in the ready queue. QT1 is compared to QT2 and the greatest is use as the QT in that round (Pallab, et al., 2012)

In Average Mid Max Round Robin (AMMRR), quantum time is the mean of the summation of the average and the maximum burst time of the processes in the ready queue in each cycle.

As for Ascending Quantum Minimum and Maximum Round Robin (AQMMRR), QT is calculated by multiplying the summation of the minimum and maximum CPU burst by 80 percent (Dawood, 2012). Multi Dynamic Quantum time Round Robin (MDQTRR) uses two different QT in single cycle. Up to the median process, the quantum time used is gotten using the median quartile formula MQT (Median Quantum Time) while for the succeeding processes, the Upper Quartile formula is used to calculate the quantum time, UQT (Upper Quartile Quantum time) (H.S, et al., 2011)

4. Proposed Approach

The HLVQTRR as proposed executes a process completely in two iterations by using half the burst time of each process as the QT for that process execution, hereby leading to *n* number of context switching for n number of processes. The GHLVQTRR is an optimisation of the HLVQTRR. The proposed approach aims to reduce number of context switching and improve throughput maintaining the two-iteration properties of HLVQTRR. Processes in the ready queue are grouped recursively based on processes' burst time, average *avg* of the ungrouped processes is computed, for processes with burst time less than the average are grouped together and *1/2avg* is the QT dedicated to the grouped processes, this implies that process that are less than 1/2 the average of the processes will execute without context switching, thus improving throughput of the system, the grouping will repeat recursively until the all processes are grouped and assigned a QT for execution.

5. Illustration

For analysis, the proposed GHLVQTRR will be compared with the traditional HLVQTRR

Algorithm 1: GHLQTRR

Case 1:
Set of processes in the ready queue.
{p1 =88; p2=89; p3=85; p4=93; p5=90; p6=84; p7= 90}
Retrieved from the HLVQTRR (Ashiru, et al., 2014)

HLVQTRR
The algorithm calculates the QT of each process by taking half its burst time
P1: TQ = 88/2 = 44, P2: TQ= 89/2 ≈ 45, P3: TQ= 85/2 ≈ 43, P4: TQ = 93/2 ≈ 47, P5: TQ= 90/2 = 45, P6: TQ= 84/2 = 42, P7: TQ = 90/2 = 45

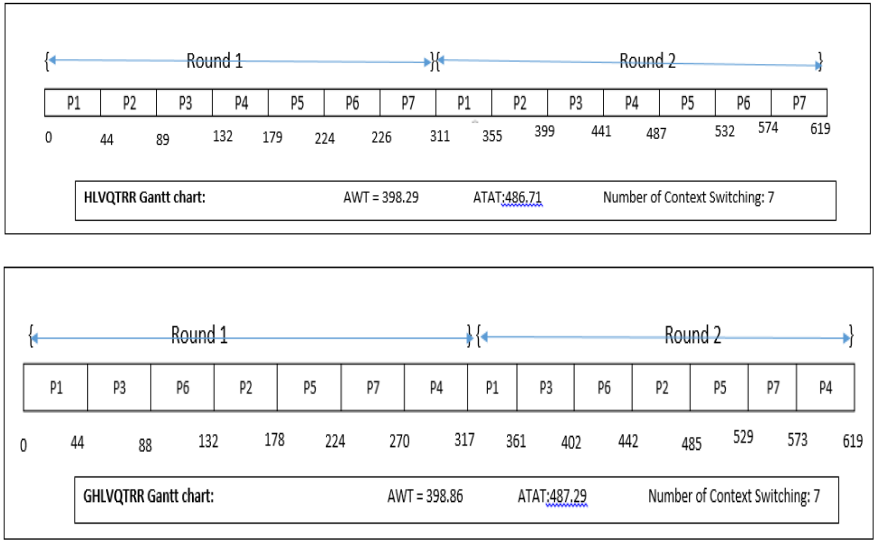


Figure 2: Gantt chat for case 1

Upon execution, each process is run for half its burst time as its QT, in the second round system behaves like FCFS approach.

GHLVQTRR

For this approach processes are grouped and each group assign a QT.
Average = 88
P1, P3, and P6 are less or equal to 88 therefore there QT is average / 2 = 44
P2, P4, P5, and P7 make an average of 91, hence P2, P5, and P7 grouped with a QT of 46
Finally, P4 has QT = 47
The limitations of the HLVQTRR algorithm is not visible with the sample process burst time where the processes with less variance.

ANALYSIS

Comparing both algorithms based on the calculated criteria the algorithms can be assumed of to be the same in terms of waiting time, turnaround time, number of context switching. In case 2 where the mean deviation is very high the difference is more defined.

