DETECTION AND PREVENTION OF PHISHING WEBSITE

Author : G.Vignesh,Paavai Engineering College
G.Sanjay Prasanth, Paavai Engineering College
Pamidi Praharsha, Paavai Engineering College

Guided by : Dr.M.Ramesh Kumar,
Paavai Engineering College
Edited by : G.Vignesh

ABSTRACT

Phishing is a new word produced from 'fishing', it refers to the act that the attackersallure users to visit a faked Web sites. It is a new type of network attack; the attacker copies the content from the website of a well-known company or a bank and creates a phishing website. The attacker keeps a visual similarity of the phishing website similar to the corresponding legitimate website to attract more users. The phishers must duplicate the content of the target site and they must use tools to (automatically) download the Web pagesfrom the target site.

This project helps the user to detect and identify the phishing websites. This projectuses the web crawler to crawl the hyperlinks in the website. The website does not contain any hyperlinks, and then the website will move to the black list. Those sites are fake and the user does not wish to access.

In this proposed web crawler can verify the user's inputted URL. If the result is phished, then it warns the user to block the website by adding it to a black list. The user canscan the website, if it is a phishing site, then it will give an alert message to prevent the user.

OBJECTIVE

An objective of this topic is to develop a technique or method that can be easily usedby everyone to detect whether the website is legitimate or non-legitimate accurately in real-time. It provides knowledge and insight to inexperienced web users in identifying phishing URLs.

PROJECT INTRODUCTION

Phishing techniques used so far imitate the features and characteristics of emails and it makes them look like the original ones. It appears like that of a legitimate URL source. Theuser considers this email has come from a real company or an organization. Thus, it lures theperson to visit the website by link given in that phishing email. Also, the attacker makes the user fill up the personal information by giving warning messages so that they fill up the required particular information or data which can be used by attackers to misuse the data.

They make such a situation that the user has to visit their fake website. Phishing is acyber-crime, the main reason behind the attacker doing this kind of crime is that it is not hardto perform; it is effective and does not cost anything. It has become so easy to find anyone’s email id nowadays and you can send or receive an email from anyone as it is freely availableacross the world.

These attackers put very little cost and effort to get valuable information easily. Thesecybercriminals are interested in the data which consists of crucial information of the user suchas OTP, password, CVV, credit/debit card number, medical data, sensitive data related to business, confidential data, personal information, etc. Sometimes these criminals also gather the information that can give them direct access to exploiting the weakness found in the system at the user’s end. For example, a system or an account may be technically secure and safe enough for password theft.

AIM

The aim of preparing this paper is to make it easier for organizations and companies and the major implications of these web attacks affect the financial transactions over the internet. Phishing is one of the most used popular methods that are performed to gain the advantages of security flaws in the system. Detecting, blocking, and preventing a phishing attack is an extremely important aspect to preserve the personal security and confidentiality ofan individual over the world of the internet.

The machine learning and heuristic combined to detect the phishing website which has higher output on detection. This method doesn't focused on List-based, and Deep Learning approaches. Phishing is a deception technique that utilizes a combination of social engineering and technology to gather sensitive and personal information, such as passwords and credit card details by masquerading as a trustworthy person or business in an electronic communication. Phishing makes use of spoofed emails that are made to look authentic and purported to be coming from legitimate sources like financial institutions, ecommerce sites etc., to lure users to visit fraudulent websites through links provided in the phishing email. The fraudulent websites are designed to mimic the look of a real company webpage. The phishing attacker's trick users by employing different social engineering tactics such as threatening to suspend user accounts if they do not complete the account update process, provide other information to validate their accounts or some other reasons to get the users to visit their spoofed web pages..

There are two major techniques for phishing detection: the first is the list-based (blacklistor white list) and the other is heuristic based approaches [3, 49, 50]. In the blacklist based methods, the suspicious domain is matched with some predefined stored phishing domains which are blacklisted [51–53]. The negative aspect of this scheme is that blacklist usually doesnot cover all phishing websites because a newly launched fraud website takes the substantial amount of time to get added in the blacklist record. In addition, 47% to 83% of fake URLs updated in the blacklist after twelve hours [51]. Heuristic based approaches match the heuristic design of the website with predefined rules. However, attackers can forge such features. Heuristic based approaches detect the phishing webpage by matching the features like the keywords, IP address, URL features, popup

