Analysis of dynamic content of social networking sites for the purpose of monitoring and to bring out the suspicious cyber security issues

Ritu Dahiya, Jagdish Pandya

Abstract: The cyber world is susceptible to sniffing, snooping, eavesdropping, botnets, Distributed Denial of Service attack, SQL injection and many other types of security threats. With the private data of users travelling across the network and various freely available and easy to use data mining and sniffing tools, there is growing concern for cyber security. With the growing users of social networking sites and enormous data flowing through the network, monitoring of the content across network and correct analysis of the OSN sites content flow can detect any suspicious activities from security threat point of view in the cyber world. This helps us to better understand what is happening on a network. A packet sniffer, a tool used to capture raw network data going across the wire, does this. [5]

Keywords: Social Networking (SN), Cyber Security, Content Monitoring, Sniffing, Intelligence Gathering

Introduction

In this paper capture and analyze the dynamic content flow across SN sites by using appropriate tools and software available in market. By tracing the pattern of conversations across the SN sites, a graph model can be done with nodes representing the communities within the network. Appropriate correlation of theoretical principals of network functioning along with the pattern of graphs obtained by using the analysis tools can guide us to beneficial monitoring of the network. We study suspicious flow of protocols in the network to detect and avoid any security threat. Content analysis includes packet sniffing or protocol analysis. It includes the process of capturing and interpreting live data as it flows across a network. Study of „Network Basics" and concepts in „Advanced Networking", which includes understanding of various protocols and the content flow and routing of packets across the networks. [6] Study broadly the algorithms and graphs in „Data Structures" to understand the theory applied while developing the various content search tools within the network. The applications of web services have wide usage in e commerce, Internet banking, online transactions, e purchase etc. With the free flowing packets across the network, including private data, there is concern for security. Hence detailed study of „Cyber Security" concepts. Study of various research papers in the field of social networking to understand the work done in this field and to clearly define work plan of content analysis for social network monitoring for security concern. Use the available software/ tools for the practical study of the above concepts for content capture and analysis across social networks.

Procedure of Content Analysis in SN

Procedure to content analysis following steps can be performed

(a) Study the dynamic content flow in social networking sites, by means of packet capture,

(b) Study of algorithms and data structure graphs to broadly understand data mining and monitoring of content on social networking sites.

(c) Study the security concepts in detail and highlight the security loopholes in the usage of web services and applications.

For practical work, study and select tools/ software for packet capture, for content monitoring and analysis. Also study of the scanning and sniffing tools, which can be potentially used by miscreants and can cause cyber security concern. Theoretical study is done by referring to available books on the subject and the research papers on social networking. The concepts in „Basic Networking", „Advanced Networking", „Data Structures" and „Cyber Security" are studied for the subject understanding.Throughout during the study, effort is made to correlate the theory with the practical conducted by using the available software/ tools, so that the check points and limitations in this process of analysis are brought out which can help us to highlight a potential security threat.[7]

Social Network Monitoring and Analysis

Social Network Monitoring refers to tracking the conversation of people with the purpose of learning, engaging, helping and collaborating. Building a social media-monitoring dashboard can do it. Monitoring is performed on keywords basis. Relevant keywords include the brand name, product name etc. Based on keywords, monitoring system of choice goes out to the social networks one specifies, grabs the relevant articles and messages and arranges them for easier assimilation and action.[8]

Social media monitoring is used by business community, academia and nonprofits. Social network analysis has emerged as a key technique in modern sociology. It has gained a significant following in anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology and sociolinguistics and has become a popular topic of speculation and study. Social networks also play a key role in hiring, in business success and in job performance. Networks provide ways for companies to gather information, to deter competition and collude in setting prices or policies.

The input from these numerous users of SN can provide valuable insight in business applications and decision-making. Timely knowledge of content flow across web and intervention by appropriate concerned authorities can prevent build up of incorrect and biased views on a subject or brand. The concepts and algorithms in data structures are used to design the various search methodologies that can be applied across the networks for the purpose of content search. The content analysis of this captured data can yield the benefits of monitoring.[1]

Social Network Analysis can be defined as a set of techniques underpinned by statistical analysis that make visible the hidden connections that are important for sharing information, decision-making and innovation in an organization. It maps and measures formal and informal relationships to understand what facilitates or impedes the knowledge flows that bind interacting units, viz., who knows whom and who shares what information and knowledge with whom by what communication media. SNA provides both a visual and a mathematical analysis of human relationships and activities.

