JOURNAL OF INFORMATION SYSTEMS APPLIED RESEARCH

Volume 12, No. 1
April 2019
ISSN: 1946-1836

In this issue:

iscap
Information Systems & Computing Academic Professionals

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer reviewed academic journal published by ISCAP, Information Systems and Computing Academic Professionals. Publishing frequency is three issues a year. The first date of publication was December 1, 2008.

JISAR is published online (http://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (http://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of AITP-EDSIG who perform the editorial and review processes for JISAR.

2019 Education Special Interest Group (EDSIG) Board of Directors

Journal of
Information Systems Applied Research

Drone Delivery Services: An Evaluation of Personal Innovativeness, Opinion Passing and Key Information Technology Adoption Factors

Charlie Chen
chench@appstate.edu

Hoon S. Choi
choihs@appstate.edu

Computer Information Systems and Supply Chain Management
Appalachian State University
Boone, NC, United States

Danuvasin Charoen
danuvasin@nida.ac.th

NIDA Business School
National Institute of Development Administration
Bangkok, Thailand

Abstract

The commercialization of drone delivery service is facing regulatory, technical and customer issues in the United States. This study examines key factors encouraging the adoption of drone delivery services by customers. Key factors were adopted from three major theories: social network theory, utility marketing theory and information technology adoption theory. This study surveyed 157 college students and their related family members and used the collected data to analyze the relationships among five key factors: opinion passing, personal innovativeness, perceived ease of use, and perceived usefulness, and the intention to adopt drone delivery services. Perceived usefulness shows the strongest impact on the intention to adopt drone delivery services, followed by personal innovativeness. In contrast, opinion passing, and perceived ease of use have no effect on the intention. Academic and practical implications were provided based on the findings. Future research directions and limitations were also discussed to conclude the study.

Keywords: Drone, personal innovativeness, opinion passing, perceived ease of use, perceived usefulness, behavioral intention

1. INTRODUCTION

A growing number of industries, including retail, food, and healthcare, are embracing drone delivery services because they can potentially streamline their supply chain operations. For instance, the healthcare industry is experimenting with using drones to deliver medical supplies to medically-underserved populations in rural areas, as well as to deliver prescriptions to the front door of patients’ houses in urban areas in order to reduce long wait times

(Lin, Shah, Mauntel, & Shah, 2018). With supply chain disruption due to natural disasters, non-government organizations (NGOs) have mobilized drones to deliver emergency commodities to disaster-affected regions and to minimize the distribution costs of disaster-relief operations (Chowdhury, Emelogu, Marufuzzaman, Nurre, & Bian, 2017). In the private sector, the first delivery service started by Flytrex in Reykjavik, Iceland in 2017 (Gilchrist, 2017). In Shanghai, China, ele.me, one of the major online food delivery platforms in China, began a drone delivery service in 2018, delivering take-out meals across the city (Whittaker, 2018). In the U.S., major advocates of drone delivery services, such as Amazon, UPS, and Domino's, are continuously experimenting and fine-tuning the technology. According to the Wall Street Journal, U.S. federal regulators and industry officials expect to start drone delivery service for limited packages in 2018 (Pasztor, 2018). Given these changes in the market, the adoption and diffusion of the technology must be just around the corner although the current drone delivery services are mostly justified for public use. Therefore, it is important to understand the relevant factors to adoption of technology in the perspective of its potential users. This study adopts a customer-driven approach in order to understand the key factors that can motivate users to adopt drone delivery services.

The customer-driven approach considers the adoption of drone delivery services from the perspective of the end user. Customers are more likely to be attracted and loyal to a product or service when it responds to their needs (Luigi, Oana, Mihai, & Simona, 2011). A simple way to understand the adoption cycle of new technology is to break it into three stages: awareness, consideration, and adoption. Since many customers are still not aware of the usefulness of drone delivery services, it is important to generate and disseminate lead sources (e.g. customer success stories, product comparisons, etc.) to potential users in order to trigger them into stepping into the process (Koshner, 1997). Marketing theory has shown that the word-of-month (WOM) marketing strategy is particularly useful at the early product introduction stage because it can help a company identify target customers and offer incentives to increase their interest in adopting the new technology (Kulviwat, Bruner Ii, Kumar, Nasco, & Clark, 2007). The interpersonal behavioral approach sometimes can be more effective than the technology acceptance approach in terms of predicting the behavioral use of new technology (Huang, 2017). With the growing number of customers consuming product information via social media, WOM is becoming even more important than ever due to its network effect, and because of its salient influence, this study considers opinion passing, one key WOM element, as a critical factor in the acceptance of drone delivery services at the early stage.

Personal innovativeness is an important individual trait that has marketing implications in terms of information technology adoption. The inclusion of personal innovativeness can help one understand how to accelerate the technology diffusion process during the inception of a new technology (Ritu & Jayesh, 1998). With limited resources, it is more effective to accelerate the technology diffusion process by identifying individuals with high personal innovativeness as key change agents and opinion leaders (Rogers, 1983). For companies interested in offering drone delivery services with limited resources, they will need to first and foremost learn to identify individuals with high personal innovativeness traits and rely on them to convince the majority of users to adopt this innovative service.

Perceived ease of use (PEOU) and perceived usefulness (PU) are two central constructs of the technology acceptance model (TAM), as they mediate the relationship between external variables (e.g. subjective norm, perceived risks, job relevance) and the behavioral intention to use new technology (Fred D. Davis, Richard P. Bagozzi, & Paul R. Warshaw, 1989). PEOU is a user's subjective belief about the degree of ease of using a particular technology, and PU is a user's subjective belief about the ability of using particular technology to enhance his/her job performance. While these two central factors are distinct from each other and have a direct effect on behavioral intention (Kulviwat et al., 2007), PEOU as a primary belief factor has a direct effect on the secondary belief factor of PU (Abdullah, Ward, & Ahmed, 2016).

Using survey data collected from college students, this study integrates social network, marketing, and technology acceptance theories, and explores the potential influence of the key factors of each theory as regards the behavioral intention to adopt drone delivery services. In addition, the study assesses the relative influence of each factor on the intention of users to adopt this innovative service. With improved understanding, the study can provide insights into how to maximize the use of limited resources in order to promote the diffusion of drone delivery services into the pubic.

The remainder of this paper is organized as follows. The literature related to behavioral intention to use drone delivery services and its marketing, technology adoption, and personal innovativeness antecedents will be examined, followed by a research model and hypotheses based on the theoretical foundations. The research methodology is discussed with respect to the research design, data collection, and analysis method. Data analysis results is reported, suggesting the theoretical and practical implications drawn from the results. Finally, research directions and limitations are presented to conclude the study.

2. CONCEPTUAL FORMATION

Social network theory asserts that social relationships are a network structure that emerges from the interactions of social actors. As such, the social actors within the same network often affect each other in the decision-making process, including new technology adoption (Vannoy & Palvia, 2010). Social or peer influence can effectively encourage members of an online community for example to engage in a series of co-innovation activities (Wang, Hsiao, Yang, & Hajli, 2016).

In terms of drone delivery services, during the early adoption phase, most users are not aware of the existence of these services. The proliferation of social network services (SNS), such as Snapchat, Twitter, Instagram, and Facebook, provides a platform for users to connect with those innovative users that are knowledgeable about a wide variety of drone delivery services. A durable social network can be constructed with the emphasis of increasing the awareness of more innovative users about drone delivery services. These users will then expand their personal networks and help more innovative network participants with otherwise unattainable resources, such as access to information about drone delivery services and their potential benefits. A growing number of studies have tried to construct an "interaction network" or "conversation map" with Twitter and other social media in order to identify and profile the main constituencies (e.g. influencers) discussing a specific topic or a specific product or service (Kwak & Kim, 2017). One of such main constituencies could be users with personal innovativeness traits.

The Effect of Personal Innovativeness Traits on Increasing Passing Opinions about Drone Delivery Services

One prevalent social networking phenomenon is the growing use of Electronic word-of-mouth (eWOM) to disseminate product and service information. eWOM in SNS is often conceptualized into three elements: opinion seeking, opinion giving, and opinion passing (Chu & Kim, 2011a). The first two elements are common to online and offline WOM activities. However, the third element is unique in eWOM as social actors need to be willing to share content with others after identifying and generating useful content. Social influence is particularly effective at boosting network externalities for communities consisting of numerous small sub-networks (Kwak & Kim, 2017). Innovative users are more likely to pass personal opinions to other social actors as a social influence method and to increase network externality.

Drone delivery services are currently favored by scattered, small groups of communities. It is important for their providers to focus on using opinion passing as a social influence method in order to increase network size and externalities. A larger installed base can become an incentive for more innovative users to join the network and to learn more about drone delivery services (Henkel & Block, 2013). For instance, many companies have attempted to explicitly incentivize opinion-passing activities by rewarding customers that can help cross-sell and up-sell products (Godinho de Matos, Ferreira, & Krackhardt, 2014). AT&T for example offers the Buy One, Get One for customers interested in buying an iPhone X if they add a new line in addition to upgrading their current phone line. This example shows that it is important to identify innovative users and to encourage them to pass opinions and to motivate others in SNS to consider adopting drone delivery services. Thus, the following hypothesis is proposed:

H1: Personal innovativeness has a positive effect on opinion passing concerning the adoption of drone delivery services.

The Positive Effect of Personal Innovativeness on the Perceived Usefulness of the Adoption of Drone Delivery Services

Users with high personal innovativeness traits are more willing to try out new products or services (Lu, 2014). When these novelty seekers have positive experiences with new technology, they tend to have more positive perceptions of its usefulness (Ritu & Jayesh, 1998). Personal innovativeness is an effective predictor of the adoption of information technology, such as virtual reality simulation (Mary, Carol, & Vivek, 2012) and mobile commerce (Daştan & Gürler, 2016), and such a correlation is also likely to be present for the adoption of drone delivery services. As novelty seekers try out drone delivery services, they would become enthusiastic at communicating their usefulness to others. As such, users with a high degree of innovativeness are more likely to participate in co-creation activities, often perceived as useful. Thus, the following is proposed:

H2: Personal innovativeness has a positive effect on the perceived usefulness of drone delivery services.

The Effect of Opinion Passing on Increasing the Intention of Users to Adopt Drone Delivery Services

Social influence is a strong predictor of the intention to use, rather than performance expectations regarding new technology (Jennings, Arlikatti, & Andrew). Opinion passing by innovative users is an important social influence activity because it can influence less innovative users to adopt new products/services. For instance, farmers often trust the opinion of other farmers whom they consider successful and innovative in their farming operations even though they have different professional and personal characteristics (Genius, Koundouri, Nauges, & Tzouvelekas, 2014).