LITERATURE SURVEY

This method of identifying phishing sites using heuristics involves examining URLs forcharacteristics such as the domain, primary domain, subdomain, and path domain. These features are analyzed and used to detect phishing sites based on the gathered information. Heuristic method of phishing detection helps to find a html errors quickly and helps to preventuser data from fraudlent websites. Heuristic method didn't use machine learning and has someaccuracy problems. Heuristic approach defines that it may produce results by themselves, or they may be used in conjunction with optimization algorithms to improve their efficiency (e.g.,they may be used to generate good seed values). With the limitation in the existing system we are introducing additional features through the heuristic approach which is simpler and effective than the earlier approaches. This is mainly used for real-world applications and one of this is used in fraud detection on an online platform. Internet environment and diversificationof available web services, web attacks have increased in quantity and advanced in quality. Heuristics approach through machine learning underlie the whole field of Artificial Intelligence and the computer simulation of thinking, as they may be used in situations where there are no known algorithm. The heuristic-based detection technique analyses and extracts phishing site features and detects phishing sites using that information. It is imperative to detect and act on such threats in a timely manner. However, blacklists cannot be exhaustive, and lack the ability to detect newly generated malicious URLs. To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years.

This Project presents a novel approach for Detecting Phishing Websites and SpamContent using Machine Learning algorithms. Specifically, we focus on using the K-NeighborsClassifier algorithm to accurately identify and classify suspicious websites and content. Ourproposed approach involves training the model on a large dataset of known Phishing Websiteand Spam Content, and then using the model to classify new websites and content based ontheir features and characteristics. We demonstrate the effectiveness of our approach through extensive experimentation and evaluation on a Real-World Dataset. Our results show that ourapproach is highly effective in detecting and classifying phishing websites and spam content,achieving high accuracy and low false positive rates. In this project, we are using Kneighbors Classifier and SVM algorithm. When we are using the Kneighbors Classifier algorithm then itsaccuracy is best. In this project, we have provided URL and Phishing website data as windows, SSL certificates, external hyperlinks, andso forth. Sometimes attacker also constructs a website in such a manner that the features are not matched with the predefined list of features. As we see some heuristic based approaches have high false positive rate [29, 54–57]. The following are the two primary advantages of visual similarity based approaches.() In order to avoid phishing detection technique, attackers usually insert images, Flash, ActiveX, and Java Applet in place of HTML text. Visual similarity based detection approaches can quickly detect such embedded objects present in phishing webpage.() Visual similarity based techniques use a signature to identify phishing webpages. The signature is created by taking common features from the whole website rather than a singlewebpage. Therefore, one signature is sufficient to detect various targeted webpages of a singlewebsite or different versions of a website..

SYSTEM ANALYSIS

EXISTING SYSTEM

Prevention and detection of phishing websites is the analysis of the detecting phishingsites, mails, messages, emails, etc. The paper detects the attacks and takes preventions from phishing sites. This existing system of Anti-phishing approaches uses feature-based MLtechniques or Blacklist Methods. These methods are Fails to recognize new phishing attacks and also get a high positive rate. The machine learning-based existing techniques extract features and data from the search engine, third party, etc. Therefore, these methods are intricate,slow, and imperfect for the real-time environment. The solution for this problem is, this paper has represented ML-based Novel Anti Phishing apps that extract features of the client-side. Thevarious attributes of malicious and non-malicious websites in-depth are examined and identified five new amazing features to differentiate the phishing websites from non-phishing ones. [4] Phishing websites are common entry points in the online human and social engineering attacks, including various ongoing web scams.

MODULE DESCRIPTION

LOGIN

The Login page will be displayed to user when user enters after registering their details.Then user can enter an Email-id and Password. To start the process, the user has to login by clicking the login button.

REGISTRATION

In this registration module, the user has to give their details such as User name, Email-id,Password, Retype input.Then detect the output whether the provided input is a Phishing website or Not. And also we detect the output whether the text data is Spam or Not..

