# Introduction

## Cloud Computing

Computing and technology continuously advance in the Information Technology industry to improve processes that assist in meeting business and consumer objectives and improving consumer satisfaction. However, innovated resources do tend to have higher upfront costs. Therefore, not many businesses or consumers can afford to purchase these resources. In addition, with the emergence of the internet in this generation, the expense of storage as well as the consumption of power by computing components has increased (Duman et al., 2016). Thus, research into the idea of utilizing resources from an innovative and powerful vacant computer remotely immerged. This concept was referred to as Cloud Computing.

Cloud Computing according to the National Institute of Standards and Technology (NIST) is a computing model that exists for the enablement of “ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” (Mell & Grance, 2011). In hindsight, cloud computing can be viewed as applications that are delivered through the internet as a service to the consumer as well as the reference to hardware and software of data centers that supply these services (Armbrust et al., 2010). The services are categorized into the following different models according to NIST (Mell & Grance, 2011):

* Software as a service (SAAS), is applied through either a program interface or a web browser to use the supplier’s applications that are running on the infrastructure of the cloud.
* Platform as a Service (PaaS), is applied where the consumer deploys a consumer-created application that is supported by the supplier, whereby the user has control of application configuration.
* Infrastructure as a Service (IaaS), is applied when the consumer is equipped with deployment and running of arbitrary software and can, therefore, control the deployed software and storage, in addition, can also have limit control to network components.

For the purpose of this paper, we will look at Software as a Service in closer detail.

Cloud computing model is subdivided into 4 deployed infrastructures:

* Public Cloud, exists in the premises of a cloud provider which is open to use of the general public.
* Private Cloud, can exist in either the premises of an exclusive organization or off on a third party organization, which is used exclusively by a single organization or jointly with a third party organization. As long as it's not open to use by the public.
* Community Cloud, is utilized by organizations that share common interests, security or concerns. The environment exists either on the premise of either of the organizations or a third party organization rendering the cloud services to this community of organizations.
* Hybrid Cloud, is the combination of either of the previously stated deployed infrastructures distinctively.

The public cloud is of particular interest to the purpose of this paper, especially when looking into the commercial service platform that exists in the public cloud such as Google, Amazon and Microsoft. The public cloud service infrastructure has rapidly grown, therefore has also brought forth an increase in unavoidable security issues. In essence, cloud security is a vital vulnerability that is of great concern, and thus even today we still find many organizations, that have not utilized the available cloud services. According to Gartner, a researcher that conducted an investigation around information security issues that are considered in cloud computing, listed that privileged access is one of the issues highlighted that public cloud vendors need to consider (Shaikh & Haider, 2011). For the end consumer, sensitive data should only be available to them however cloud vendor administrators should have access to cloud environment for maintainability and security of environment. Thus sensitive data may also need to be available to administrators.

## Impersonation

To enable the availability to end user, Cloud computing service needs to provide an identity management. This is however a great concern in the public cloud environment as the cloud providers identity management are not consistently integrating their platforms with identity services (Brodkin, 2010). This is also due to the constant evolution of SaaS. This opens to a renowned cyber crime attack that is probable in exposing the vulnerability. This crime is known as Phishing. Phishing is executed through someone impersonating as a legitimate and reputable organisation or individual targeting any individual or organization to providing sensitive data (Aravindhan et al., 2016). Therefore, attackers performing Phishing attacks usually perform them through email, where they prompt target into giving sensitive information. Attackers also target public cloud infrastructure through PaaS where they inject links that users using Cloud platform as a SaaS would click the link that would redirect user to a fraudulent web page.

Anti-phishing tools have been developed and utilized by public cloud providers in attempt to prevent Phishing attacks. Tools like Envelope content splitting (ECS) tool is used to authenticate the attacker sending a link of a fraudulent website, and so protects the recipient from phishing (Aravindhan et al., 2016). There exists tools that also detect Phishing attacks, such as a Phishing tank whereby a database containing Phishing sites are used by other anti-phishing tools to detect phishing. The literature review will focus on some of the anti-phishing tools that were researched for the purpose of detecting Phishing, particularly in the Public Cloud. This will be followed by the identifying the gap in detection of Phishing attacks that this research paper will address. Authorship detection is the proposed solution based of a Stylometric analysis approach that this paper will investigate.

# Problem Statement

# Literature Review

A paper done by Nazmul Islam, Mohammed Moshiul Hoque and Mohammad Rajib Hossain on the “Automatic Authorship Detection from Bengali Text using Stylometric Approach” (Islam, Hoque & Hossain, 2017) investigates writing styles of Bengali writters by collecting writings and blogs from their sample writers. Through analysis they discovered n-gram features that were useful to detect certain authors. N-grams, also reffered to as shingles, is an adjacent sequence of n items from a given sample of text (Broder et al., 1997). Thus can be viewed as a probabilistic model in a language base that predicts the next item in a sequence derived from the Markov model (Brill, 1995). The investigation used unigram, bigram as well as trigram together with parts of speech features such as conjunctions and pronouns on the sample sets of texts provided. Then the use of three machine learning algorithms on final dataset, namely Naïve Bayes, Decision Tree and Random Forest Classifier, were used, with the explanation of how each step in the Random Forest was conducted and the results that got passed into a new document. This is due to the 96% accuracy they received from their research in the Random Forest Classifier. The strength of this research was dataset was reliable was they had selected texts randomly to minimize biasness of the data. It also reached a very high accuracy of 96% from the 3125 literary passages of a sample of 10 prominent writers. However, the approach has a limitation in that was narrowly focussed on blog writings and so proposed method has still need to address writings found in news articles, emails, tweets and other texts found in public cloud environments that are of interests to this research paper.