Opinion passing can have impact on the intention of users to adopt different products or services. In hospital environments, for example, the attitude of senior medical doctors can affect the intention of young medical doctors to use clinical informatics systems (Abyaomi, Evans, & Ocholla, 2017). In higher education institutions, the opinions among students can increase their intention to adopt information, communication, and technology (ICT) tools (Rosaline & Wesley, 2017). In face of the option of using advanced driver assistance systems (ADAS) to reduce the number of crashes, and enhance driver comfort, positive opinions among one’s peers can also exert a significant influence on the adoption decision of drivers (Rahman, Lesch, Horrey, & Strawderman, 2017). Drone delivery services offer enticing reasons (e.g. cheaper and faster shipping at anytime and anywhere) for users to adopt them, and an increasing number of users are recognizing the positive reasons for doing so. The positive opinions of these users can potentially increase the intention of others to adopt these services; thus the following is proposed:

H3: Opinion passing has a positive effect on increasing the intention of users to adopt drone delivery services.

The Positive Effect of Personal Innovativeness on the Intention of Users to Adopt Drone Delivery Services

Personal innovativeness has a positive influence on utilitarian and hedonic values (Hong, Hsieh, & Lin), and utilitarian values include perceived usefulness and expected performance, while hedonic values concern the fun and playfulness perceptions of users (Lu, 2014). Along with the increase in the utilitarian and hedonic values of technology comes the increased intention of users to adopt this technology. The causal effect of personal innovativeness on the increase of adoption intention can be found in many information technologies, such as mobile hotel booking systems (Ozturk, Nusair, Okumus, & Hua, 2016) and smart watches (Hong, Hsieh, & Lin, 2017). It is plausible that drone delivery services could also be susceptible to the positive influence of personal innovativeness. Thus, the following is proposed:

H4: Personal innovativeness can increase the intention of users to adopt drone delivery services.

The Positive Effect of Perceived Usefulness on the Intention of Users to Adopt Drone Delivery Services

Perceived usefulness is an important determinant of beginning and continuation of emerging technology, such as mobile commerce applications (Daştan & Gürler, 2016; Eun-Yong, Soo-Bum, & Yu Jung Jennifer, 2017), food delivery applications (Eun-Yong et al., 2017), and mobile tourism applications (Chen & Tsai, 2017). Drones are useful at reducing delivery times and human labor, as well as possible to deliver customers a wide variety of products, such as small parcels, medications, pizza, etc. When users perceive the benefits of using drones to deliver items to them, they are more likely to adopt these services. Thus, the following is proposed:

H5: Perceived usefulness can increase the intention of users to adopt drone delivery services.

The Positive Effect of Perceived Ease of Use on the Intention of Users to Adopt Drone Delivery Services

Perceived ease of use (PEOU) is a person's belief about using a particular system without spending too much effort (Fred D. Davis et al., 1989). PEOU is a strong predictor of the intention to use rather than performance expectations regarding new technology (Jennings et al.). PEOU has a positive effect on the intention of users to adopt different technologies, such as virtual reality learning (Hsiu-Mei & Shu-Sheng, 2018), cloud computing (Shana & Abulibdeh, 2017), educational video games (Sánchez-Mena, Martí-Parreño, & Aldás-Manzano, 2017), and electronic health records (Tubaishat, 2017). Flying a drone to deliver goods to customers requires a team of drone experts and GPS technology; however, receiving drone delivery services is free of effort, as users only need to know where to pick up their delivered items after a drone drops off them. Because of the perceived ease of use, users are more likely to adopt drone delivery services; thus the following is proposed:

H6: Perceived ease of use can increase the intention of users to adopt drone delivery services.

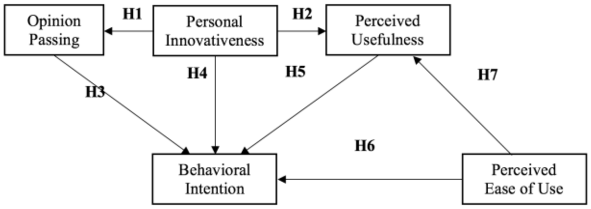
The Positive Effect of Perceived Ease of Use on the Perceived Usefulness of Drone Delivery Services

Although both perceived ease of use and perceived usefulness are usability factors (Ifinedo, Pyke, & Anwar, 2018), they are distinct from each other and the former factor has a positive influence on the latter factor (Fred D. Davis et al., 1989). By increasing the PEOU of the adopted system, users are more likely to perform better and contribute their improved performance to the increased PEOU. In order to enjoy the benefits of using drone delivery services, users will need to know how to place an order online. For instance, they turn on the location service for the drone delivery application, receive an alert for product delivery information, and use GPS to identify the delivery location. Simplifying these tasks can directly increase the users' PEOU, thereby improving their PU for the drone delivery service. Thus, the following is proposed:

H7: Perceived ease of use has a positive influence on the perceived usefulness of drone delivery services.

The above discussion leads to the development of the present research model (Figure 1).

Figure 1. Theoretical Model for the Adoption of Drone Delivery Services



3. RESEARCH METHODOLOGY

Survey

In order to test the hypotheses proposed, an online survey was conducted to collect data. Participants of the survey watched a drone delivery video to have a clear understanding on the service before answering questions. They answered the questions based on the assumption that a drone delivery service will be introduced soon. Total 182 undergraduate students participated in the survey. They were taking a required business course, *Information Technology in the Organization*, in a public university in the USA. Those students voluntarily participated in the study in order to receive 0.5% of their final grade as extra credit. We used a total 157 responses for analysis, excluding 25 invalid, incomplete responses. Table 1 below illustrates the profile of the respondents.

Table 1. Profile of Respondents

Category	Group	Frequency	Portion
Gender	Males	91	57.9
	Females	66	42.1
Age	18-22	97	61.8
	23-30	29	18.4
	31-50	16	10.2
	over 50	15	9.6

Reliability and Validity of Survey Instrument

Existing items were employed to measure the major constructs of this study (Appendix 1). The questions for the constructs were placed on a 5-point Likert scale, ranging from 1 "strongly disagree" to 5 "strongly agree." Table 2 presents the constructs and their sources, including their loadings. We discarded items with loadings lower than 0.7 in order to ensure their indicator reliability (Chin, 2010).

Additional tests were performed to ensure the validity and reliability of the constructs. Cronbach's α coefficients for the measurement were higher than the acceptable cut-off value of 0.7 (Chin, 2010; Hair, Sarstedt, Pieper, & Ringle, 2012), suggesting internal consistency reliability. Convergent validity was examined with composite reliability and average variance extracted (AVE) and all of the values for composite reliability exceeded the recommended threshold of 0.7 (Fornell & Larcker, 1981), with the smallest AVE being 0.547, which is larger than the cut-off of 0.5 (Fornell & Larcker, 1981; Hulland, 1999). In addition, the square root of the construct's AVE was greater than the correlations with other constructs, ensuring discriminant validity of the measurement (Chin, 2010). Lastly, we checked the variance inflation factors (VIFs) of the constructs in order to determine the degree of multicollinearity. The VIFs ranged from 1.55 to 4.61, which was far below the recommended threshold of 10 (Chin, 2010), suggesting no significant multicollinearity in the model. Table 2 and Table 3 respectively summarize the model quality indicators and the correlations with square root of AVE on the diagonal discussed.

Table 2. Quality Indicators

Construct	CA	CR	AVE	VIF
OP	0.773	0.969	0.562	1.832
PEOU	0.898	0.989	0.698	3.147
PU	0.822	0.979	0.642	2.156
PI	0.748	0.935	0.600	1.554
BI	0.924	0.992	0.645	4.161

※ **CA:** Cronbach's α, **CR:** Composite Reliability, **AVE:** Average Variance Extracted, **VIF:** Variance Inflation Factor

Table 3. Correlations with Square Root of AVE on the Diagonal

Construct	OP	PEOU	PU	PI	BI
OP	**0.750**				
PEOU	0.110	**0.834**			
PU	0.310	0.701	**0.740**		
PI	0.445	0.110	0.190	**0.777**	
BI	0.274	0.506	0.611	0.311	**0.837**

Structural Model and Hypothesis Test

Structural Equation Modeling (SEM) with Partial Least Squares (PLS) was employed to test the proposed hypotheses. SEM is a reliable technique to test multiple causal relationships (Henseler, Ringle, & Sinkovics, 2009), and is not sensitive to the issues about population, scale of measurement, and residual distribution (Chin, 1998; Fornell & Bookstein, 1982). Table 4 and Figure 2 summarize the results of the hypothesis tests.

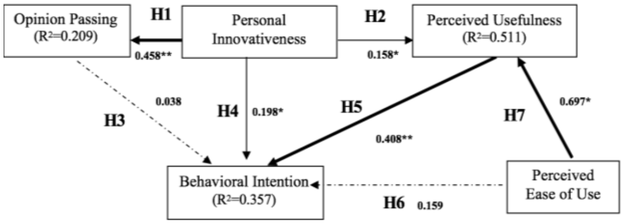
Personal Innovativeness (PI) explained 20.9% of the variance in Opinion Passing (OP) ($R^2 = 0.209$). PI had a positive influence on OP at the 99% confidence level ($\beta = 0.458$; $t = 5.33$), supporting Hypothesis 1. Hypothesis 2 was supported at the 95% level ($\beta = 0.158$; $t = 1.99$), suggesting a positive impact of PI on Perceived Usefulness (PU). PI and Perceived Ease of Use (PEOU) explained 51.1% of the variance in PU ($R^2 = 0.511$). Hypothesis 4 was supported at the 95% level, indicating a positive effect of PI on Behavioral Intention (BI) to use drone delivery. The effect of PU on BI was positive and statistically significant at the 99% level, supporting Hypothesis 5. OP, PI, PU, and PEOU together explained 35.7% of the variance in BI ($R^2 = 0.357$). Finally, Hypothesis 7 was supported at the 99% level, suggesting a positive impact of PEOU on PU. Different from the prediction, however, was Hypothesis 3 and Hypothesis 6, which were found to be not statistically significant. This suggests that OP and PEOU does not have a relationship with BI. One potential reason that ease of use may not be a good measure of behavioral intention can depend on how drone delivery is implemented. For instance, if it is just a third option on the screen (e.g. 1-day shipping, ground or 2-hour drone), ease of use may never really have an impact on behavioral intention.

Table 4. Results of Hypothesis Testing

Hypothesis	Coeff.	t-stat.	Result
H1: PI → OP	0.458	5.33**	*Supported*
H2: PI → PU	0.158	1.99*	*Supported*
H3: OP → BI	0.038	0.42	*Not Supported*
H4: PI → BI	0.198	2.11*	*Supported*
H5: PU → BI	0.408	3.52**	*Supported*
H6: PEOU→BI	0.159	1.40	*Not Supported*
H7: PEOU→PU	0.697	13.06**	*Supported*

※ **Significance:** *p < 0.05, **p <0.01

Figure 2. Research Model with Results of Hypothesis Testing



※ **Significance:** *p < 0.05, **p <0.01

4. DISCUSSION AND IMPLICATIONS

This study examined how marketing, technology adoption, and personal innovativeness factors affect the users' intention to adopt drone delivery services. Five out of seven proposed hypotheses were supported. Support of Hypotheses 1, 2, and 4 indicate that personal innovativeness plays a central role in the adoption of drone delivery services by increasing opinion passing, perceived usefulness, and behavioral intention. This finding confirms previous study concerning the strong predictive power of personal innovativeness regarding adoption intention, particularly in the domain of information technology (Mary et al., 2012).