The Web is fed with experiences about phishing, which aim at sharing phisher techniquesand behaviours. We argue that this textual information can be turned into knowledge, exploitable to prevent such attacks. Unlike anti-phishing works that aim at detecting phishing traces, this work is the first attempt to design a tool to retrieve web pages which have phishing contents. The expected crawler is dedicated to extract phishing feeds. Existing crawlers mainlyrely on building vector space models (VSM) from pages while exploiting Term frequency inverse - document frequency (Tf-idf) and cosine similarity to compute Web page similarities related to a given query. Considering the fact that vector modelling ignores the order of appearance of terms in the document as well as proximity and the connections between terms, we introduce Crawl-shing, an improved search model based on isomorphic graphs, which given two documents evaluate their similarity degree by seeking the largest common subgraph. Experimental results with phishing Web pages show that Crawl-shing presents a harvest rate better than the Breadth First Search (BFS) approach. Crawl-shing has been found more precise during exploration compared to the approaches based on vector modelling.

PROBLEM STATEMENT

The current existing systems are fully functional but when security flaws, safety, reliability of an individual are concerned; it is very disappointing to provide efficiency and afterevaluating the performance and accuracy of the systems. The prime concerns of every user are safety, security issues, confidentiality, time concerns, and personal data loss to anyone.

PROPOSED SYSTEM

The idea of the proposed system solution has to make an effective system that will give maximum accuracy. Here the decision tree classifier will be used for data fitting and a random forest algorithm will be used for classification. The planned system has the following features: a) Monitor all “HTTP” traffic of the end-user system by creating a browser extension. The benefits of extension over any software or application are that the system will be based onThe Unique in the same time at real-time and also quite agile in sending the output. b) Comparing each URL domain with the white-list of trusted domains and also the black-list of illegitimate domains. c) If the domain of the URL is found under the white list, mark the URLas innocent (Exact Matching), else go further and use the other approaches. d) The whole website analysis would now be done by considering various details. The set of features selectedare: Website protocol (secure or insecure), Password, Date of birth and Gender through the user can login. If the user is anexisting user then they shall continue the process bygiving their authentication details. If the user is not an existing user then they have to register by giving their details. MD5 algorithm is used to secure the user's profileby encrypting the password.

WEB CRAWLER

Crawlers look at Web Pages and follow links on those pages. In this modulewe can give some website URL. If it is not a phishing site, it will crawl the links of that website; otherwise thatlink will add to the blacklist.

BLACKLIST

The black list contains the phishing sites**.** Which works on list based detection, known phishing sites which are in black list. And then these phishing sitesare added to black list.

CHECK WEBSITES

In this module user can enter the URL of website, to check if it is in the blacklist or not.If it is in the black list, those sites are fake and the user does not wish to access and then it will give the alert message to the user.

FEEDBACK FORM

User can give their feedback of the application. User can fill their detailslike Name, Email-id and then give the feedback.

BACK END-MY SQL

A database is a separate application that stores a collection of data. Each database has oneor more distinct APIs for creating, accessing, managing, searchingand replicating the data it holds.

MySQL is an open-source relational database management system (RDBMS).MySQL is a central component of the LAMP open-source web application softwarestack. . MySQL is also usedin many high-profile, large-scale websites, including Google, Facebook, Twitter, Flickr, and YouTube. MySQL is a very powerful program in its own right. It handles a large subset of the functionality of the most expensive and powerful database packages. MySQL Server can run comfortably ona desktop or laptop, alongside user other applications, web servers, and so on, requiring little or no attention. If user dedicates an entire machine to MySQL, user can adjust the settings to take length of the URL, number of hyphens (-) in URL,number of @ symbol in URL, number of dots in the URL, using the direct IP address or not, daily page view, registration, and expiration date of website, daily unique visitor, favicon iconsimilarity and Google indexing.

SOFTWARE DESCRIPTION

PHP

Hypertext Preprocessor. PHP is a widely-used, open source scripting language and its scripts are executed on the server.. In its early development by a guy named Erasmus Leadoff, it was called Personal Home Page tools. When it developed into a full-blown language, the name changed to be more in line with its expanded functionality.

The PHP language's syntax is similar to the syntax of C, so if you have experience with C, you'll be comfortable with PHP.PHP is actually simpler then C because it doesn't use some ofthe more difficult concepts of C.PHP also doesn't include the low-level Programming capabilities of C because PHP is designed to program Web sites and doesn't require those capabilities. PHP isparticularly strongin its ability to interact with database.

PHP handles connecting to the database and communicating with it. You don't need to know the technical details for connecting to a database or for exchanging message with it .you tellPHP the name of the database and where it is, and PHP handles the details .it connects to the database, passes your instructions to the database, and returns the database response to you. Technical support is availablefor PHP.

