Examining Login URLS to Identify Phishing Threats

M. Sravan Kumar Babu[1], A. Chandana[2], A. Anusha[2], K. Harika[2], P. Jhansi[2]

[1]Assistant Professor,[2] UG Scholar, [1,2] Department of CSE-Cyber Security

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

Abstract

Phishing refers to a type of cyberattack known as social engineering, in which criminals trick users into revealing their credentials by utilizing a deceptive login form that submits the information to a malicious server. In this project, we compare machine learning techniques to propose a method for effectively detecting phishing websites through URL analysis. Most current state-of-the-art solutions for phishing detection consider homepages without login forms as the legitimate class. However, we differ in our approach by incorporating URLs from the login pages into both classes. We believe this approach better reflects real-world scenarios and demonstrate that existing techniques yield a high false-positive rate when tested with URLs from legitimate login pages. Furthermore, we employ datasets from different yearsto illustrate how models experience a decline in accuracy over time. We train a base model using outdated datasets and evaluate its performance using recent URLs. Additionally, we conduct a frequency analysis of current phishing domains to identify the various techniques employed by phishers in their campaigns. To support our claims, we introduce a new dataset called Phishing Index Login URL (PILU-90K), which consists of 60,000 legitimate URLs encompassing index and login websites, along with 30,000 phishing URLs. Lastly, we present a Logistic Regression model that, when combined with Term Frequency - Inverse Document Frequency (TFIDF) feature extraction, achieves an accuracy of 96.50% on the provided login URL dataset.

Keywords: URLs, Phishing attacks, Phishing index login URL.

1. Introduction

In the last years, web services usage has grown drastically due to the current digital transformation. Companies motivate the change by providing their services online, likee-banking, ecommerce, and Software as a Service [1].

Identifying phishing sites through their HTTP protocol is no longer a valid rule. In the 3rd quarter of 2017 [7], the APWG reported that less than 25% of phishing websites were hosted under HTTPS protocol, whilst this amount has increased to 83% in the 1st quarter of 2021 [8]. These websites provide secure end-to-end communication, which transmits a false safe impression to the user while making an online transaction [9]. Furthermore, the Anti-Phishing Working Group (APWG) [10] has reported a significant increase in phishing attacks, i.e. from 165; 772 to 611; 877 websites, just between the first quarter of 2020 and 2021 respectively. A reason behind this increase might be that people have resorted (and still are) to online services during the COVID- 19 pandemic. One of the most popular solutions for phishing detection is the list-based approach, which analyses the requested URL against a phishing database [11]. Some examples of this solution are Google SafeBrowsing,1 PhishTank,2 OpenPhish3, or SmartScreen.4 If a requested URL matches any record, the request is blocked, and a warning is displayed to the user before visiting the website. However, despite the capabilities of the list-based approach, it would fail if the phishing URL was not reported previously [12][14], and it will require a continuous effort to update the database with newer phishing data. Bell and Komisarczuk [11] observed that many phishing URLs were removed after day five from Phish tank while Open Phish removed all URLs after seven days from its report. This issue allows attackers to reuse the same URL when it is removed from different list.

Due to the mentioned drawbacks with the blacklist-based methods, the automatic detection of phishing URLs based on machine learning has attracted attention in research [15], [16]. These approaches can be grouped into four classes according to the type of data used for the detection: the text of the URL, the page content, the visual features, and networking information. Methods based on the page content and visual features require visiting the website to collect the source code and render it, which is a time-consuming task. Other availability limitations can be found in studies that rely on networking and 3rd party information such as WHOIS or search engine rankings. To overcome these limitations, we focus on phishing detection through URL since it implies advantages such as fast computation -because no websites are loaded- and 3rdparty and language-independent since features are extracted only from the URLs.