To monitor and analyze the content of social networks from cyber security perspective, it is important to understand the terminologies in the subject. Subsequent paragraphs, lists them with their corresponding brief introduction.

Cyber Security in Social Networks

Security is a major issue with the Internet because it is public domain. Since packets pass through machines over which one has no control, someone can potentially see confidential information. Any hacker with a network data scope can get credit card numbers, social security numbers and other confidential information from the transmissions.[2] We need to design for these potential security leaks. For cyber security, we look from the bottom of the protocol stack (the physical wire) all the way up to end-user applications. Any level can be attacked therefore every level needs some security mechanisms in place. The following are the list of security risks:-

(a) Trojans - It is a program in which the malicious or harmful code is contained inside apparently harmless programming or data in a way that it can get control and cause damage. Attacker gets access to the stored passwords in the attacked computer and is able to read personal documents, delete files and display pictures.

(b) Keyloggers - It is program that runs in the background and allows remote attackers to record every keystroke.

(c) Spyware - it is a type of malware that allows attackers to secretly gather information about a person or organization.[10]

(d) Fast Flux Botnets - These can be used to launch various types of Denial of Service (DoS) and other web-based attacks which may lead to business downtime and significant loss of revenues. Botnet infect a large number of computers across a large geographical area to create a network of bots that is controlled through a control center.

(e) Phishing - An illegitimate email falsely claiming to be from a legitimate site attempts to acquire the user‟s personal or account information.

(f) Social Engineering - Convincing people to reveal the confidential information.[9]

(g) Viruses - A virus is a self-replicating program that produces its own code by attaching copies of self into other executable codes.

(h) Worms - These are malicious programs that replicate, execute and spread across the network connections independently without human interaction.

(i) Dumpster Diving - Searching for sensitive information at the user‟s trash bins, printer trash bins and user desk for sticky notes.

(j) Cyber Espionage - Gathering information through cyber technology.

(k) Transportable data (USB, laptops, backup tapes).

(l) Zombie Networks - Networks with DDoS attacks.

Experimentation and Results

The content flow of the University Campus is captured by installing Wireshark on one of the machines in the lab. General layout of the university LAN is sketched as below. One of the computers (with Wireshark installed on it) in the lab is configured as a gateway and the entire internet traffic of one lab is routed though this gateway. The packets are captured, filtered and analyzed at this machine.