In the survey, the subjects were asked to report the personal reasons why they would or would not like to use drone delivery services as soon as they become available from companies, such as Amazon, Domino's Pizza, and Netflix. Based on the responses, it was clear that some users had a high degree of personal innovativeness and these users were more likely to consider drone delivery services useful and to adopt them as soon as they become available. Innovative users expressed such opinions as the following:

"I want to try it out and see if it is for me or not;" "It will be interesting to see the package delivered and much faster and more reliable that with other modes of transportation."
"It seems very futuristic," and "I think it would be cool."

Thus, it is important to identify users with high personal innovativeness because this factor can promote technological opinion leadership and gadget lovers (Thakur, Angriawan, & Summey, 2016).

On the other hand, some participants showed reservations about embracing drone delivery services and would like to see more users adopt the service before considering adopting it. These users were considered less innovative and they made such comments as the following:

"I would only use it if absolute dire need situations where a delivery must be done in a matter of a few hours as in emergency situation."
"I would likely not use them right at first, but after seeing how well they work out I may consider using them depending on price and things."
"I just don't like the idea. Too many liability issues if the package doesn't get there or is damaged on delivery. I feel companies would be less likely to return my money to me even though it would be their fault."

These remarks affirm our finding about the importance of identifying users with personal innovativeness traits and encouraging them to become opinion leaders to help pass on opinions about the usefulness of drone delivery services to others.

Perceived usefulness had the largest impact on the user's intention to adopt drone delivery services based on our findings. This corroborates previous study where perceived usefulness had a direct positive effect on behavioral intention to adopt different technological services, such as mobile commerce (Kalinic & Marinkovic, 2016) and mobile instant messaging (Yoon, Jeong, & Rolland, 2015). The participants in this study considered drone delivery services to be useful mostly because of their fast delivery and convenience. The following comments confirm with our findings:

"From a business perspective, if I needed something ASAP (machine down, etc.), drone service would be beneficial."
"I would love 30min delivery!"
"Faster delivery."
"Get items quickly instead of having to wait or drive out to get them. Save Time!"
"It could be useful for time sensitive delivery."
"I would like to receive shipments sooner."
"You can get what you need with a short wait time."
"Convenient and fast delivery."
"Seems fast and convenient."

Perceived ease of use appeared not to be important to users without prior experience with using drone delivery services. When asked for their personal reasons why they would or would not adopt drone delivery services, most users emphasized usefulness, or expressed concerns about the services. The top positive reasons for the perceived usefulness included convenience, delivery speed, and the top negative reasons included security, noise, privacy, safety, price, and location. No users in our survey had comments directly related to the perceived ease of use for drone delivery services. Experience is one of the best predictors for the perceived ease of use of technology (Abdullah et al., 2016). After experiencing various drone delivery services, users may be able to make a better judgment about the importance of perceived ease of use. However, the subjects in this study lacked prior experience and could not validate the assumption.

Although perceived ease of use might not have a direct effect on the user's intention to adopt drone delivery services, it had a significant influence on perceived usefulness. This indicates a strong mediation effect of perceived usefulness in the relationship between perceived ease of use and the adoption intention of drone delivery services. This finding corroborates previous study, where perceived ease of use was able to improve the perceived usefulness of food delivery apps (Eun-Yong et al., 2017). However, their joint effect shown in the adoption of other technologies was not present in the behavioral adoption of drone delivery services.

Theoretical Contributions

One major theoretical contribution of this study is to construct an integrative research model and to empirically test the impacts of the marketing (opinion passing), personality traits (personal innovativeness), and technology adoption (perceived ease of use and perceived usefulness) factors on the intention to use drone delivery services. All of these factors together were found to explain approximately 35% of the variance in the decision to use these delivery services. This finding is theoretically valuable in the context of drone delivery services while it indicates that other factors that could better explain the remaining larger variance need to be further explored.

Another theoretical contribution is that prior work has mainly emphasized both perceived ease of use and perceived usefulness as two important antecedents of technology adoption. This study departs such prior work and suggests that perceived usefulness rather than perceived ease of use is a more important technology adoption factor with respect to its influence on the intention to adopt drone delivery services.

Moreover, this study found that personal innovativeness is a salient factor having a direct effect on perceived usefulness, opinion passing, and behavioral intention. This confirms the importance of identifying innovative users as opinion leaders for increasing the intention to use drone delivery services. Effective opinion leaders are the best marketing choice as they can not only increase the velocity of the technological diffusion process, but also the maximum cumulative number of adopters (Cho, Hwang, & Lee, 2012). Users with the personal innovativeness trait not only have a strong intention to adopt these services, but also have the tendency to tell others about their usefulness. Recruiting effective innovative users as opinion leaders is one effective way to promote the adoption and diffusion of drone delivery service (Jingyuan & Salmon, 2018) and to reduce the time to reach a critical mass in the domain (Cho et al., 2012).

Contrary to general expectations about the importance of opinion passing as a marketing factor, and perceived ease of use as a technology adoption factor, the finding of this study suggests that they do not have a direct impact on the adoption of drone delivery services. At the early adoption stage of these services, it is imperative to improve the perceived usefulness of innovative users so that they can influence others, thereby increasing their intention to adopt the services.

Practical Contributions

This study offers practical insights into promoting the adoption of drone delivery services by identifying innovative users and improving their perception concerning their usefulness. First, as our findings show, it is a valuable strategy for firms to identify and to encourage innovative users (individuals with the high personal innovativeness trait) in both online and offline social communities to pass their positive opinions to others concerning their drone delivery services. Digital opinion leaders' persuasive messages, for instance, are effective in changing the attitudes of followers and influencing them in their adoption and purchasing decisions (Huhn Nunes, Sabino de Freitas, & Leão Ramos, 2018). These innovative users are not only early adopters but they also can help articulate the usefulness of drone delivery services to other users.

Most users in the study perceived the usefulness of drone delivery services as fast and convenient. Firms providing the services should explore methods to improve the user's perception of their usefulness as well as improving the perceived ease of use of the services.

Opinion passing is an important eWOM activity; however, its impact on the adoption of drone delivery services is not as salient as was expected. Firms may want to consider using traditional word-of-mouth methods (e.g. hosting events, information sessions) (Chu & Kim, 2011a) to seek opinions and to provide feedback to prospective users of drone delivery services.

Limitations and Future Work

The findings of this study warrant careful interpretations because of certain limitations. First, the participants in the study were mostly students, even though they were encouraged to ask their family members to complete the same survey. Since the majority of the participants were students, the findings can be best generalizable to student users. Future research could recruit a larger sample size that represents different age and income groups. In this way, the findings could be more generalizable to the general population.

Second, all three factors (marketing, technology adoption, and personal innovativeness factors) examined in this paper could only explain 35.7 % of the variance in adoption intention of drone delivery services. Future research may want to explore other factors in the same areas or other relevant factors in other research areas. For instance, researchers can investigate whether traditional marketing methods and other technology adoption factors (e.g. perceived playfulness) can contribute to the remaining larger variance in the adoption of the intention to use drone delivery services. Some consumers are more concerned about the risk of a package being damaged by a drone even though it can be delivered quickly. Future research may want to focus on acceptance of drone delivery services from the risk/benefit perspective.

Third, drone delivery services vary with industries, and the users in the domains have different degrees of technology self-efficacy. All of these uncontrollable factors can possibly affect users' intention to adopt drone delivery services. Future work may want to divide a larger, general sample into different groups based on their technology self-efficacy and types of drone delivery services (e.g. delivering medical supplies, groceries, souvenirs, and food). Fourth, although innovative users can adopt drone delivery services themselves, passing their opinions on to others alone has no effect on the intention to adopt these services. WOM consists of three major elements: opinion seeking, opinion giving, and opinion passing. When these three elements work in concert, they can have a positive impact on the satisfaction level of users in the online community (Nagy, KemÉNy, SzŰCs, Simon, & Kiss, 2017). However, this study only investigated a single eWOM element: opinion passing. Future research may want to assess the relative and joint influence of these three essential WOM elements on the adoption intention of drone delivery services.

5. CONCLUSIONS

The demand for drone delivery services is growing as they have become more mature and their value propositions are attractive to not only business organizations but also users. In order to understand the driving factors in the adoption of drone delivery services, this study employed an interdisciplinary approach by combing marketing, technology adoption, and social network factors into an integrative research model.

Empirical survey data from 157 subjects examined the proposed hypotheses. The findings indicated that perceived usefulness had the largest impact on the intention to adopt drone delivery services, followed by personal innovativeness. In addition, personal innovativeness was found to play a central role not only in directly affecting the intention to adopt these delivery services, but also the users' perceived usefulness and opinion passing. These findings not only contribute to research on the adoption of drone delivery services, but also inform practitioners regarding the utilization of different methods to promote the use of drone delivery services.

6. REFERENCES

Abdullah, F., Ward, R., & Ahmed, E. (2016). Investigating the influence of the most commonly used external variables of TAM on students' Perceived Ease of Use (PEOU) and Perceived Usefulness (PU) of e-portfolios. *Computers in Human Behavior, 63*, 75-90. doi: 10.1016/j.chb.2016.05.014

Abyaomi, O. K. y. y. c., Evans, N. D. e. u. a. z., & Ocholla, D. N. O. u. a. z. (2017). FACTORS THAT INFLUENCE MEDICAL DOCTORS' BEHAVIOURAL INTENTION TO USE CLINICAL INFORMATICS. *Mousaion, 35*(1), 130-154. doi: 10.25159/0027-2639/2321

Chen, C.-C., & Tsai, J.-L. (2017). Determinants of behavioral intention to use the Personalized Location-based Mobile Tourism Application: An empirical study by integrating TAM with ISSM. *Future Generation Computer Systems*. doi: 10.1016/j.future.2017.02.028

Chin, W. W. (1998). The partial least squares approach to structural equation modeling.

Modern methods for business research, 295(2), 295-336.

Chin, W. W. (2010). How to write up and report PLS analyses *Handbook of partial least squares* (pp. 655-690): Springer.

Cho, Y., Hwang, J., & Lee, D. (2012). Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach. *Technological Forecasting & Social Change, 79*, 97-106. doi: 10.1016/j.techfore.2011.06.003

Chowdhury, S., Emelogu, A., Marufuzzaman, M., Nurre, S. G., & Bian, L. (2017). Drones for disaster response and relief operations: A continuous approximation model. *International Journal of Production Economics, 188*, 167-184. doi: 10.1016/j.ijpe.2017.03.024

Chu, S.-C., & Kim, Y. (2011a). Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. *International Journal of Advertising: The Quarterly Review of Marketing Communications, 30*(1), 47-75. doi: 10.2501/IJA-30-1-047-075

Chu, S.-C., & Kim, Y. (2011b). Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. *International journal of Advertising, 30*(1), 47-75.