2. Proposed System

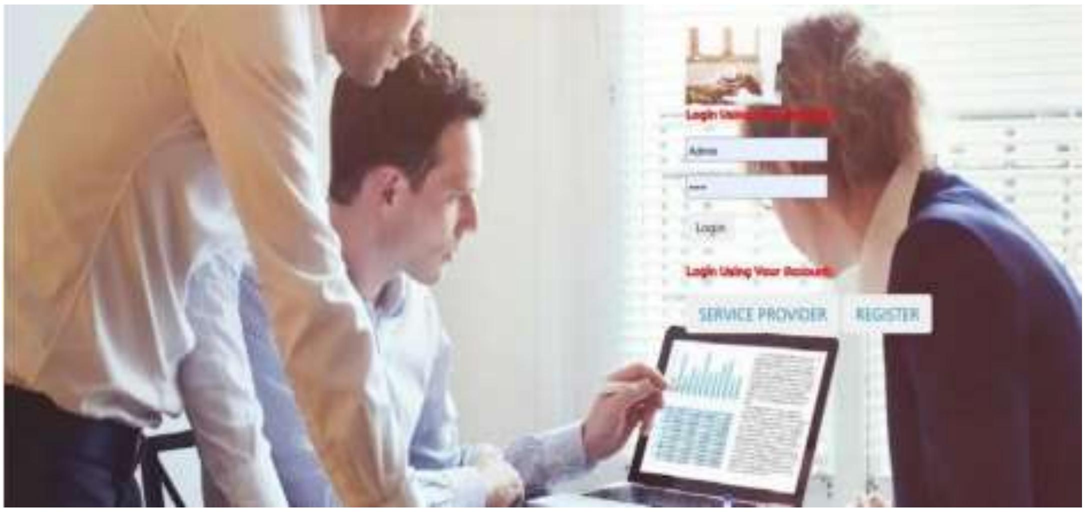
This paper presents a phishing URL dataset using legitimate login websites to obtain the URLs from such pages. Then, we evaluate machine and deep learning techniques for recommending the method with higher accuracy. Next, we show how models trained with legitimate homepages struggle to classify legitimate login URLs, demonstrating our hypothesis about phishing detection and legitimate login URLs. Additionally, we show how the accuracy decreases with the time on models trained with datasets from 2016 and evaluated on data collected in 2020. Finally, we provide an overview of current phishing encounters, explaining attacker tricks and approaches. We extended our previous dataset PILU-60K (Phishing Index Login URL), from60K to 90K URLs equally distributed among three classes: phishing, the legitimate home page, and legitimate login. We make this extended dataset, PILU-90K, publicly available for research purposes Using PILU-90K, we implemented and evaluated three pipelines for URL phishing detection: (i) we use the 38 handcrafted feature descriptors for training eight supervised machine learning classifiers and also (ii) automatic feature extraction using Term Frequency Inverse Document Frequency (TFIDF) at character N-gram level combined with Logistic Regression (LR) algorithm, and (iii) a Convolutional Neural Network (CNN) at character level too. We demonstrated empirically how an URL phishing detection model struggles in classifying login URLs when it was trained on the URLs of the homepage of phishing and legitimate URLs. We evaluated the robustness of the proposed phishing detection over time. We trained the model on a dataset collected between March 2016 and April 2016, and we evaluated the model on other datasets collected between 2017 and 2020. Phishing websites were analyzed using domain frequency. We found six different phishing domains depending on the service hired by the attacker.