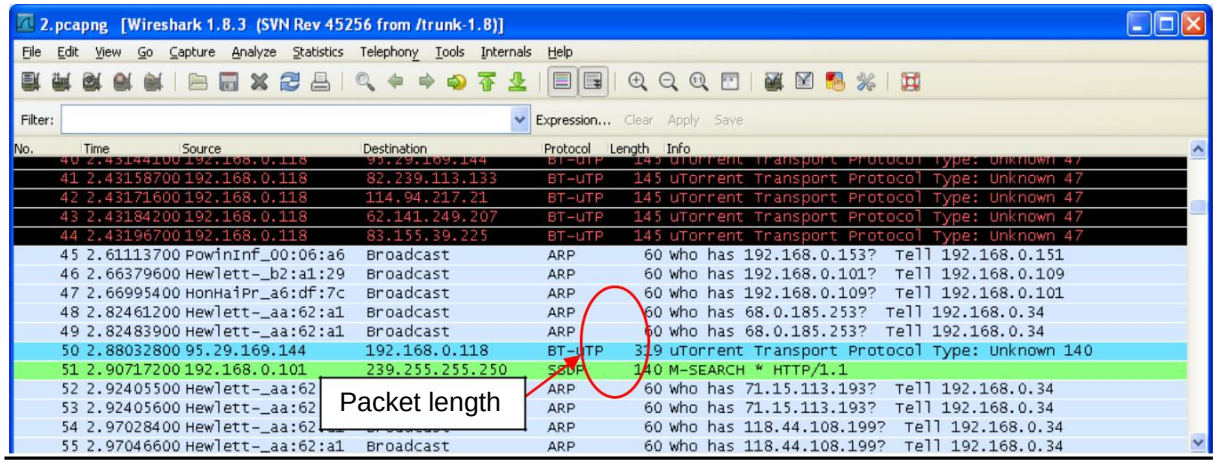

Figure 1. Content Analysis by Comparison of Packet Length

Network activities can be determine by viewing the length of the packets that are flowing across the network. Type of protocol flow can be determined from the packet length. Their correlation with the IP addresses of the network users can further assist to detect any suspicious activities. IP, TCP, UDP, ICMP are the lower layer protocols. DHCP, DNS, HTTP are most common upper layer protocols.[3,4] The graph below shows the percentage distribution of the protocols in the network traffic that is captured using Wireshark. Samples of the protocol distribution results in a given network at different times can help the analyst to study normal network activities in a given environment. Any deviations from the same may be viewed with suspicion.

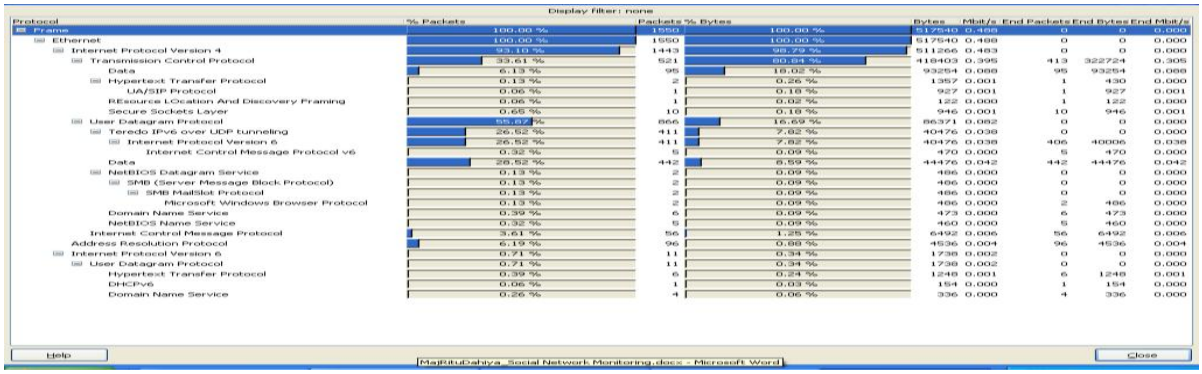

Figure 2. Analysis of Network Traffic by Study of Protocol Flow

Filters are created for traffic and the resulting graph indicates the throughput trends between these protocols. The protocol strength versus time is captured below. Comparative flow of various protocols at any given time can be seen.