Case 2:

Set of processes in the ready queue
{P1 = 5; P2 = 7; P3 = 8; P4 = 10; P5 = 9; P6 = 7; P7 = 3; P8 = 5; P9 = 4; P10 = 8; P11 = 10; P12 = 9; P13 = 2; P14 = 6; P15 = 14; P16 = 13; P17 = 11; P18 = 16; P19 = 20; P20 = 21}

HLVQTRR

For quantum time.

P1: TQ = 5/2; P2: TQ = 7/2; P3: TQ = 8/2; P4: TQ = 10/2; P5: TQ = 9/2; P6: TQ = 7/2; P7: TQ = 3/2; P8: TQ = 5/2; P9: TQ = 4/2; P10: TQ = 8/2; P11: TQ = 10/2; P12: TQ = 9/2; P13: TQ = 2/2; P14: TQ = 6/2; P15: TQ = 14/2; P16: TQ = 13/2; P17: TQ = 11/2; P18: TQ = 16/2; P19: TQ = 20/2; P20: TQ = 21/2.

Evaluation

Average waiting time (AWT) = 127.33

Average Turnaround Time ATAT = 133.53

Number of context Switching = 20

CPU Utilisation = 100%

GHLVQTRR

Iteration 1:

Average burst times = (5+7+8+10+9+7+3+5+4+8+10+9+2+6+14+13+11+16+20+21) / 20 == $\lceil 9.4 \rceil$ 10,

QT $\lceil 9.4 \rceil$ 1 = 5

Grouped processes: P1, P2, P3 … P14 have burst times of at most 10.

QR = P15, P16, P17, P18, P19, P20

Iteration 2:

Average burst times of QR = (14+13+11+16+20+21) / 6 = $\left\lceil \frac{95}{6} \right\rceil$ = 16

Q $\left\lceil \frac{95}{6} \right\rceil$ T2 = 8.

Group processes: P15, P16, P17, and P18 have burst times less than or equal 16.

RQ = P19 and P20.

Iteration 3:

Average or RQ = (20 + 21) / 2 = $\lceil 20.5 \rceil$ = 21, Q $\lceil 20.5 \rceil$ T3 = $\lceil 10.5 \rceil$ = 11.

G $\lceil 10.5 \rceil$ roup processes: P19 and P20.

RQ is empty.

Evaluation

Average waiting time (AWT) = 108.15

Average Turnaround Time ATAT = 127

Number of context Switching = 15

CPU Utilisation = 100%

ANALYSIS

In this case, the processes tend to have larger variance with sizes ranging between 2 and 21. Using the HLVQTRR every process will experience context switching thus 20 context switching happened while executing processes. Whereas for pro-

posed GHLVQTRR three groups created at the end of grouping. While executing the processes, after the first iteration, all process with burst times less than the average were executed fully, this results to less context switching than HLVQTRR. The average waiting time for HLVTQTRR was 127.33 and 108.15 for GHLVQTRR, this result from relatively smaller processes execute completely in the first round of execution reducing their total waiting time those processes. Furthermore, a good improvement was observed in the turnaround time implying that the proposed GHLVQTRR has higher throughput compare to HLVQTRR.

Discussion

This study was made to optimise the HLVQTRR algorithm which was proposed and was proven to be better than the classic RR using average as the QT. after comparing the proposed optimised algorithm GHLVQTRR and HLVQTRR, it was observed that the variance of the burst times of the processes affects the behaviour of the algorithms, as seen in the first case, the HLVQTRR was better in terms of AWT by 0.57 time unit and GHLVQTRR has 0.58 of time unit worse than HLVQTRR with the same number of context switching, the processes range between 88 and 93 in case 1. For the second case processes with wider range and larger variance was considered and it was observed that the proposed optimisation was very much better than the initial algorithm as discussed in the latter sub heading. Comparing the two cases outcomes the GHLVTQRR should be considered as an optimisation of the original HLVQTRR as it behaves better with process set.

6. Conclusion

An optimisation of HLVQTRR was proposed and proven to be optimised using sample data of different variance. Some limitations and trade-offs of the proposed GHLVQTRR are as follows:

- Overhead computing for the grouping process.
- Reduced multi programming compared to the initial HLVQTRR
- Behaves like HLVQTRR with narrow variance.
- Batch system of processing was used for both case, that is, all process arrived at RQ at time 0.

7. Recommendation

For future work, researcher should consider the following:

- Processes with different arrival times.
- Using a computer simulation to evaluate the algorithm.

Correspondence
Salisu Ibrahim Yusuf
Department of Computer Science
Baze University Abuja Nigeria
Email: salisu.yusuf@bazeuniversity.edu.n

S E Abdullahi PhD
Department of Computer Science
Nile University of Nigeria
Email: Saleh.abdulllahi@nileuniversity.edu.n

References

Abbas, N., Ali, K. & Seifedine, K., 2011. A New RR based Schedulling Algorithm for Operating Systems: Dynamic Quantum Using the Mean Average. *IJCSI International Journal of Computer Science Issues,,* 8(3), pp. 224-229.