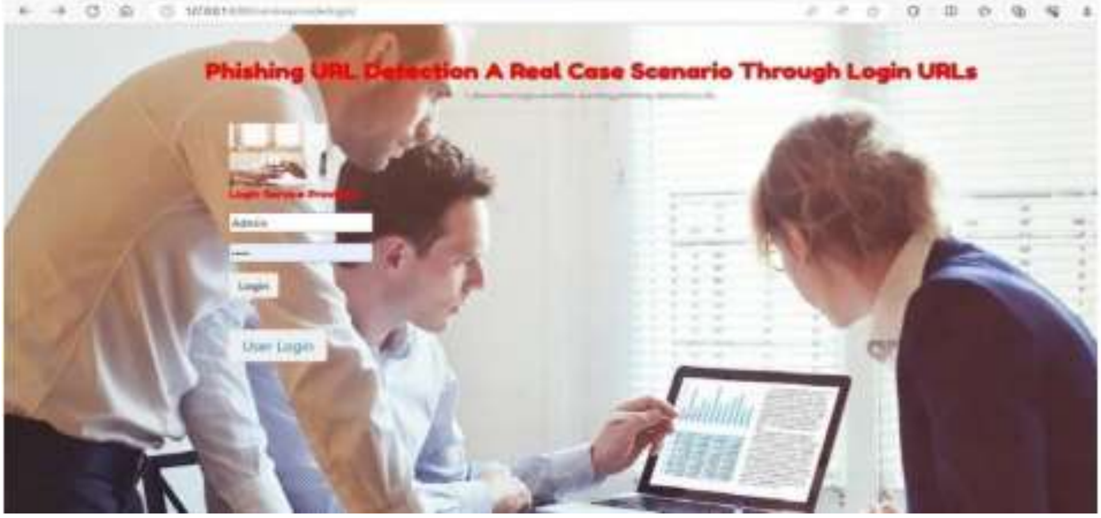
2.1 Advantages

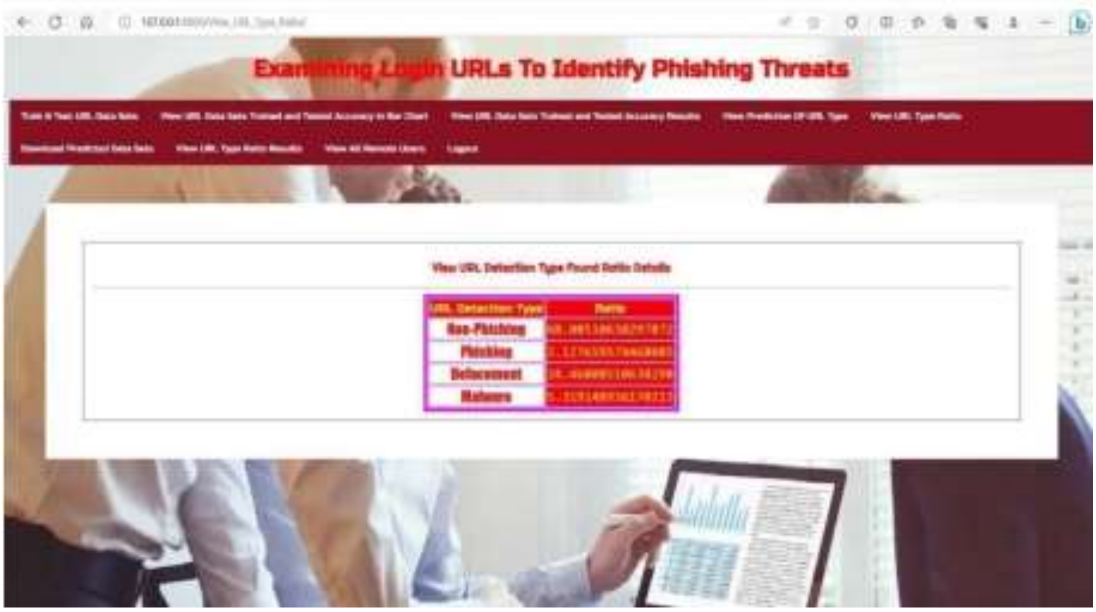
Machine learning models to detect unreported phishing encounters. Depending on their input data, these approaches can be classified into two categories: URL-based and content based.

Present an extended version of the Phishing Index Login URL (PILU- 60K) dataset and we name it PILU-90K. PILU-90K contains 90K URLs divided into three classes.