Daştan, İ., & Gürler, C. (2016). Factors Affecting the Adoption of Mobile Payment Systems: An Empirical Analysis. *EMAJ: Emerging Markets Journal, 6*(1), 16-24. doi: 10.5195/emaj.2016.95

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management science, 35*(8), 982-1003.

Eun-Yong, L. E. E., Soo-Bum, L. E. E., & Yu Jung Jennifer, J. (2017). FACTORS INFLUENCING THE BEHAVIORAL INTENTION TO USE FOOD DELIVERY APPS. *Social Behavior & Personality: an international journal, 45*(9), 1461-1473.

Fornell, C., & Bookstein, F. L. (1982). Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research (JMR), 19*(4), 440-452.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of marketing research*, 382-388.

Fred D. Davis, a., Richard P. Bagozzi, a., & Paul R. Warshaw, a. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management science*(8), 982.

Genius, M., Koundouri, P., Nauges, C., & Tzouvelekas, V. (2014). Information Transmission in Irrigation Technology Adoption and Diffusion: Social Learning, Extension Services, and Spatial Effects. *American Journal of Agricultural Economics, 96*(1), 328-344. doi: https://academic.oup.com/ajae/issue

Gilchrist, K. (2017). World’s first drone delivery service launches in Iceland. Retrieved August 6, 2018, from https://www.cnbc.com/2017/08/22/worlds-first-drone-delivery-service-launches-in-iceland.html

Godinho de Matos, M., Ferreira, P., & Krackhardt, D. (2014). PEER INFLUENCE IN THE DIFFUSION OF IPHONE 3G OVER A LARGE SOCIAL NETWORK. *MIS quarterly, 38*(4), 1103-A1115.

Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012). The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. *Long range planning, 45*(5-6), 320-340.

Henkel, J., & Block, J. (2013). Peer influence in network markets: a theoretical and empirical analysis. *Journal of Evolutionary Economics, 23*(5), 925-953. doi: 10.1007/s00191-012-0302-4

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing *New challenges to international marketing* (pp. 277-319): Emerald Group Publishing Limited.

Hong, J.-C., Hsieh, P.-C., & Lin, P.-H. - The effect of consumer innovativeness on perceived value and continuance intention to use smartwatch.

Hong, J.-C., Hsieh, P.-C., & Lin, P.-H. (2017). - The effect of consumer innovativeness on perceived value and continuance intention to use smartwatch. *Computers in Human Behavior, 67*, 264-272.

Hsiu-Mei, H., & Shu-Sheng, L. (2018). An Analysis of Learners' Intentions Toward Virtual Reality Learning Based on Constructivist and Technology Acceptance Approaches. *International Review of Research in Open & Distance Learning, 19*(1), 91-115.

Huang, C.-C. (2017). Cognitive factors in predicting continued use of information systems with technology adoption models. *Information Research, 22*(2), 1-29.

Huhn Nunes, R., Sabino de Freitas, A., & Leão Ramos, F. (2018). The effects of social media opinion leaders' recommendations on followers' intention to buy. *Revista Brasileira de Gestão de Negócios, 20*(1), 57-73. doi: 10.7819/rbgn.v20i1.3678

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strategic management journal*, 195-204.

Ifinedo, P., Pyke, J., & Anwar, A. (2018). Business undergraduates' perceived use outcomes of Moodle in a blended learning environment: The roles of usability factors and external support. *Telematics and Informatics, 35*, 93-102. doi: 10.1016/j.tele.2017.10.001

Jennings, E., Arlikatti, S., & Andrew, S. - Determinants of Emergency Management Decision Support Software Technology: An Empirical Analysis of Social Influence in Technology Adoption.

Jingyuan, S., & Salmon, C. T. (2018). Identifying Opinion Leaders to Promote Organ Donation on Social Media: Network Study. *Journal of Medical Internet Research, 20*(1), 16-16. doi: 10.2196/jmir.7643

Kalinic, Z., & Marinkovic, V. (2016). Determinants of Users' Intention to Adopt M-Commerce: An Empirical Analysis. *Information Systems and e-Business Management, 14*(2), 367-387. doi: https://link.springer.com/journal/volumesAndIssues/10257

Koshner, E. L. (1997). A Market-Focused and Customer-Driven Approach to Growth. *Human Resource Planning, 20*(2), 9-14.

Kulviwat, S., Bruner Ii, G. C., Kumar, A., Nasco, S. A., & Clark, T. (2007). Toward a unified theory of consumer acceptance technology. *Psychology & Marketing, 24*(12), 1059-1084.

Kwak, D., & Kim, W. (2017). Understanding the process of social network evolution: Online-offline integrated analysis of social tie formation. *PLoS ONE, 12*(5), 1-16. doi: 10.1371/journal.pone.0177729

Lin, C. A., Shah, K., Mauntel, C., & Shah, S. A. (2018). Drone delivery of medications: Review of the landscape and legal considerations. *American Journal of Health-System Pharmacy, 75*(3), 153-158. doi: 10.2146/ajhp170196

Lu, J. (2014). Are personal innovativeness and social influence critical to continue with mobile commerce? *Internet Research*(2), 134. doi: 10.1108/IntR-05-2012-0100

Luigi, D., Oana, S., Mihai, T., & Simona, V. (2011). PURSUING A CUSTOMER-DRIVEN APPROACH FOR INNOVATION AND MARKETING EXCELLENCE. *Studies in Business & Economics, 6*(2), 19-26.

Mary, F., Carol, K., & Vivek, P. (2012). Exploring the adoption of a virtual reality simulation : The role of perceived ease of use, perceived usefulness and personal innovativeness. *Campus-Wide Information Systems*(2), 117. doi: 10.1108/10650741211212368

Nagy, A., KemÉNy, I., SzŰCs, K., Simon, J., & Kiss, V. (2017). ARE OPINION LEADERS MORE SATISFIED? RESULTS OF A SEM MODEL ABOUT THE RELATIONSHIP BETWEEN OPINION LEADERSHIP AND ONLINE CUSTOMER SATISFACTION. *Society & Economy, 39*(1), 141.

Ozturk, A. B., Nusair, K., Okumus, F., & Hua, N. (2016). The role of utilitarian and hedonic values on users' continued usage intention in a mobile hotel booking environment. *International Journal of Hospitality*

Management, 57, 106-115. doi: 10.1016/j.ijhm.2016.06.007

Pasztor, A. (2018). Coming Soon to a Front Porch Near You: Package Delivery Via Drone. from https://www.wsj.com/articles/coming-soon-to-a-front-porch-near-you-package-delivery-via-drone-1520798822

Rahman, M. M., Lesch, M. F., Horrey, W. J., & Strawderman, L. (2017). Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems. *Accident Analysis and Prevention, 108*, 361-373. doi: 10.1016/j.aap.2017.09.011

Ritu, A., & Jayesh, P. (1998). A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology. *Information systems research*(2), 204.

Rogers, E. M. (1983). *Diffusion of innovations*: New York : Free Press ; London : Collier Macmillan, c1983.
3rd ed.

Rosaline, S., & Wesley, J. R. (2017). Factors Affecting Students' Adoption of ICT Tools in Higher Education Institutions: An Indian Context. *International Journal of Information and Communication Technology Education, 13*(2), 82-94.

Sánchez-Mena, A. a. s. u. e., Martí-Parreño, J. j. m. u. e., & Aldás-Manzano, J. j. a. u. e. (2017). The Effect of Age on Teachers' Intention to Use Educational Video Games: A TAM Approach. *Electronic Journal of e-Learning, 15*(4), 355-365.

Shana, Z. z. y. c., & Abulibdeh, E. S. e. g. c. (2017). Cloud Computing Issues for Higher Education: Theory of Acceptance Model. *International Journal of Emerging Technologies in Learning, 12*(11), 168-184. doi: 10.3991/ijet.v12.i11.7473

Sim, L. L., & Koi, S. M. (2002). Singapore's Internet shoppers and their impact on traditional shopping patterns. *Journal of Retailing and Consumer Services, 9*(2), 115-124.

Thakur, R., Angriawan, A., & Summey, J. H. (2016). Technological opinion leadership: The role of personal innovativeness, gadget love, and technological innovativeness. *Journal of Business Research, 69*, 2764-2773. doi: 10.1016/j.jbusres.2015.11.012

Tubaishat, A. (2017). Perceived usefulness and perceived ease of use of electronic health records among nurses: Application of Technology Acceptance Model. *Informatics For Health & Social Care*, 1-11. doi: 10.1080/17538157.2017.1363761

Vannoy, S. A., & Palvia, P. (2010). The Social Influence Model of Technology Adoption. *Communications of the ACM, 53*(6), 149-153. doi: 10.1145/1743546.1743585

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

Wang, Y., Hsiao, S.-H., Yang, Z., & Hajli, N. (2016). The impact of sellers' social influence on the co-creation of innovation with customers and brand awareness in online communities. *Industrial Marketing Management, 54*, 56-70. doi: 10.1016/j.indmarman.2015.12.008

Whittaker, S. (2018). Alibaba's Ele.me Starts Food Delivery by Drone in Shanghai First. Retrieved August 6, 2018, from https://dronebelow.com/2018/05/30/world-first-alibaba-eleme-food-delivery-drone-shanghai/

Yoon, C., Jeong, C., & Rolland, E. (2015). Understanding individual adoption of mobile instant messaging: a multiple perspectives approach. *Information Technology & Management, 16*(2), 139-151. doi: 10.1007/s10799-014-0202-4

Editor's Note:

This paper was selected for inclusion in the journal as the CONISAR 2018 Best Paper The acceptance rate is typically 2% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2018.