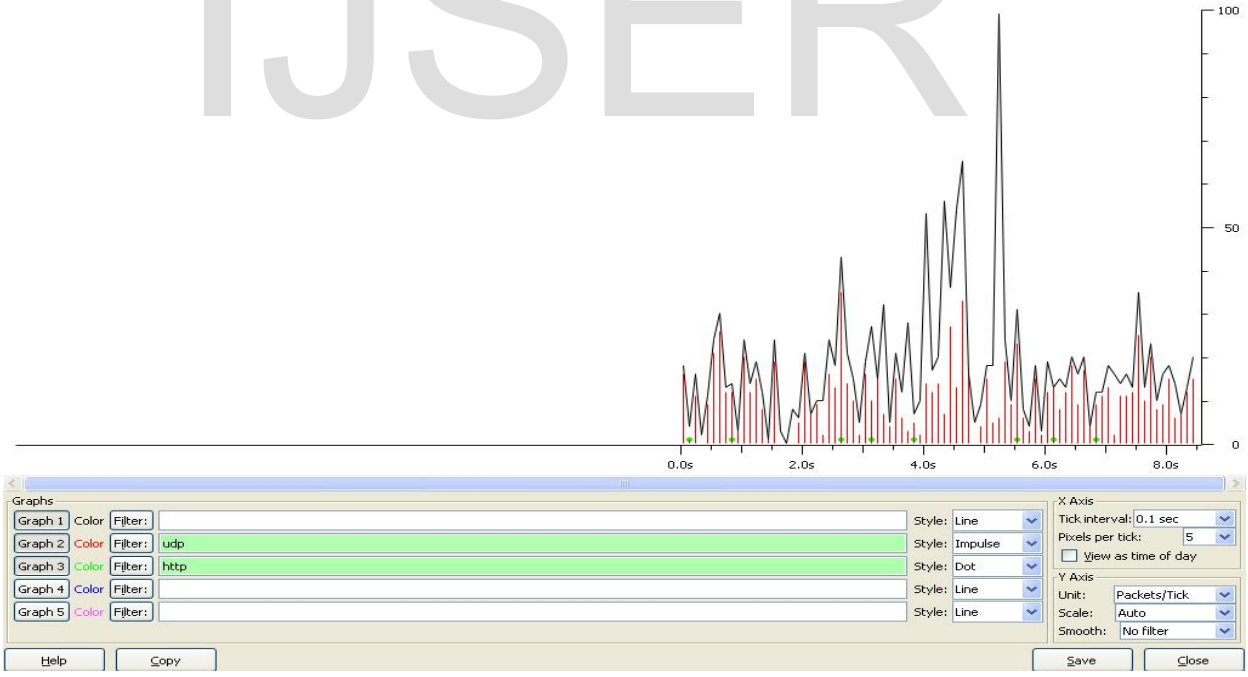

Figure 3. Analysis of Network Traffic by Creating Filters

Round Trip Times (RTT) for a captured file can be viewd.[11] RTT is the time taken for a packet to receive an acknowledgement. It can help us to determine slow points or any latency or bottlenecks in communication. It indicates the download rate and hence can assist in determining the performance of a particular social networking site.

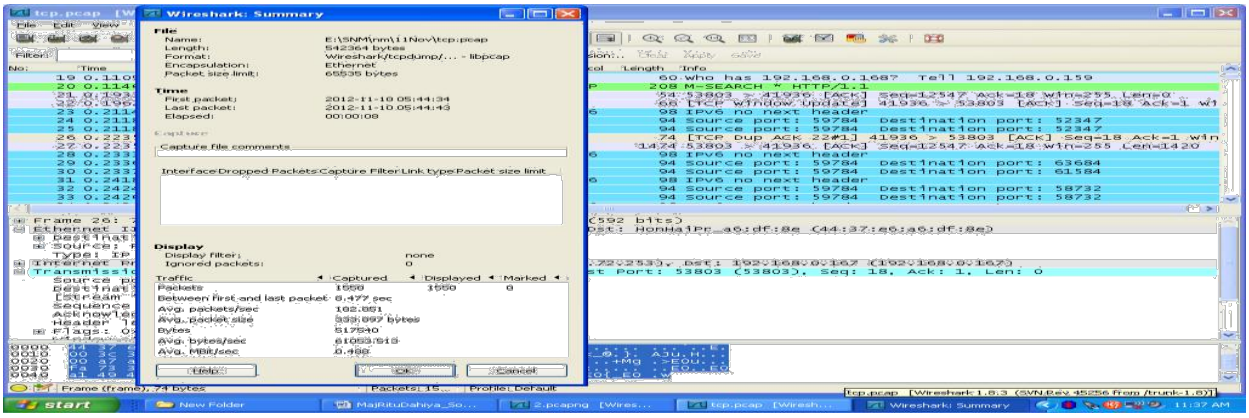

Figure 4. Summary of Captured Data File

Flow graph displays the flow of data over time. It assists in visualizing connections. This is particularly advantageous to troubleshoot or do in-depth analysis of any suspicious activity. By applying appropriate filters on IP addresses and protocol types, certain type of cyber security issues like DDoS and man-in-the middle attack can be verified.