3. Results

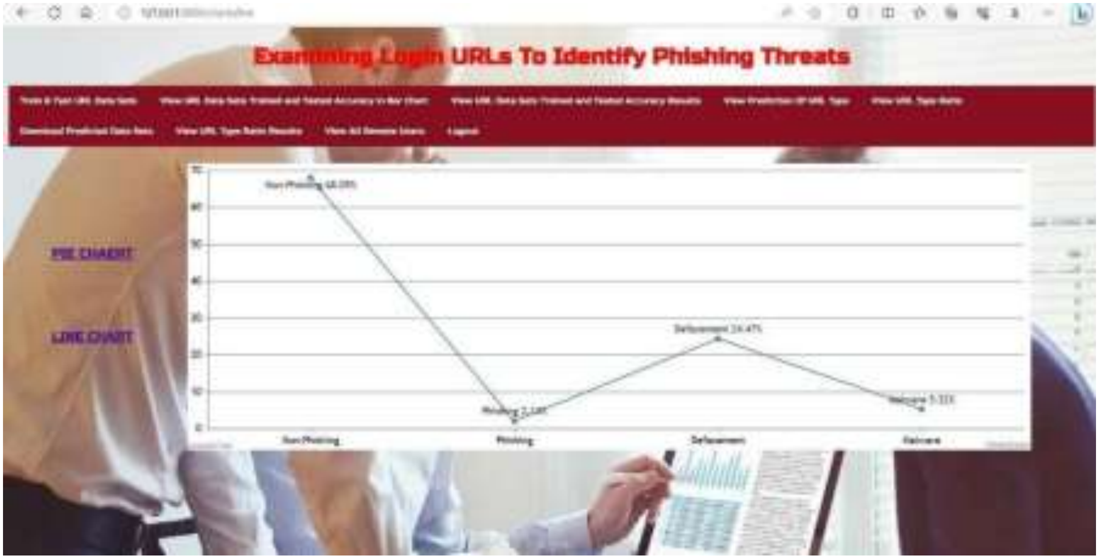


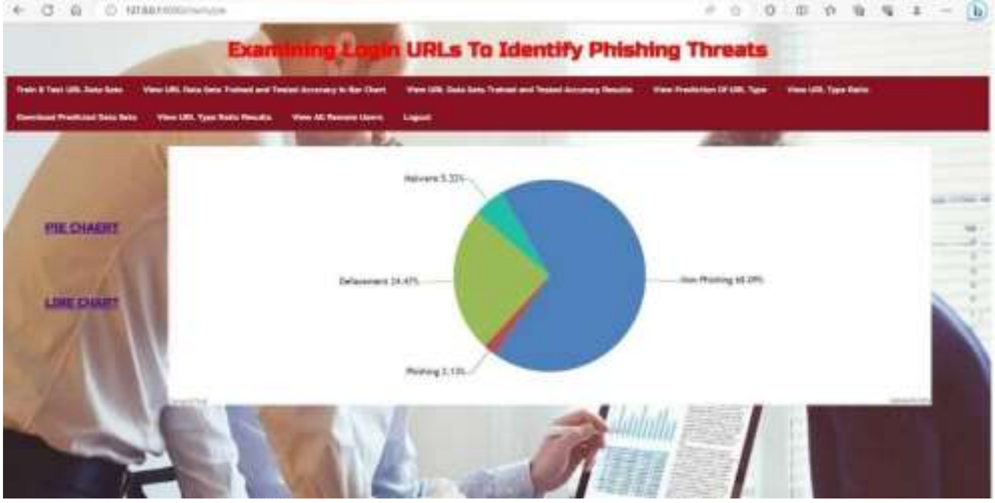

Phishing URL Detection A Real Case Scenario Through Login URLs
Login Service Provider
Admin
Login
User Login

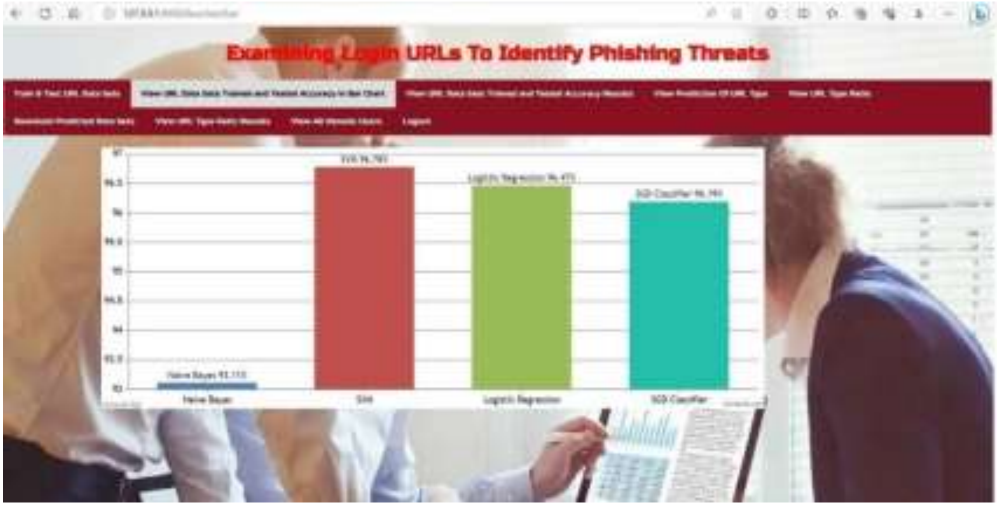

Examining Login URLs To Identify Phishing Threats
Train & Test URL Data Sets
View URL Data Sets Trained and Tested Accuracy in Bar Chart
View URL Data Sets Trained and Tested Accuracy Results
View Prediction Of URL Type
View URL Type Ratio
Download Predicted Data Sets
View URL Type Ratio Results
View All Remote Users
Logout