Appendix.1: Constructs and Items

Construct	Item	Source
Opinion Passing (OP)	When I receive product related information or opinion from a friend, I will pass it along to my other contacts on social networks (0.893) On social networks, I often influence my contacts' opinions about products (0.735) I like to pass along interesting information about products from one group of my contacts on my 'friends' list to another on social networks (0.855)	(Chu & Kim, 2011b)
Perceived Ease of Use (PEOU)	It is easy for me to become skillful in using drone delivery service (0.847) I find drone delivery service easy to use (0.915) I find it easy to use drone delivery service to do what I want it to do (0.877) Learning to use drone delivery service is easy for me (0.869)	(Davis, Bagozzi, & Warshaw, 1989)
Perceived Usefulness (PU)	Using drone delivery service enhances my daily productivity (0.838) I find drone delivery service useful in my daily activities (0.852) I believe drone technology would make my life easier (0.829) I believe drones can deliver packages to me faster than other forms of transportation (0.704)	(Davis et al., 1989)
Personal Innovativeness (PI)	I often try new brands before my friends and neighbors do (0.894) When I see a new brand on the shelf, I often buy it to see what it is like (0.893)	(Sim & Koi, 2002)
Behavioral Intention (BI)	I intend to use a drone delivery service in the next months (0.850) I predict I would use a drone delivery service in the next months (0.864) I will try to use a drone delivery service in my daily life (0.954)	(Venkatesh, Morris, Davis, & Davis, 2003)

The use of Snap Length in Lossy Network Traffic Compression for Network Intrusion Detection Applications

Sidney C. Smith
Sidney.c.smith24.civ@mail.mil
Computational Information Sciences Directorate
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD 21005, U.S.A

Robert J. Hammell II
rhammell@towson.edu
Department of Computer and Information Sciences
Towson University
Towson, MD 21252, U.S.A

Abstract

In distributed network intrusion applications, it is necessary to transmit data from the remote sensors to the central analysis systems (CAS). Transmitting all the data captured by the sensor would place an unacceptable demand on the bandwidth available to the site. Most applications address this problem by sending only alerts or summaries; however, these alone do not always provide the analyst with enough information to truly understand what is happening on the network. Since lossless compression techniques alone are not sufficient to address the bandwidth demand, applications that send raw traffic to the CAS for analysis must employ some form of lossy compression. This lossy compression may take the form or dropping entire sessions, packets, or portions of packets. In this paper we explore impact of compressing network traffic by dropping portions of packets. This is accomplished by truncating packets through adjusting the snap length.

Keywords: compression, network intrusion detection, snap length, Snort, Tcpdump

1. INTRODUCTION

Distributed Network Intrusion Detection Systems (NIDS) allows a relatively small number of highly trained analysts to monitor a much larger number of sites; however, they require information to be transmitted from the remote sensor to the central analysis system (CAS) as pictured in Figure 1. Unless an expensive dedicated NIDS network is employed, this transmission must use the same channels that the site uses to conduct their daily business. This makes it important to reduce the amount of information transmitted back to the CAS to minimize the impact that the NIDS has on daily operations as much as practical.

Smith and Hammell (2017) proposed that it should be possible to create a lossy compression tool using anomaly detection techniques to rate each session and a modification of the Kelly criterion (Kelly, 1956) to select how much traffic from each session to return as seen in Figure 2.

Once the determination of how much traffic to return is made, it is necessary to understand the best way to reduce that traffic. One could carve entire sessions out of the network traffic as Long and Morgan did. (2007) One could drop individual packets as Smith, Hammell, and Neyens did. (2017) Or one could truncate packets as Long did with the “snapper” tool. (2004) This research will consider the implications of the last method adjusting the snap length which truncates packets to achieve lossy compression.

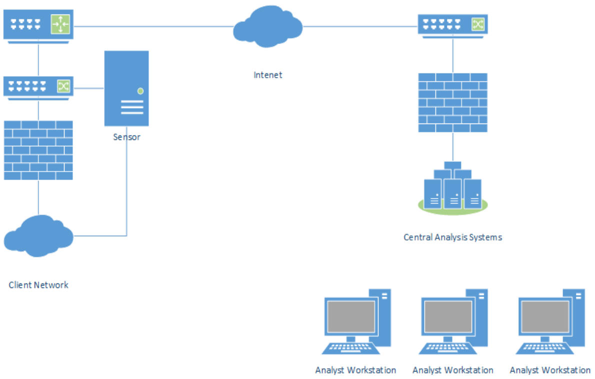


Figure 1. Distributed network intrusion detection

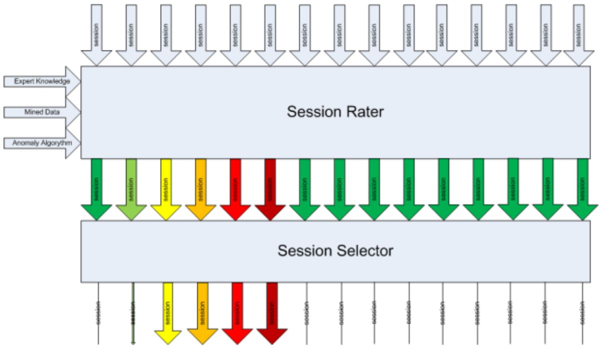


Figure 2. Kelly compressor

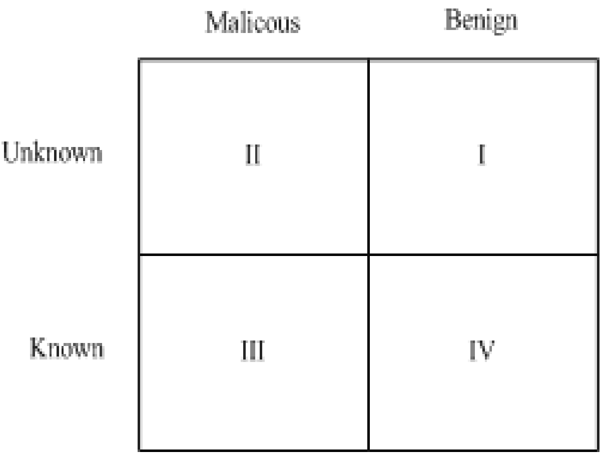
The remainder of this paper is organized into the following sections: Section 2 provides background, Section 3 outlines the approach chosen to address this problem, Section 4 presents our results, and finally, Section 5 provides a conclusion and discussion of future work.

2. BACKGROUND

One popular strategy for implementing a distributed NIDS is to do all of the intrusion detection on the sensor and send only alerts or logs to the CAS. (Roesch, 1999) (Paxson, 1999) A second strategy might be to use lossless compression to reduce the size of the data returned to the CAS. A third strategy is to implement some form of lossy compression algorithm to send back relevant portions of traffic.

There are three problems with the first strategy. The first is that it has the potential to over burden the sensor's central processing unit (CPU) and introduce packet loss. Smith *et al.* discovered that the impact of packet loss can sometimes be quite severe for even small rates of packet loss. (2016a) (2016b) The second problem is that the alerts by themselves often do not contain enough information to determine whether the attack was successful. The third problem is that these systems are most often implemented with signature-based intrusion detection engines. Signature-based systems may be tuned to produce few false positives; however, they are ineffective at detecting zero-day and advanced persistent threats. (Kremmerer & Giovanni, 2002)

The problem with the second strategy is that lossless compression alone simply is not capable of reducing the amount of traffic enough. Using GNU Zip to compress the 2009 Cyber Defense Exercise dataset provides a compression ratio of 2:1. (Smith, Neyens, & Hammell, 2017) Compression ratios of better than 10:1 are required to minimize the impact of NIDS on day-to-day operations.



	Malicous	Benign
Unknown	II	I
Known	III	IV

Figure 3. Network traffic composition

The third strategy is to use lossy compression to provide a solution. Network traffic may be considered to be composed of sessions that span spectrums from known to unknown and malicious to benign as illustrated in Figure 3. Quadrant III, the known malicious quadrant, is the domain of intrusion prevention systems as described by Ierace, Urrautia, and Bassett (Ierace, Urrutia, & Bassett, 2005). This research is most interested in quadrant II, the unknown malicious quadrant, because that is the quadrant where evidence of zero-day and advanced persistent threat attacks will be found. In 2004, Kerry Long described the Interrogator Intrusion Detection System Architecture (2004). In this architecture, remotely deployed sensors, known as Gators, collect network traffic and transmit a subset of the traffic to the analysis level. Interrogator employs "a dynamic network traffic selection algorithm called Snapper'". (2004). Long and Morgan describe how they used data mining to discover known benign traffic that they excluded from the data transmitted back to the analysis servers (2007).

3. APPROACH

Tcpdump (Jacobson, Leres, & McCanne, 2017) is a very popular network capture tool. The data format use by tcpdump to store the captured network traffic has become the *de facto* standard format for network capture data. Snort (Roesch, 1999) is a very popular signature based network intrusion detection tool. Both tcpdump and snort support an option to set the snap length. This option is used to set the maximum size of any packet collected. Packets larger than the snap length will be truncated. It is primarily used to improve efficiency when the maximum transmission unit of the network is known. One might suspect that conducting several iterations of these experiments would be as simple as repeatedly executing one of the commands seen in Figure 4.

```
$ tcpdump -r ${DATASET} -s
${SNAPLEN} \ > -w - |
> snort -N -c ${RULESET} -k none -r - -l
.

$ snort -r ${DATASET} -k none \
> –c ${RULESET}\
> --snaplength ${SNAPLEN} -l .
```

Figure 4. Command line

The authors of tcpdump pulled the packet capture routines out of tcpdump into a standard alone library known as the packet capture library or libpcap (Jacobson, Leres, & McCanne, 2015). Today both tcpdump and snort leverage this library. It turns out that both tcpdump and snort implement snap length by passing the option to libpcap (Jacobson, Leres, & McCanne, 2015) which only implements this feature for live traffic capture. To use the snap length features of either tcpdump or snort we needed to leverage an experimental environment similar to the one seen in Figure 5. Replaying a dataset several times at some multiple of the original speed small enough to ensure that packets are not lost in transmission requires a significant amount of time. We conducted this experiment only twice to gain a baseline. We developed a tool that will implement the snapping in software. We tested it against the baseline we established using the experimental environment. The validated tool was then be used to quickly test of impact of snap length across multiple datasets.

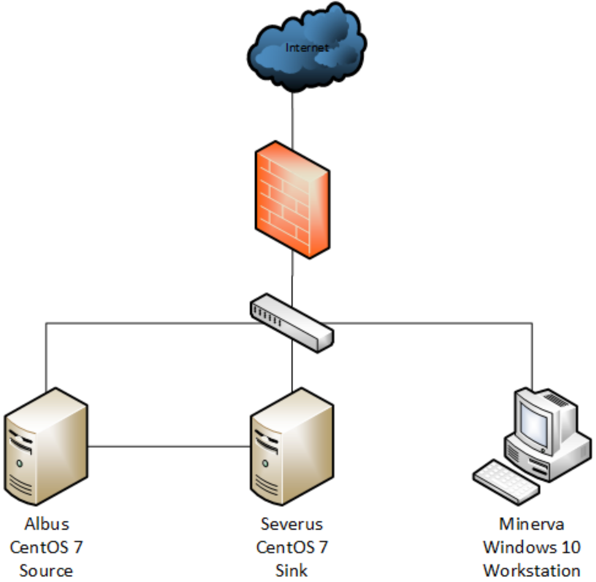


Figure 5. Experimental environment

Experimental Baseline

The experimental environment seen in Figure 5 consists of two workstation class systems with Gigabit Ethernet cards directly connected to each other. We did not configure the interfaces of these cards to prevent any extraneous traffic from appearing on this network. Albus is designated as the source, and tcpreplay (Turner & Bing, 2013) was used to replay the traffic. Severus was designated as the sink and tcpdump and snort were used to collect and analyze the traffic. Several iterations were conducted changing the snap length. The snap length used, the percentage for the original size of the data set, and the number of alerts are collected and plotted.