XAMPP

XAMPP is a free and open source cross-platform web server solution stack package developed by Apache Friends, consisting mainly of the Apache HTTP Server, Maria DB database,and interpreters for scripts written in the PHP and Perl programming languages.XAMPP stands forCross-Platform (X), Apache (A), Maria DB (M), PHPand Perl (P). It is a simple, light weight Apache distribution that makes it extremely easy for developers to create a local web server for

testing and deploymentpurposes. Everything needed to set up a web server – server application(Apache), database (Maria DB), and scripting language.

advantage of all the memory, CPU power, and I/O capacity available.

MySQL works very quickly and works well even with large data sets. MySQLis very friendly to PHP, the most appreciated language for web development. MySQL supports large databases, up to 50 million rows or more in a table. The default file size limit for a table is 4GB, but you can increase this (if your operatingsystem can handle it) to a theoretical limit of 8 millionterabytes (TB).

The MySQL server software itself and the client libraries use dual-licensing distribution. Support can be obtained from the official manual. Free support additionally is available in differentIRC channels and forums. Oracle offers paid support via its MySQL Enterprise products. They differ in the scope of services andin price. Additionally, a number of third party organizations existto provide supportand services, including MariaDB and Percona.

INPUT DESIGN

Input design is one of the most expensive phases of the operations of any computerized system and is often the major problem of a system. A larger number of problemswith a system can usually be traced back to fault input design and method. Needless to say, therefore that the input data is the life block of a system and has to be analyzed and designedwith the most consideration.

The objectives considered during input design are:

- Nature of input processing.
- Flexibility and thoroughness of validationrules.
- Handling of properties within the inputdocuments.

System analysts decide the following input design details like, what data item to input,what medium to use, how the data should be arranged or coded data items and transactions needing validations to detect errors and at last the dialogue to guide users in providing input.

Input data of a system may not be necessarily raw data captured in the system from scratch. These can also be the output of another system or subsystem. The design of input covers all phases of input from the initial data to actual data entering to the system for processing. The design of inputs involves identifying the data needed, specifying the characteristics of each data

CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

In this paper, a Machine learning algorithm is applied progressively and successivelyto predict the phishing of an URL of a website. Thus, the proposed system enables internet users to have safe browsing and safe transactions. It helps users to save their important privatedetails that should not be leaked. User only needs to provide an active internet connection to check the legitimacy of an URL. Providing our proposed system to users in the form of an extension makes the process of delivering our system much easier. A challenge in this domain is that criminals are constantly making new strategies to counter our defense measures.

To succeed and achieve efficient results, there is a need for algorithms that continually adapt to new threats, examples, new phishing techniques and features of phishing URLs. And thus, online learning algorithms are used. This paper provides a system that gives a result with maximum accuracy. Using random forest with a decision tree classifier enhances the evaluationmetrics of the system and provides an efficient protection system for the internet user. In the future system is designed to work efficiently and successfully to detect false-negative and false-positive results, more accuracy in the detection and new problems in the features of an URL.

FUTUREENHANCEMENT

As a part of the future scope, deep learning models like Recurrent Neural Network(RNN) or Generative Adversarial Network (GAN) can be used. Adding to that, we would liketo notifythe original users related to their phishing websites; for example, if a phishing website similar to www.microsoft.com is created, say www.microsooft.com, so, we would like to inform the original creator of the website regarding the phishing website that has been createdunder their name. Along with that, the type of websites which showcase a fake product and take away the money without receiving any products in exchange can be detected using webscraping and sentimental analysis.

OUTPUT DESIGN

Output design generally refers to the results and information that are generated by thesystem for many end-users; the score report is the main output for Detecting phishing websitesand the basis on which they evaluate the usefulness of the application.

The output is designed in such a way that it is attractive, convenient and informative.

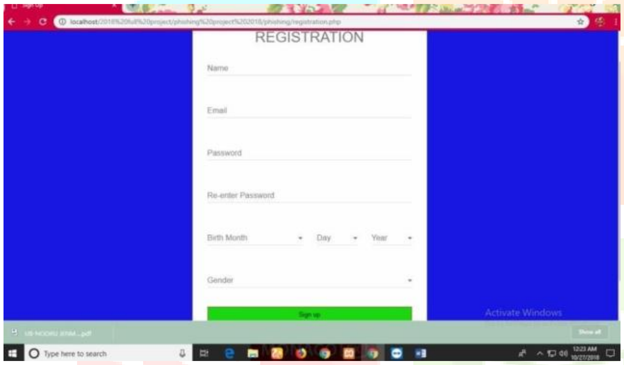
Pages are designed in PHP with various features, making the console output more pleasing.