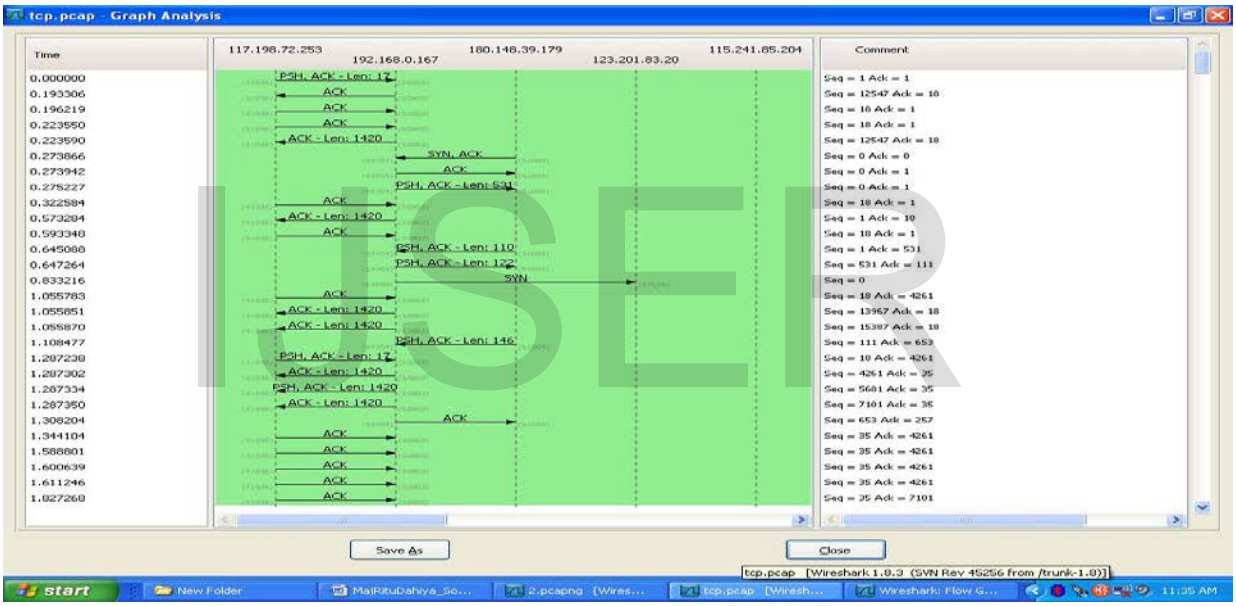

Figure 5. Tracing Communication between Users

Dissector of protocols can be useful to create alerts. In the various types of cyber attacks the hacker carries out scanning to detect any open ports. This is done by pinging sequence of IP addresses, port scanning and then abruptly closing an established connection. The analysis tools have codes to detect these basic types of unusual events during network usage. [12]

Ideally every connection would end gracefully with TCP teardown. In reality the connections may end abruptly. If a potential attacker is performing a port scan, TCP packets with RST flag is used. If a machine is attempting to communicate with another on port 80 which has no web interface configured i.e. no service is listening on that port, then in response RST flag is used to indicate that no communication is possible. HTTP represents the highest percentage in the social networking. The procedural flow of the HTTP packets for the communication over social networking sites is listed below:-

(a) Communication begins with a 3-way handshake between the client and the server.

(b) Once communication is established, the first packet is marked as a Hyper Text Transfer Protocol (HTTP) packet from the client to the server.

(c) The HTTP packet is delivered over Transfer Control Protocol to the server's port 80.

(d) HTTP packets are recognized by one of the eight different request methods, which point out the action that the packet transmission will be performed on the receiver.

(e) In next step the host transmits information about itself to the web server. This information includes the user agent (browser) being used, languages accepted by the browser and cookie information.

(f) The server uses this information about the host to determine the type of data to return to the client so as to ensure compatibility. HTTP is used only to issue application layer commands between client and server.

(g) Data is sent from the server in packets.

(h) An acknowledgement is sent from client.

(i) The data is transferred as TCP segments, rather than HTTP packets.

Once the data is transferred, a reassembled stream of the data is sent. HTTP uses a number of predefined response codes to indicate the results of a request method. The packet also includes a time stamp and some additional information about encoding of the content and configuration parameters of the web server. On receipt of this packet, the transaction is deemed to be complete.
To upload data, user uses POST method. After initial three way handshake, client sends an HTTP packet to the web server. Once data is transmitted in POST, an ACK packet is sent.
The server responds with packet consisting of response code 302 meaning "found". 302 response code is a common means of redirection in HTTP. The location field in this packet specifies where the client is to be directed. That is the place on the originating web page where the comment was posted. Finally, the server transmits status code 200 meaning successful request method.
The page‟s content is sent over the next several packets to complete the transmission process.
We can analyze the traffic of the social networking websites or web services by capture and study of protocol flow on network. Subsequently we describe and compare the protocol flow on two of the social networking sites. To maintain confidentiality the names of these social networking sites are referred to as SNS-1 and SNS-2. We login in to the websites and capture the traffic from the login process.