Snapping Tool

There are three length fields in libpcap files. The first is a global length field. We set this field by passing the new snap length to the pcap_open_dead() function when we created the pcap_t structure which we passed to pcap_dump_open(). The other two length fields are contained in the pcap_pkthdr structure. These are caplen and len. The len field is the original length of the packet, and the caplen field is the number of bytes actually stored in the libpcap file.

In previous research, we developed the pcapcat program (Smith S. C., 2013). This program simply takes the list of libpcap file names on the command line and reads each file writing it to standard output. This provided a necessary first step for any tools which will manipulate libpcap files, and a convenient method to join several libpcap files into one file. We took this program and added a snap length option. Implementing this option involved setting the global snap length when we created the output handle, and setting the caplen value of the pcap_pkthdr.

Datasets

In the following section we provide a brief summary of the various datasets that were used in our experiment. Table 1 provides a summary of the duration and packet count for each of these datasets.

- DARPA Datasets

As part of their evaluation of intrusion detection systems, Lippman *et al.* created a dataset of synthetic network traffic (2000). We used the small sample dataset which was provided before the experiment to give the participants examples of the data that they would be provided in the evaluation. This dataset is about 10 min long and was used to validate that the tools were working correctly. They also created the four hour dataset. This dataset was used to evaluate the efficiency of the intrusion detection techniques. We used it because it is large enough to provide a good baseline but small enough to allow us to conduct our experiment in a reasonable amount of time. We used it to compare the results of using the snap length options of tcpdump and snort to our snapping tool. Finally we used the testing data from Wednesday and Friday of week 2. We selected these two days because Wednesday contains the smallest number of alerts and Friday contains the largest number of alerts.

- Cyber Defense Exercise 2009

In 2009 the National Security Agency/Central Security Service (NSA/CSS) conducted an exercise pitting teams from the military academies of the United States and Canada against teams of professional network specialists to see who best defended their network. Data from this exercise was captured and used by Sangster *et al.* in his efforts to generate labeled datasets (2009). Two network traffic sensors were employed in the exercise: gator-usama010 and gator-usama020. We used the pcapcat program to consolidate the individual hours of for network traffic collected by each sensor into two libpcap files.

- Mid-Atlantic Collegiate Cyber Defense Competition

Based upon the pattern of the Cyber Defense Exercises, a group of industry academics created the collegiate cyber defense competitions (Carlin, Manson, & Zhu, 2010). We used the network capture data for the Mid-Atlantic Collegiate Cyber Defense Competitions from 2010 and 2011 which is available from:
https://www.netresec.com/?page=MACCDC.

- Real World

We were able to collect real world network traffic from the top level architecture of one site of a research laboratory on the Defense Research Engineering Network.

Table 1. Datasets

Name	Seconds	Packets
DARPA98ss	624	14,523
DARPA984h	19,258	233,428
DARPATestW2Wed	86,400	2,026,473
DARPATestW2Fri	90,432	2,177,646
CDX09_usama010	378,000	5,218,144
CDX09_usama020	345,600	42,293,657
MACCDC2010	275,666	264,973,151
MACCDC2011	165,243	134,465,786
Live Data	138,895	2,256,633,016

4. RESULTS

First we will review the results of our validation exercises. Then we will present the results of our validated snapping tool.

Validation in the Experimental Environment

The first step in the process is to ensure that our snapping tool provides the same results as we obtained using tcpdump. To do this we will take the DARPA98 Four Hour and DAPRA98 Small Sample datasets and replay them in the experimental environment seen in Figure 5. We automated 30 iterations of Albus replaying the traffic using the tcpreplay tool while Severus used tcpdump with using snap lengths ranging from 1542-42. These captured files were then analyzed with snort.

- DARPA 98 Four Hour

To ensure that this experiment using the four hour dataset completed in a reasonable amount of time, we replayed the traffic at 10 times the original speed. In Figure 6, we have plotted the percentage of the original file size using triangles. We have plotted the alert loss rate (ALR) as a percentage in circles. We have also plotted the packet loss rate (PLR) as a percentage in crosses. In Figure 7, we have plotted the results of using the snapping tool on the same dataset. Comparing the graphs, we find that the relationship between the ALR and the snap length for the experimental environment and the snapping tool is very similar. The differences between the relationship between the compression and the snap length between the two experiments may be attributed to the PLR.

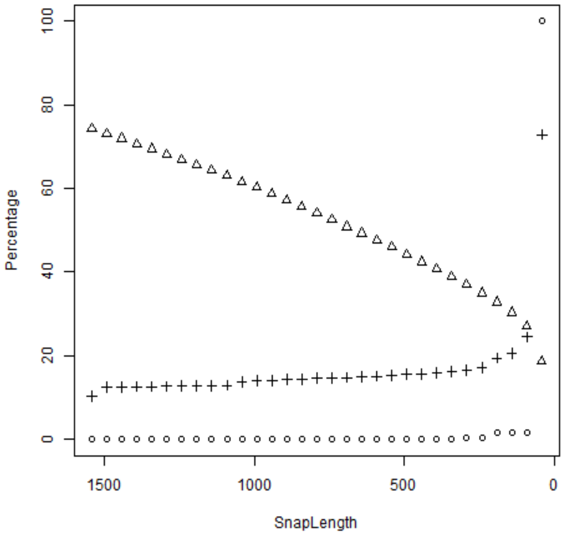


Figure 6. Results of using tcpdump to snap the packets of the DARPA 98 Four Hour dataset in the experimental environment

- DARPA 98 Small Sample

To further assure that our snapping tool is performing correctly, we repeated the experiment with the DAPRA 98 Small Sample dataset. This dataset is about 10 min long allowing us to replay the traffic at the original speed and still complete the experiment in a reasonable amount of time. In Figure 8, we have plotted the percentage of the original file size using triangles. We have plotted the ALR as a percentage in circles. We have also plotted the PLR as a percentage in crosses. One thing of note is that packet loss is completely from packets that snort has discarded. Since these packets were discarded and not dropped, they are not subtracted from the size when the percentage of the original size is computed. In Figure 9, we have plotted the results of using the snapping tool on the same dataset. Again the results are very similar and from this we conclude that our snapping tool is truncating the packets in the

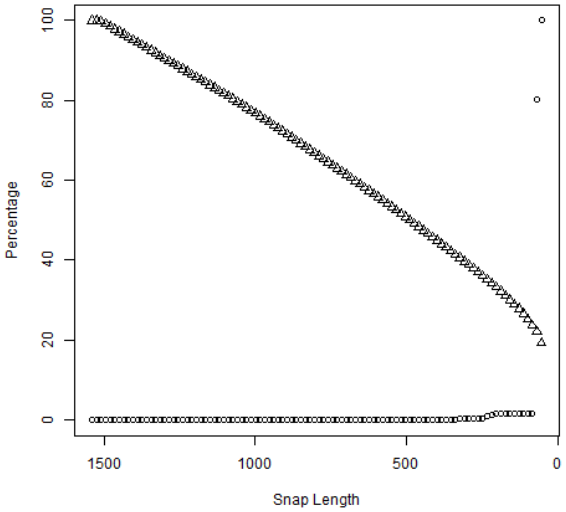


Figure 7. Snap length verses the ALR and Compression of the DARPA 98 Four Hour datasets using the snapping tool

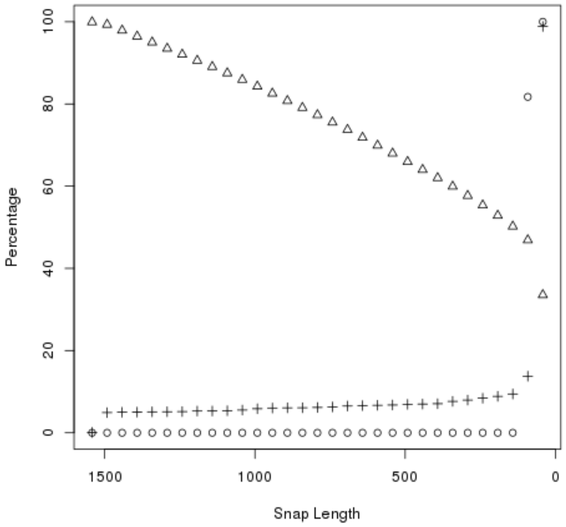


Figure 8. Results of using tcpdump to snap the packets of the DARPA 98 Small Sample dataset in the experimental environment

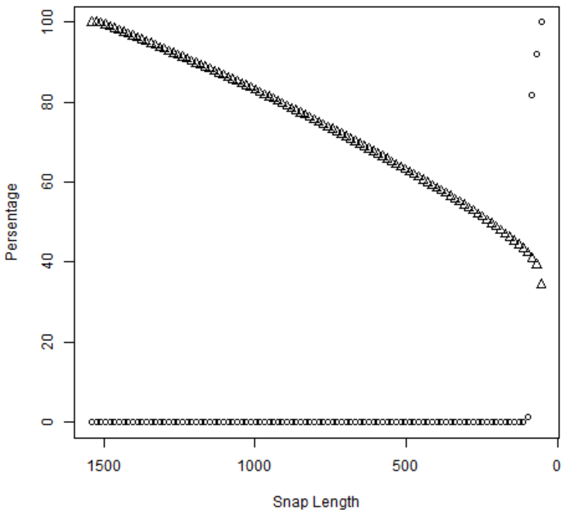


Figure 9. Snap Length verses the ALR and Compression of the DARPA 98 Small Sample dataset using the snapping tool

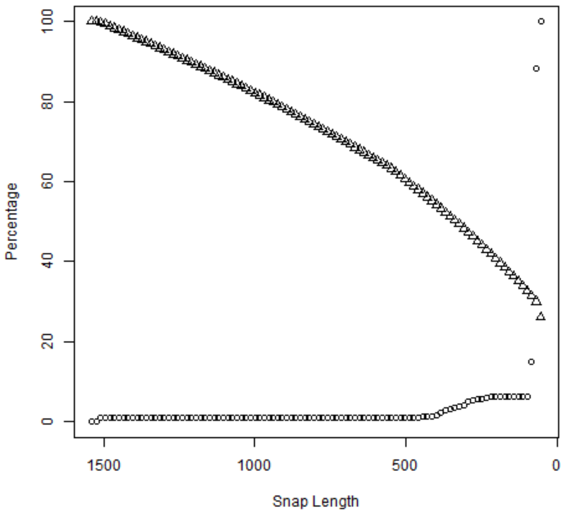


Figure 10. Snap length verses the ALR and Compression of the DARPA 98 testing week 2 day 3 datasets using the snapping tool

Experimentation with the Snapping Tool

Having validated that the snapping tool performs in the same manner as the snap length option to tcpdump, we may forgo further use of the experimental environment. We created a shell script to automate the snapping and analysis of the remaining datasets.

- DARPA 98 Testing Week 2 Wednesday

In Figure 10 and Figure 11 we see the results of using our snapping tool on the 2 days we selected from the DARPA 98 Testing dataset. Notice that for each of these 2 datasets, we are able to gain a significant amount of compression by snapping packets with little or no increase in the ALR. The same may be said for the datasets that we used to validate the snapping tool.