Rakesh Verma and Nirmala Rai proposed the “Phish-IDetector: Message-Id Based Automatic Phishing Detection” (Verma & Rai, 2015) which primarily focuses on email headers. The research focussed on observing less than 10 legitimate emails and phishing emails as the research was more drawn to the Message-ID field that is a universally unique string. The research utilises Machine Learning algorithms together with the properties of Message-IDs onto the n-gram analysis of the Message-IDs. The Random Forest Classifier algorithm performed the best in the research and so results found were utilised with the SMO algorithm (The Sequential Minimal Optimization). The researchers ensured that 100% of the experimental dataset legitimate emails had Message-IDs. This experiment relies on the Message-ID field and so if not existing the experiment will not work, fortunately it will still raise red flags to the email security. The research reached 99% True Positive Rates. It is important to note that in literature, there is no phishing detection of 100%, therefore the result of 99% was due to the smaller finite sample set of data, and is indeed noted that the exponential increasing of the sample email set together with a higher order of n-grams is difficult to run with different Machine Learning classifiers without the specialized use of big data approaches. This therefore shows a limit in the automatic detection system.

Following from this research, Rakesh Verma and Ayman El Aassal state its better to include users in the detection process or send warning to users to warn them of possible attack and thus include user training, when they were introducing “A Correlation-based Analysis and User Participation method for Detecting Phishing Email” (Verma et al., 2015). In this research, a comprehensive method to determine phishing off an email was designed, whereby information extracted off the email header is relevant to the information contained in the email body. This introduced method was executed using two algorithms, namely Header Analysis Algorthm and Matching Algorithm. Within the Matching Algorithm, multiple header fields and sender identity are authenticated using digital signatures in the Domain-Key Identified Mail (Herzberg, 2009). A sender Policy Framework is imployed to enable verification that the sending mail server is authorized in the domain that appears in the “mail from” address. Other methods like URL Analysis and Semantic Analysis are also used in the research. The research primarily focuses on the header analysis and so if the results pass the rules set for the header, then the email is passed off as legitimate. This deduction, however, is not legitimate as the attacker can still get hold of a legitimate email address by hacking into the organisations mail server and impersonate as a member of the organisation. Hence future work involves the analysis of email body.

A paper on a Content-Based Authorship Identification Framework that is used on the detection of spear phishing (Khonji, Iraqi & Jones, 2011). The paper introduces a novel framework called the Anti-Spear phishing Content-based Authorship Identification which analyses the message body of the message sender without relying on the senders ID. Its important to note that this is used specifically for Spear Phishing attacks. Bulk Phishing attacks are generic and target many users, and so many cloud providers contain several software classifiers that have the ability to detect generic nature of bulk phishing attacks. Spear Phishing attacks are targeted and thus difficult to detect to detect due to unique nature. The paper highlights a key point in its motivation into using Content based, that is User ID-based authentication is not helpful in detection as Users read text in parallel and not sequentially, thus can fall into typo-squatting and cousin-naming. Users also introduce weak authentication in systems, particularly in cloud infrastructures whereby they set very weak passwords as its easier to remember for them but also easy for a brute force algorithm to hack. In addition, users passwords can be stolen through Keyloggers, and so attackers can impersonate as the user in the system if they have the users credentials. The research aims to provide a software framework, that was otherwise not introduced in previously mentioned literature. The research also takes a whitelist approach based off of email mining techniques constructed from emails sender stylometric profiles. Thus a unique approach compared to other literatures that were a derivation of the black-list approach. 2 proposed modes to calculate similarities between the claimed and the predicted identities are namely the passive and active mode. In the passive mode, both identities are presented to the end-user. In the active mode, in a scenario where a mismatch between identities is found, the content of the message is blocked to the end-user with a warning of a mismatch occurred. The issue with this research is false positives can result if multiple users contribute to an email which the writeprints can be altered from original senders writeprint. Active mode was not explored in this paper as it increases implementation complexity and Software might not have been as accurate as the end-user. The paper followed the Security Content Automation Protocol (SCAP) methodology (Mackenzie et al., 2011) due to its high achievement of 100% of classification accuracy on datasets as it makes use of byte-level n-grams and hence useful for natural languages. For this research, a dataset of 289 emails from 12 authors were used. An Accuracy rate was used to measure performance as dataset was evaluated with the use of 10-fold cross-validation. The results found in the research show a maximum accuracy of 83% for a non-greedy n-gram ranking method and 87% maximum accuracy rate of 87% with focus to n-gram ranking methods. This framework had setbacks and limitations such as the writeprint extraction whereby if no message was previously sent by the original author then the SCAP would attempt to map it to the closest author profile match, which is not desired.

# Aim of this study

# Methodology

# Planning

# References