SNS-1

For login to SNS-1, initial three packets constitute TCP handshake between local device and remote server. The remote server listens on port 443 which is associated with SSL over HTTP (i.e. HTTPS). Hence we assume that it is SSL traffic.
The packets that follow the handshake are part of SSL encrypted handshake. SSL relies on keys. Keys are strings of characters used to encrypt and decrypt communication. During handshake, formal transmission of keys between hosts and transmission of encryption characteristics takes place.
The encrypted packets that transfer the data are identified as „Application Data‟ in the „Info‟ column of „Packet Details‟ pane of Wireshark. Expanding it displays encrypted application data in unreadable form. [13]
Initially there is transfer of the user name and password during login. The authentication continues until the connection begins its teardown process with FIN/ACK.
After authentication, the host browser is redirected to SNS-1 home page. Fresh handshake process sets up new connection to the same remote server on port 80 instead of 443. After establishing connection, HTTP GET request is sent by host for the root directory of web server. The server acknowledges the request and begins transmitting data over next several packets.

If we submit a chat message and capture the packets, transmission is seen to begin with a handshake between the host machine and remote machine. Expanding the HTTP header in „Packet Details" pane, POST is used with URL /status/update. Host field contains SNS-1.com, hence we conclude that this is a packet from the chat message. The data is contained in the packet"s Line based Text Data field. Inside it, a field named Authenticity Token followed by status field in a URL contains the chat message text data. The value of the status field is the chat message in an unencrypted plaintext.

SNS-2

Initially during login process, there occurs TCP handshake over port 443. After handshake, SSL handshake occurs and login credentials are submitted.

Here we note the difference between SNS-1 handshake and that of SNS-2. There is no authentication connection teardown following the transmission of login credentials. Instead there is GET request for /home.php in HTTP header. The connection used for authentication is torn down after the content of home.php is delivered. First, the HTTP connection over port 80 is torn down and then HTTPS connection over port 443 is torn down.

While private message is sent from one account to another SNS-2 account. HTTP traffic responsible for transmission of message, uses POST method with reference to URL string which includes reference to AJAX. AJAX is a client side approach used for creating interactive web applications that retrieve information from server. Unlike SNS-1, after the message is sent to the client"s browser, the session is not redirected to another page. The message is sent by some interactive pop-up. There is no redirection or reloading of content. [14]

The way in which the authentication is done in a social networking site, can prevent certain security threats or make them difficult. Also the process determines why certain services operate slowly.

The UDP based DNS traffic is in form of queries and responses. A request to view a web page can break down to touch various servers. The social networking can be affected by various networking issues.

The snapshots below show the viewing of yahoo pipes. The hops are viewed by traceroute. Same is seen by packet capture using Wireshark. The connection initiation, route taken, TCP handshake, transfer of data, conversation between host and server and finally termination of connection are studied using Wireshark and other packet analyzing tools like PRGT and Colasoft Capsa.