- Cyber Defense Exercise

In Figure 12 and Figure 13 we see the results of using our snapping tool on the Cyber Defense Exercise 2009 datasets. These graphs show a much earlier rise in ALR.

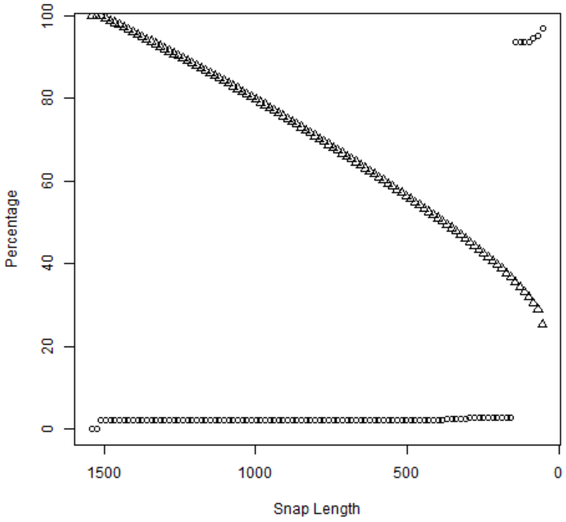


Figure 11. Snap length verses the ALR and Compression of the DARPA 98 testing week 2 day 6 datasets using the snapping tool

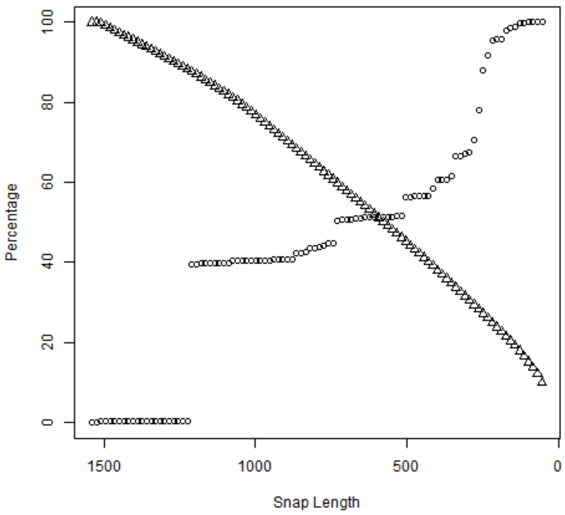


Figure 12. Snap Length verses the ALR and Compression for CDX2009 usama010

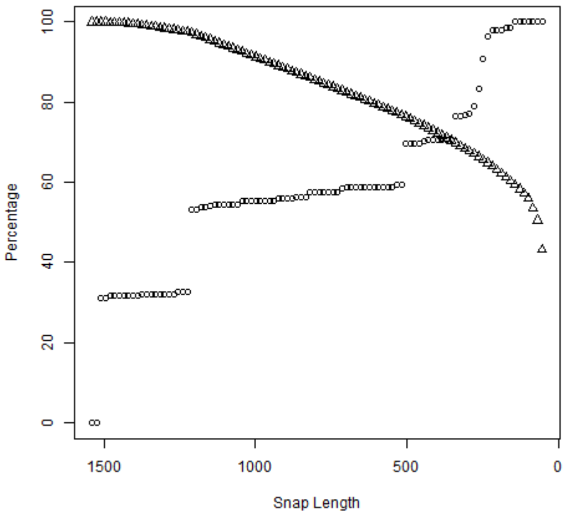


Figure 13. Snap Length verses the ALR and Compression for CDX2009 usama020

- Mid-Atlantic Collegiate Cyber Defense Competition Datasets

In Figure 14 and Figure 15 we see the results of applying our snapping tool to the Mid-Atlantic Collegiate Cyber Defense Competitions of 2010 and 2011. With the 2010 data we see more dramatic rise in ALR, but not as dramatic as the rise we saw in the CDX data. With the 2011 data we see that it is possible for the snapping process to create alerts in the data that did not exist previously. The creation of false positive alerts was not one of the anticipated outcomes.

- Live Data

The results of the experiment using live data may be seen in Figure 16. It would appear that that data set had a small number of very large packets, but once the snap length reached about 1500 the size started falling steadily, but the ALR raised quickly only to level off.

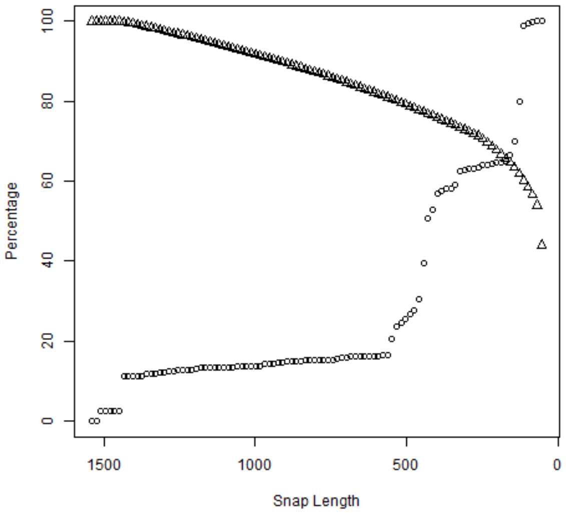


Figure 14. Snap length verses the ALR and Compression of the MACCDC 2010 dataset using the snapping tool

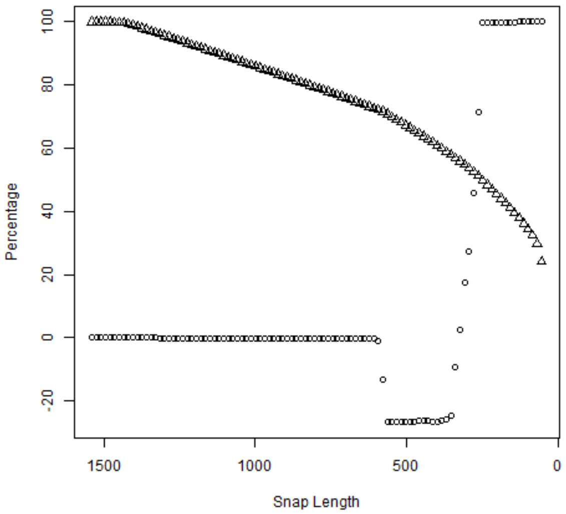


Figure 15. Snap length verses the ALR and Compression of the MACCDC 2011 dataset using the snapping tool

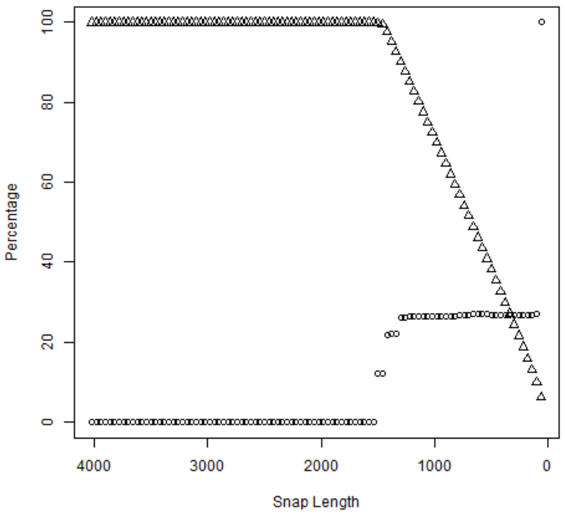


Figure 16. . Snap length verses the ALR and Compression for live data captured from an operational network using the snapping tool

5. CONCLUSIONS

Looking at our results from the DARPA datasets it would appear that employing snap length as a compression tool has the potential to reduce the size of the traffic that must be transmitted from the sensor to the CAS. Our results from the Cyber Defense Exercise data indicate that this might be a very dangerous technique as the ALR rises rapidly with the decrease in snap length. Our results from the Mid-Atlantic Collegiate Cyber Defense Competition and live data seem to occupy the middle ground with the caveat that the technique may introduce false positive alerts.

It might appear that the malicious content in new traffic is deeper in the packet than malicious content in older traffic; however, an examination of the traffic reveals that this is not the case. In every packet that we examined that triggered an alert in the original data, but did not trigger an alert in the abridged data, the string in the rule existed in the abridged packet. The explanation for our results lies in the number of discarded packets observed in the experiment using the DARPA 98 Small Sample dataset in our experimental environment. Even though we used the option to instruct snort not to validate the checksums, it is discarding truncated packets. We are not seeing that the malicious nature is deeper into the packets in new traffic. We are seeing that packets with a malicious nature are larger in newer traffic, and a detection tool that does not discard truncated packets would have detected the malicious traffic. Also analysts reviewing the truncated traffic based upon alerts generated by snort seeing the unabridged traffic would be able to use the truncated traffic to conduct their analysis.

Although tools like snort are best run on the sensor where they may have a full view of the network traffic, there is value in running tools like this on the CAS where the size of the ruleset will not negatively impact of the amount of traffic which may be collected. In future work it will be necessary to explore other methods of lossy compression that might not have the same issues. Alternatively snort could be altered to accept truncated packets or a similar tool could be developed that would accept truncated packets.

6. REFERENCES

Carlin, A., Manson, D. P., & Zhu, J. (2010). Developing the Cyber Defenders of Tomorrow with Regional Collegiate Cyber Defense Competitions (CCDC). *Information Systems Education Journal*, 3-10.

Ierace, N., Urrutia, C., & Bassett, R. (2005). Intrusion Prevention Systems. *Ubiquity*, 2-2.

Jacobson, V., Leres, C., & McCanne, S. (2015, March 8). *PCAP -- packet capture library*. Retrieved from Tcpdump/Libpcap: http://www.tcpdump.org/manpages/pcap.3pcap.1.html

Jacobson, V., Leres, C., & McCanne, S. (2017, February 2). *tcpdump -- dump traffic on a network*. Retrieved from TCPDUMP & LIBPCAP: http://www.tcpdump.org/manpages/tcpdump.1.html

Kelly, J. L. (1956). A new interpretation of information rate. *Information Theory, IRE Transactiosn on*, 185-189.

Kremmerer, R. A., & Giovanni, V. (2002). Intrusion detection: a brief history and overview (supplement to Computer magazine). *Computer*, 27-30.

Lippmann, R. P., Fried, D. J., Graf, I., Haines, J. W., Kendall, K. R., McClung, D., . . . Zissman, M. A. (2000). Evaluating intrusion detection systems: the 1998 {DARPA} off-line intrusion detection evaluation. *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings* (pp. 12-26). Hilton Head, SC: IEEE.

Long, K. S. (2004). *Catching the Cyber Spy: ARL's Interrogator.* Aberdeen Proving Ground: Army Research Laboratory.

Long, K. S., & Morgan, J. B. (2007). *Using data mining to improve the efficiency of intrusion detection analsysis.* Army Research Laboratory. Aberdeen Proving Ground (MD): Army Research Laboratory.

Paxson, V. (1999). Bro: a system for detecting network intruders in real-time. *Computer Networks*, 2435-2463.