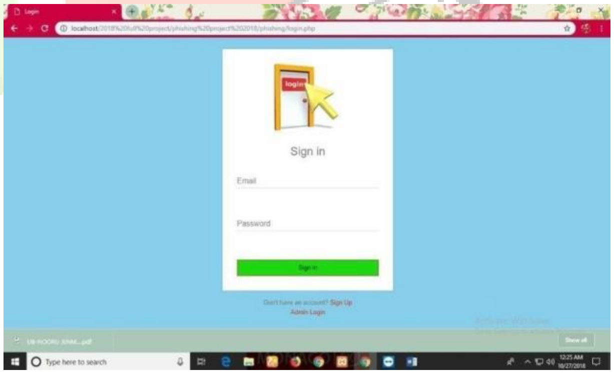
Output design is a very important concept in the computerized system, without reliableoutput the user may feel the entire system is unnecessary and avoids using it. The proper output design is important in any system and facilitates effective decision-making.

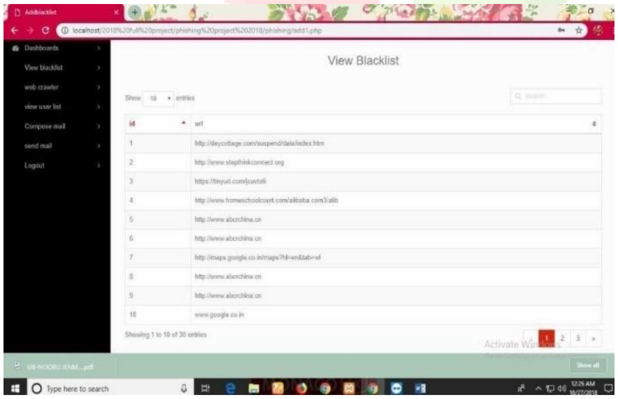
OUTPUT DESIGN

Hence it is necessary to design output so that the objectives of the system are met in the best possible manner. The input to the system needs to be accurate for the output of the system to be accurate. And only if the output is accurate the system will be recognized.

As the outputs are the most important sources of information to the users, better design should improve the system's relationships with us and also will help in decision-making. In this system, output is in the detecting and preventing websites using web crawler. When the user enters some website URL and clicks the scan button, it will scan and produce the result whetherthe site is phishing or not. This process produces the result based on a blacklist. It can preventthe user from giving the alert message.




Login
localhost/2018%20full%20project/phishing%20project%202018/phishing/login.php
login
Sign in
Email
Password
Sign in
Don't have an account? Sign Up
Admin Login
Type here to search
12:25 AM
10/27/2018

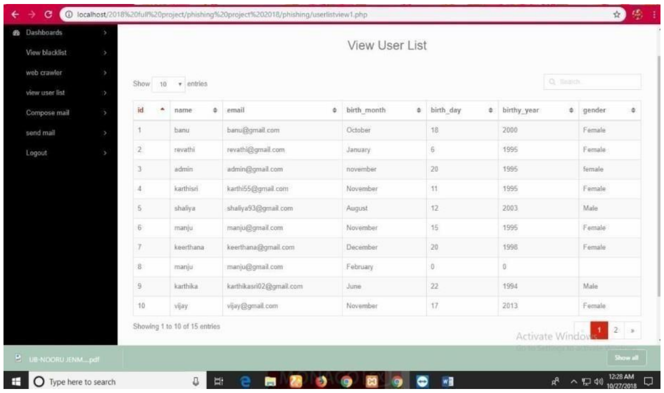


View Blacklist

Show 10 entries

id	url
1	http://deycottage.com/suspend/data/index.htm
2	http://www.stopthinkconnect.org
3	https://tinyurl.com/jcuvto6
4	http://www.homeschoolcourt.com/alibaba.com3/alib
5	http://www.abcrchina.cn
6	http://www.abcrchina.cn
7	http://maps.google.co.in/maps?hl=en&tab=wl
8	http://www.abcrchina.cn
9	http://www.abcrchina.cn
10	www.google.co.in