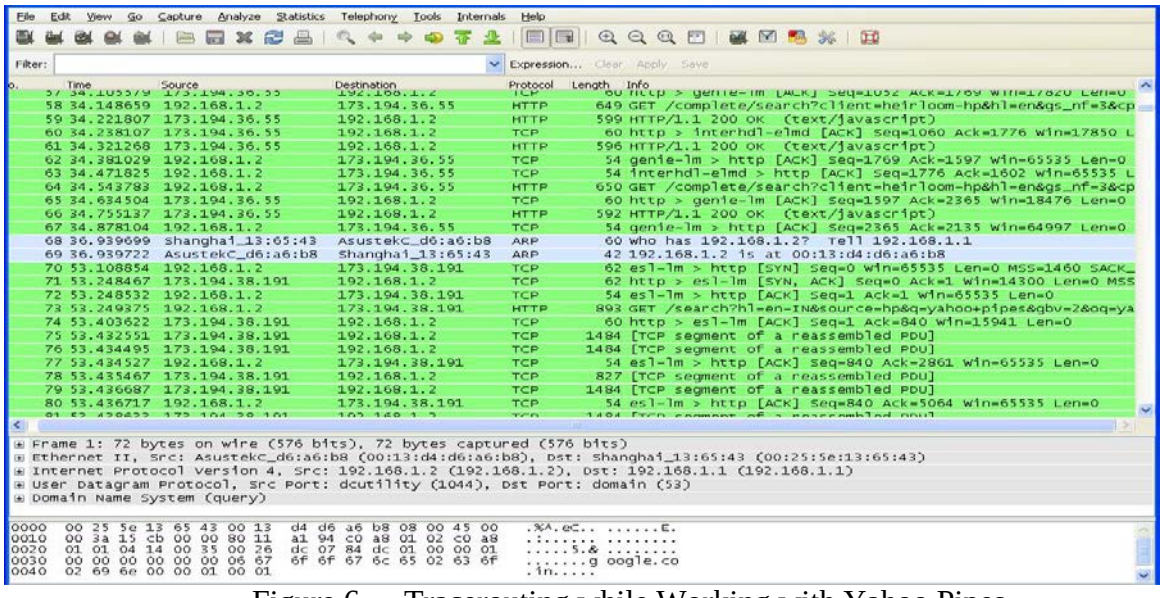

Figure 6. Tracerouting while Working with Yahoo Pipes

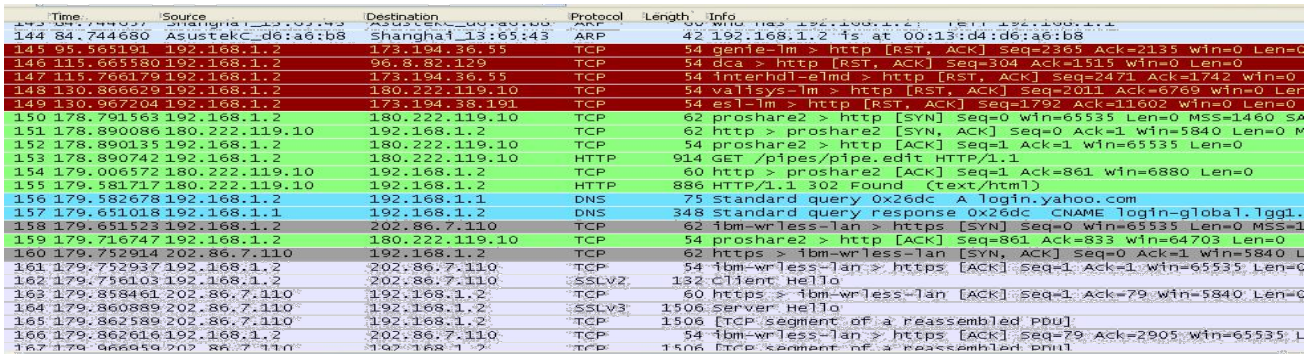

Figure 7. Large Packets show Data Transfer during Conversations

Figure 8. Transfer of Pictures while Fetching of Yahoo Pipes

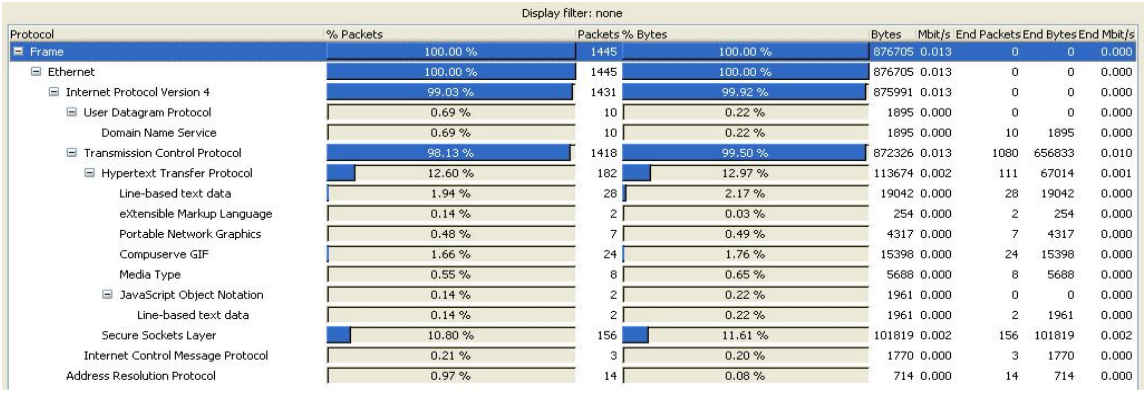

Figure 9. Percentage Distribution of Various Protocols while Working with Yahoo Pipes

Conclusion

There is enormous data flowing across the web. It demands tremendous knowledge on the subject, colossal efforts and practice to selectively extract the relevant and desired information from these free flowing packets of information. Study and analysis of protocol flow on network, help us to monitor and spot suspicious activities on the web. To conclude anomalies by analysis of packet traffic, it is firstly required to understand the normal flow on web. Examining packet lengths is a way to get bird‟s eye view of capture. Small packets consist of protocol control commands. Large packets indicate data transfer.