View URL Detection Type Found Ratio Details

URL Detection Type	Ratio
Non-Phishing	[illegible]
Phishing	[illegible]
Defacement	[illegible]
Malware	[illegible]


Examining Login URLs To Identify Phishing Threats
Train & Test URL Data Sets
View URL Data Sets Trained and Tested Accuracy in Bar Chart
View URL Data Sets Trained and Tested Accuracy Results
View Prediction Of URL Type
View URL Type Ratio
Download Predicted Data Sets
View URL Type Ratio Results
View All Remote Users
Logout
PIE CHART
LINE CHART
70
60
50
40
30
20
10
0
Non Phishing 68.09%
Phishing 2.11%
Defacement 24.47%
Malware 5.33%
Non Phishing
Phishing
Defacement
Malware


Examining Login URLs To Identify Phishing Threats
Train & Test URL Data Sets
View URL Data Sets Trained and Tested Accuracy in Bar Chart
View URL Data Sets Trained and Tested Accuracy Results
View Prediction Of URL Type
View URL Type Ratio
Download Predicted Data Sets
View URL Type Ratio Results
View All Remote Users
Logout
PIE CHART
LINE CHART
Malware 5.33%
Defacement 24.47%
Non Phishing 68.09%
Phishing 2.11%


Examining Login URLs To Identify Phishing Threats
Train & Test URL Data Sets
View URL Data Sets Trained and Tested Accuracy in Bar Chart
View URL Data Sets Trained and Tested Accuracy Results
View Prediction Of URL Type
View URL Type Ratio
Download Predicted Data Sets
View URL Type Ratio Results
View All Remote Users
Logout
97
96.5
96
95.5
95
94.5
94
93.5
93
SVM 96.78%
Logistic Regression 96.47%
SGD Classifier 96.14%
Naive Bayes 93.11%
Naive Bayes
SVM
Logistic Regression
SGD Classifier

4. Conclusion

The phishing detection mechanism aims to enhance existing blacklist methods and protect users from malicious login forms. Our research work introduces a new dataset called PILU-90K, which researchers can utilize to train and evaluate their approaches. This dataset consists of legitimate login URLs, which are highly representative of real-world phishing detection scenarios. In our study, we explored various URL-based detection models that employed deep learning and machine learning solutions. These models were trained using both phishing and legitimate home URLs. One significant advantage of our approach is its ability to achieve a low false-positive rate when classifying this type of URL. Among the different models we evaluated, the SVM algorithm yielded the best results, with an accuracy of 96.78%. This outperformed the current state-of-the-art methods. We demonstrated that phishing URL detection systems trained with legitimate landing page URLs struggle to classify legitimate login URLs correctly. Even the best-performing models we tested could only achieve a 69.50% accuracy in classifying these URLs, resulting in a high false positive rate. Therefore, we recommend that phishing detectors intended for real-world usage should be trained using legitimate login websites, such as our PLU-60K dataset, instead of homepages. Although using login websites for training slightly reduces overall accuracy due to the similarity between phishing and legitimate samples, this trade-off is justified considering the high false-positive rates of existing methods.