Showing 1 to 10 of 30 entries

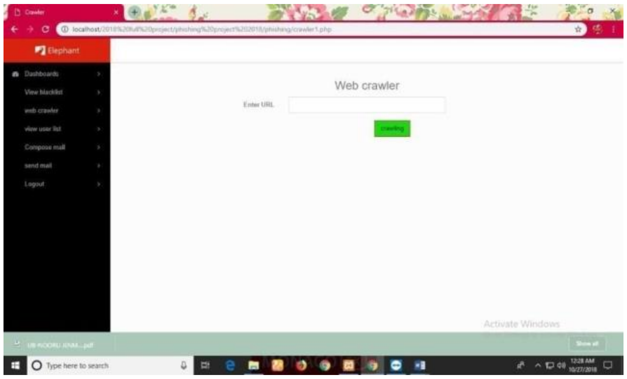


View User List

Show 10 entries

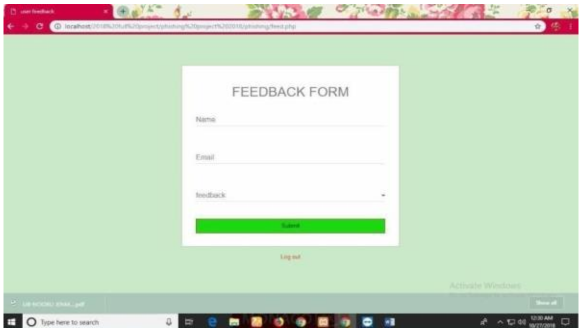
id	name	email	birth_month	birth_day	birthy_year	gender
1	banu	banu@gmail.com	October	18	2000	Female
2	revathi	revathi@gmail.com	January	6	1995	Female
3	admin	admin@gmail.com	november	20	1995	female
4	karthisri	karthi55@gmail.com	November	11	1995	Female
5	shaliya	shaliya93@gmail.com	August	12	2003	Male
6	manju	manju@gmail.com	November	15	1995	Female
7	keerthana	keerthana@gmail.com	December	20	1998	Female
8	manju	manju@gmail.com	February	0	0	
9	karthika	karthikasri02@gmail.com	June	22	1994	Male
10	vijay	vijay@gmail.com	November	17	2013	Female

Showing 1 to 10 of 15 entries



Web crawler

Enter URL

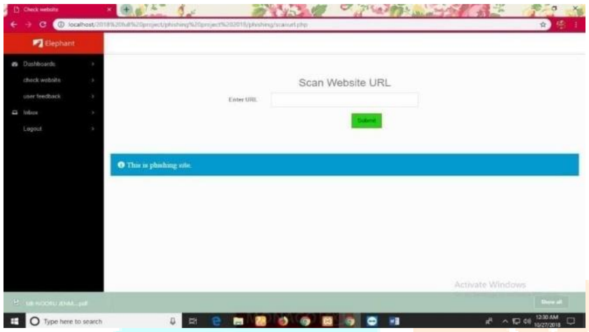


FEEDBACK FORM

Name

Email

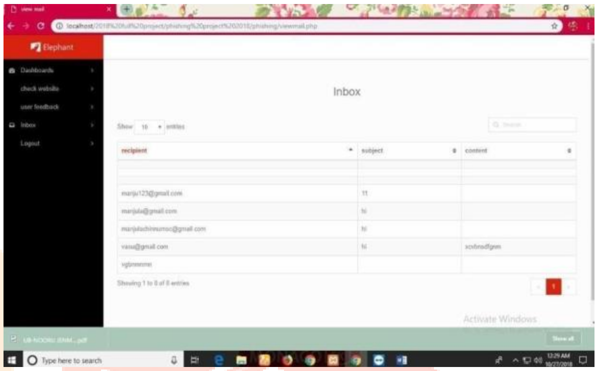
feedback



Scan Website URL

Enter URL

This is phishing site.



Inbox

Show 10 entries

recipient	subject	content
manju123@gmail.com	11	
manjula@gmail.com	hi	
manjulachinnumsc@gmail.com	hi	
vasu@gmail.com	hi	xcvbnsdfgm
vgbnmmnl		

Showing 1 to 8 of 8 entries