Abur, M., Muhammad, A., Danjuma, S. & Abdullahi, S., 2011. A Critical Simulation of CPU Scheduling Algorithm using Exponential Distribution. *International Journal of Computer Science Issues,* 8(6), pp. 201-206.

Ashiru, S., Abdullahi, S. & Junaidu, S., 2014. Half Life Variable Quantum Time Round Robin (HLVQTRR). *International Journal of Computer Science and Information Technologies,* 5(6), pp. 7210-7217.

Ashiru, S., Saleh, A. & Sahalu, J., 2014. Dynamic Round Robin with Controlled Preemtion. *International Journal of Computer Science Issues,* 11(3), pp. 109-117.

Dawood, A. J., 2012. Improving Efficiency of Round Robin Scheduling Using Ascending Quantum And Minumim-Maxumum Burst time. *J. of university of anbar for pure science,* 2012(6).

H.S, B., Rakesh, M., Sabyasachi, S. & Sourav, B. K., 2011. Comparative performance analysis of Multi-dynamic Quantum time Round Robin (MDQTRR) Algorithm with Arrival Time. *Indian Journal of Computer Science and Engineering (IJCSE),* 2(2), pp. 262-271.

Pallab, B., B., P. & D.S., S., 2012. Comparative Performance Analysis of Even Odd Round Robin Scheduling Algorithm (EORR) Using Dynamic Quantum Time with Round Robin Scheduling Algorithm using Static Quantum Time. *International Journal of Advanced Research in Computer Science and software engineering,* Volume 2, pp. 62-70.

Singh, R. I. & Deepa, G., 2012. A priority base round robin CPU schedulling algorithm for real time systesms. *International Journal of Innovations in Engineering and Technology (IJIET),* 1(3), pp. 1-11.

Yashasvini, S., 2013. Determining the Variable Quantum Time (VQT) In Round Robin and importance over Average. *International Journal of Science, Engineering and Technology Research (IJSETR),* 2(3), pp. 612-617.

Instructions for Authors
All manuscripts will be submitted via email. Papers will be reviewed in the order in which they are received.

Preparation of the Paper: There is no particular standard pattern of presentation, but a research-based paper will include elements such as an introduction, stating aims and objectives of the research and a review of previous work; a statement of the sample, procedure and tests; a factual statement of the results and a discussion of their significance, stating implications for policy or practice.

Peer Review Process: All research articles in the *Proceedings on Big Data Analytics &Innovation (peer-Reviewed)* will undergo peer review, based on initial editor screening by at least two anonymous referees. Papers accepted for publication will be subject to copyediting.

Preparation of Manuscript: A manuscript submitted to the *Proceedings on Big Data Analytics &Innovation (peer-Reviewed)* should be a Microsoft Word file. Manuscripts should be formatted into a 210 x 297mm (A4) document with one inch margins, should be Times New Roman 12 point type-face, and should not be more than 10,000 words (including references).

***Title/Abstract Page*:** The title page should contain the article name, authors' names and the address for manuscript correspondence (including e-mail address, telephone, and or fax numbers). This page should also contain a descriptive abstract of no more than 200 double-spaced words. Author's name should appear only on the title page; any indication to the author's identity in the text should be deleted.
References should be listed in alphabetical order of surnames at the end of an article and should generally accord with the **Harvard styles**. The form in which references are given should be:

For articles: Craft, A. (1998) *Educator Perspective On Creativity: An English Study*, Journal of Creative Behaviour, 32, 4, pp. 244-256.

For books: Ainley, P. and Bailey, B. (1997) *The Business of Learning: Staff and Student Experiences of Further Education in the 1990s*, London, Cassell.

For chapters within books: McNay, I. (1995) 'From the Collegial Academy to Corporate Enterprise: the Changing Cultures of Universities' in T. Schuller (ed). *The Changing University?* Buckingham, SRHE and Open University Press.

Tables/figures/illustrations/photographs: Tables should be typed out on separate sheets and not included in the actual text. One set of (finished) artwork for figures should be supplied in a form suitable for reproduction.
Authors are responsible for clearing copyrights issues of any materials used in the manuscript (e.g. figures, tables, illustrations, photographs, etc.)
Proofs will be sent to author/s if there is sufficient time to do so. They should be corrected and returned to the Editors within five working days. Major alterations to the text cannot be accepted.
Offprint (in PDF) will be supplied free of charge to authors via email.
Copyright: Individual authors. Authors may use the manuscript elsewhere after publication without prior permission from the publishers, KIE Publications, provided that acknowledgment is given to the *Proceedings on Big Data Analytics &Innovation (peer-Reviewed)* as the original source.

Papers should be sent electronically to: Conferenceteam@kiecon.org.

Papers not adhering to these submission guidelines are subject to being rejected without review.

Proceedings on

Big Data Analytics & Innovation (Peer-Reviewed)

9 781859 242223 >