Future Scope

As phishing techniques evolve, there is a need for more advanced machine learning algorithms to accurately identify phishing URLs. Future research can focus on developing and refining algorithms that can detect subtle patterns and anomalies in login URLs to improve phishing detection rates. Deep learning techniques, combined with NLP, can be employed to analyze the content and context of a login URL. By understanding the semantic meaning and intent behind the URL, it becomes possible to identify phishing attempts that employ sophisticated obfuscation techniques. Implementing real-time analysis of login URLs can significantly enhance phishing detection capabilities. By leveraging cloud computing and distributed systems, it becomes feasible to scan URLs in real-time and provide immediate warnings or block access to potentially malicious sites. Phishing attacks often target multiple organizations or individuals simultaneously. Therefore, future efforts can focus on establishing collaborative platforms that allow organizations and security researchers to share data, insights, and threat intelligence related to phishing URLs. Such collaboration can lead to a more comprehensive understanding of evolving phishing techniques and better protection for users. With the increasing use of smartphones and mobile devices, the scope of phishing threats expands to these platforms. Future research can explore techniques specifically designed for detecting phishing URLs on mobile devices, taking into account the unique characteristics and constraints of these platforms.

References

[1] Statista. (2020). Adoption Rate of Emerging Technologies in OrganizationsWorldwide as of 2020. Accessed: Sep. 12, 2021. [Online]. Available: https://www.statista.com/statistics/661164/worldwide-cio- surveyoperati%onpriorities/

[2] R. De', N. Pandey, and A. Pal, ``Impact of digital surge during COVID-19 pandemic: A viewpoint on research and practice," Int. J. Inf. Manage., vol. 55, Dec.2020, Art. no. 102171.

[3] P. Patel, D. M. Sarno, J. E. Lewis, M. Shoss, M. B. Neider, and C. J. Bohil, ``Perceptual representation of spam and phishing emails," Appl. Cognit.Psychol., vol. 33, no. 6, pp. 1296_1304, Nov. 2019.

[4] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, ``Phishing attacks and defenses," Int. J. Secur. Appl., vol. 10, no. 1, pp. 247_256, 2016.

[5] M. Hijji and G. Alam, ``A multivocal literature review on growing social engineering-based cyber-attacks/threats during the COVID-19 pandemic: Challenges and prospective solutions," IEEE Access, vol. 9, pp. 7152_7169, 2021.

[6] Alzahrani, ``Coronavirus social engineering attacks: Issues and recommendations," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 5,pp. 154_161, 2020.

[7] Phishing Activity Trends Report 3Q, Anti-Phishing Working Group, International, 2017. Accessed: Sep. 12, 2021.

[8] Phishing Activity Trends Report 1Q, Anti-Phishing Working Group,International, 2021. Accessed: Sep. 14, 2021.

[9] R.Chen,J. Gaia, and H.R. Rao, ``An examination of the effect of recent phishingencounters on phishing susceptibility,'' Decis. Support Syst., vol. 133, Jun. 2020, Art. no. 113287. encounters on phishing susceptibility,'' Decis. Support Syst., vol. 133, Jun. 2020, Art. no. 113287.

[10] S. Bell and P. Komisarczuk, ``An analysis of phishing blacklists: Google safebrowsing, OpenPhish, and PhishTank,'' in Proc. Australas. Comput. Sci. Week Multiconf., Feb. 2020, pp. 1_11.

[11] Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili,A. Doupé, and G.-J. Ahn, ``Phishtime: Continuouslongitudinal measurement of the effectiveness of anti-phishing blacklists,'' in Proc. 29thUSENIX Security. Symp., 2020, pp. 379_396.

[12] L. Li, E. Berki, M. Helenius, and S. Ovaska, ``Towards a contingency approach with whitelist- and blacklist-based anti-phishing applications: What do usability tests indicate?'' Behavior Inf. Technol., vol. 33, no. 11,pp.1136_1147, Nov. 2014.

[13] N. Samarasinghe and M. Mannan, ``On cloaking behaviors of malicious websites,'' Comput. Secur., vol. 101, pp. 102_114, Feb. 2021.

[14] L. Halgas, I. Agra_otis, and J. R. C. Nurse, ``Catching the phish: Detecting phishing attacks using recurrent neural networks (RNNs),'' in Information Security Applications (Lecture Notes in Computer Science), vol. 11897. Cham, Switzerland: Springer, 2020, pp. 219_233.

[15] R. S. Rao and A. R. Pais, ``Jail-phish: An improved search engine based phishing detection system,'' Comput Secure. vol. 83, pp. 246_267, Jun. 2019.

[16] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani,``Systematization of knowledge (SoK): A systematic review of software- based webphishing detection,'' IEEE Commun. Surveys Tuts., vol. 19,no. 4, pp. 2797_2819, 4th Quart., 2017.