Future Work

Today‟s Internet is a technological mix of various devices and transmission technologies. With intelligent routing in place, it is difficult to capture entire flow of packets to look at the Internet security from a broader perspective. There is also limitation due to scalability of networks and finite computational ability of the machines. Various machines work on different platforms, further limiting the analysis capability of tools. With manual means of content analysis, it is difficult to address all aspects relating to Internet security. Detailed study of each aspect and its corresponding automation would help us to practically monitor Internet traffic and address cyber security concern.

References

1. Rudin, April. “My Personal Social Story: Why Social Media Platforms Are Fads.” The Huffington Post. TheHuffingtonPost.com, 27 Aug.

2. Grifoni Patrizia, Ferri Fernando, D‟Andrea Alessia (2013). An integrated framework for on-line viral marketing campaign planning in

3. “Social Media facts, figures and statistics 2013.” Digital Insights. N.p., n.d. Web. 3 Apr. 2014. http://blog.digitalinsights.in/social-mediafacts-and-statistics-2013/0560387.html

4. 4 Gelles, D. (2011, July 29). Inside Match.com. Financial Times. Retrieved April 11, 2014, from http://www.ft.com/intl/cms/s/2/f31cae04-b8ca-11e0-8206-00144feabdc0.html#axzz2yaf8JbDU

5. Tumblr. (n.d.). About. Retrieved April 11, 2014, from http://www.tumblr.com/about

6. Shazam - About Shazam - Advertisers. (n.d.). Shazam. Retrieved April 11, 2014, from https://www.shazam.com/music/web/advertisers.html

7. Etherington, D. (2014, February 10). Flickr At 10: 1M Photos Shared Per Day, 170% Increase Since Making 1TB Free.TechCrunch. Retrieved April 11, 2014, from http://techcrunch.com/2014/02/10/flickr-at-10-1m-photos-shared-per-day-170-increase-since-making-1tb-free/

8. Lunden, I. (2013, September 11). Disrupt SF 2013. TechCrunch. Retrieved April 11, 2014, from http://techcrunch.com/2013/09/11/githubhits-the-4m-user-mark-as-it-looks-beyond-developers-for-its-next-stage-of-growth/

9. Fidelman, M. (2013, May 19). Study: 78% Of Salespeople Using Social Media Outsell Their Peers. Forbes. Retrieved April 11, 2014, from

http://www.forbes.com/sites/markfidelman/2013/05/19/study-78-of-salespeople-using-social-media-outsell-their-peers/

10. "SEC Embraces Social Media." The Wall Street Journal. Dow Jones & Company, n.d. Web. 11 Apr. 2014. http://online.wsj.com/news/articles/SB10001424127887323611604578398862292997352.

11. Kelly, H. (2013, August 2). 83 million Facebook accounts are fakes and dupes. CNN. Retrieved May 4, 2014, from http://www.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/

12. Pick, T. (2013, August 6). 101 Vital Social Media and Digital Marketing Statistics. RSS. Retrieved April 25, 2014, from http://socialmediatoday.com/tompick/1647801/101-vital-social-media-and-digital-marketing-statistics-rest-2013.

13. Research, F. (2013, July 17). In Business, Everybody Uses Social Media For Work; The Question Is How. Forbes. Retrieved April 8, 2014, from http://www.forbes.com/sites/forrester/2013/07/17/in-business-everybody-uses-social-media-for-work-the-question-is-how/

14. Social Not-Working: Block, Allow, or Manage? Clearswift Adaptive Cyber Protection. Retrieved April 8, 2014, from http://cyberhub.clearswift.com/system/files/Social-notworking-block-allow-or-manage_0.pdf