Roesch, M. (1999). Snort: lightweight intrusion detection for networks. *Proceedings of the 13th System Administration Conference (LISA '99)* (pp. 229-238). Seattle, WA: USENIX.

Sangster, B., O'Conner, T., Cook, T., Franelli, R., Dean, E., Adams, W. J., . . . Conti, G. (2009). Toward instrumenting network warfare competitions to generate labeled datasets. *Proc. of the 2nd Workshop on Cyber Security Experimentation and Test CSET09.* Montreal Canada.

Smith, S. C. (2013, May). The effect of packet loss on Network Intrusion Detection. *Towson University Institutional Repository*. Towson, MD, USA: Towson University.

Smith, S. C., & Hammell, R. J. (2017, Aug). Proposal for Kelly Criterion-Inspired Lossy Network Compression for Network Intrusion Applications. *Journal of Information Systems Applied Research, 10*(2), 43-51.

Smith, S. C., Hammell, R. J., Wong, K. W., & Carlos, J. M. (2016). An Experimental Exploration of the Impact of Host-Level Packet Loss on Network Intrusion Detection. *Cybersecurity Symposium (CYBERSEC)* (pp. 13-19). IEEE.

Smith, S. C., Hammell, R. J., Wong, K. W., & Carlos, J. M. (2016). An experimental exploration of the impact of multi-level packet loss on network intrusion detection. *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)* (pp. 23-30). Towson, MD: IEEE.

Smith, S. C., Neyens, S. R., & Hammell, R. J. (2017). The use of Entropy in Lossy Network Traffic Compression for Network Intrusion Detection Applications. *Proceedings of the 12th International Conference on Cyber Warfare and Security {ICCWS} 2017* (pp. 352-360). Reading (UK): Academic Conferences and Publishing International Limited.

Smith, S. C., Neyens, S. R., & Hammell, R. J. (2017). The use of Entropy in Lossy Network Traffic Compression for Network Intrusion Detection Applications. *Proceedings of the 12th International Conference on Cyber Warfare and Security ICCWS* (pp. 352-360). Reading (UK): Academic Conferences and Publishing International Limited.

Turner, A., & Bing, M. (2013, December 14). *Tcpreplay: Pcap editing and replay tools for *nix*. Retrieved from Syn Fin dot Net: http://tcpreplay.synfin.net

Adversarial Machine Learning for Cyber Security

Michael J. De Lucia [a,b]
Michael.j.delucia2.civ@mail.mil

Chase Cotton [b]
ccotton@udel.edu

[a] U.S. Army Research Laboratory (ARL)
Aberdeen Proving Ground, MD 21005

[b] Electrical and Computing Engineering Department
University of Delaware
Newark, DE 19716

Abstract

The security of machine learning, also referred to as Adversarial Machine Learning (AML) has come to the forefront in machine learning and is not well understood in the application to the cyber security area. AML has been largely applied to image classification but has been limited in application to the cyber security area. One of the most fundamental components of machine learning, is the features. The disparate features of the cyber security area vary and are different than in image classification. To understand the features of the cyber security area, traffic classification is selected as a use case to focus on. Additionally, we present an example of cyber security AML of a network scanning classifier. A background on AML attack types, Adversarial Knowledge, and Image Classification features is given first. Next a discussion of the Cyber security traffic analysis features and AML of the cyber security area is given. We propose the disparate features of the cyber security area, augmented with ensemble learning could lead to a defense against AML. Future research is proposed for experimentation of AML with a subset of the cyber features discussed and the development of a defense against AML.

Keywords: Adversarial Machine Learning, Cyber Security, Traffic Analysis, Features, Machine Learning

1. INTRODUCTION

The security of machine learning, also referred to as Adversarial Machine Learning (AML) has come to the forefront in machine learning and is not well understood within a cyber security context. Machine Learning has become integrated into many different technologies to include cyber security (i.e. Intrusion Detection Systems (IDS), traffic analysis, malware and network scanning detection). Adversaries will attempt to circumvent and negatively affect the classification decisions, where machine learning has been employed for protection (Laskov & Lippmann, 2010).

AML has largely been applied to image classification and spam filtering with limited understanding within cyber security (Laskov & Lippmann, 2010). AML has also been focused on Deep Neural Networks (DNN) but has also been applied to traditional machine learning algorithms such as Support Vector Machines (SVM) (Papernot, McDaniel, Goodfellow, Jha, Celik, & Swami, 2017). Thus far there has been a limited knowledge of AML to cyber security. The specific

cyber security area that will be focused on will be AML of SVM (machine learning) traffic classification and analysis methods in addition to a network scanning detection scenario.
One of the fundamental components of the employment of machine learning methods to a specific technology area is feature engineering and representation. Features employed within machine learning based cyber security network detection classifier implementations vary greatly and are developed and engineered based on network traffic characteristics. The techniques that an adversary can use to perturb network traffic such that it is misclassified by the defender's IDS or traffic classification varies greatly depending on the machine learning approach and features implemented in the IDS or traffic classification.

We propose, a greater understanding of the importance of features and inclusion of multiple disparate features to improve the defense against AML for cyber security (traffic analysis). First, a background on the attack types, levels of adversarial knowledge, image classification features and AML will be given. Next, a discussion of features of the cyber security area and AML in cyber security, followed by an investigation and results of conducting AML on a network scanning detection classifier. Lastly a conclusion and discussion of future work will be presented.

2. BACKGROUND

AML Attack Types

In AML, there are two different types of attacks an adversary could perform; Evasion and Poisoning attacks (Muñoz-González, Biggio, Demontis, Paudice, Wongrassamee, Lupu, & Roli, 2017). An evasion attack occurs when an adversary perturbs a sample at test (detection) time to cause misclassification. A poisoning attack occurs when an attacker inserts mislabeled bad or perturbed data into the training samples. The focus of this paper will be on evasion attacks.

Adversarial Knowledge

There are varying levels of an adversary's knowledge of a system, which can be leveraged as attack models (Biggio, Corona, Maiorca, Nelson, Šrndić, Laskov, & Roli, 2013). The varying levels of knowledge include Perfect (Complete Knowledge), Limited, and Zero. Perfect level knowledge is defined as the adversary having knowledge of the feature space, type of classifier, and the trained model (Biggio et al., 2013). In the limited knowledge case, the adversary knows feature representation (features included) and the type of classifier, but not the trained model (Biggio et al., 2013). Lastly, zero knowledge is when the adversary does not know any of the details (features, type of classifier, or trained model) of the machine learning system. An adversary's knowledge levels of Perfect, Limited, and Zero are analogous respectively with the traditional cyber security terms of White-box, Grey-box, and Black-box. The terms White-box, Grey-box, and Black-box will be used throughout this work to refer to the adversary's level of knowledge of the machine learning classifier.

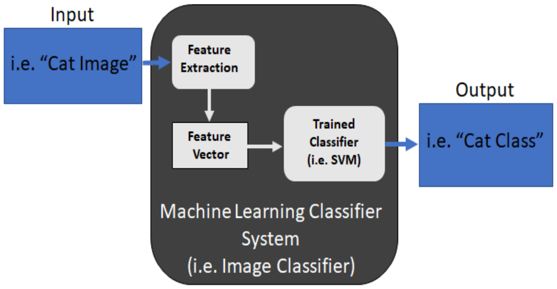


Figure 17- Machine Learning Classifier System View

Recall in the Black-box instance, an attacker has zero knowledge of the machine learning classifier. Therefore, an attacker may only have access to the input and output of the machine learning classifier. In Figure 1, it can be observed that the feature extraction and the classification decision occur within the machine learning classifier's system boundaries. Therefore, the features are unknown to the adversary. As Figure 1 depicts the Machine Learning Classifier System takes an input of the sample instance which is to be classified and the output is the class assigned. As shown in Figure 1 the adversary provides an input image of a cat to the Image Classifier and receives an output of the "Cat Class".

Many machine learning classifiers are open systems, allowing the adversary to view both the inputs (i.e. image) presented to the classifier and the resulting output class assigned (i.e. "Cat", "Not Cat"). However, there are cases where an adversary will have a partial view or no view of the input or output. An example, where an adversary will have no view of the input or output is a machine learning classifier which is executed in an isolated offline environment (not accessible). In a partial view, where only the input can be viewed, the adversary may need to infer the output class based on outside observations or knowledge. A further discussion of a partial view will be provided in a later section of AML for cyber security.

Image Classification AML

To understand transferability of AML from image classification to cyber security, we will give a brief background on the features within image classification. In image classification, an image is composed of a matrix of pixels and channels (e.g. 3 RGB channels), each representing the pixel (i.e. color) intensification (0-255). The pixels are directly extracted from an image as a feature. Additionally, the relationship between neighboring pixels can be extracted as features by using a combination of image gradient, edge detection, orientation, spatial cues, smoothing, and normalization (Zheng & Casari, 2018).

In image classification, AML is the perturbation of an image by adding noise to cause misclassification (Papernot, McDaniel, Jha, Fredrikson, Celik, & Swami, 2016). The perturbation of the image by an adversary must be applied meticulously to cause misclassification by the machine learning classifier, while still being correctly classified by the human eye (Papernot et al., 2016). AML in image classification has been primarily focused on DNN but has been demonstrated to transfer to traditional machine learning methods such as Support Vector Machines (SVM) (Papernot et al., 2017).

3. CYBER SECURITY FEATURES

A fundamental component of the machine learning development process is feature engineering. Feature engineering is defined as the process of transforming raw data into features to better represent the relationship between classes to improve machine learning performance (Susarla & Ozdemir, 2018). The features within cyber security are extracted differently compared to image classification. The features within the SVM based traffic analysis cyber security, are not always based solely on the bits within the network packet. They may be either based on each network packet or the network traffic flow.

There are many options for feature extraction directly from a network packet. Examples of features directly extracted from the network packet include the nested protocol headers or sub-fields or the packet payload (content). Inspection of the payload is often referred to as Deep Packet Inspection (DPI) or Payload based Classification (Kim, Claffy, Fomenkov, Barman, Faloutsos, & Lee, 2008).

An alternative option for feature extraction includes characteristics of a network traffic flow. A network traffic flow is often a group of network packets for a specific conversation between two endpoints. There are many characteristics of a network flow such as connection tuples (source and destination IP and Ports), inter-arrival times, sequence of packet sizes, Transport Layer Security (TLS) record sizes, offered TLS Cipher Suites, and the total bytes transferred in each direction.

Network Packet Features

An example which creates features from the packet payload is the Extremely Lightweight Intrusion Detection (ELiDe) System (Chang, Harang, & Payer, 2013). ELiDE builds an n-gram representation of the bits contained within the network packet payload to create the features for input into a binary linear classifier (Chang et al., 2013). While, the motivation for ELiDE was an Intrusion Detection System (IDS), it could also be used for fingerprinting of the payload for traffic analysis. Similarly, to image classification, an n-gram representation of the bytes contained in the network packet payload are directly extracted from the network packet as features.

However, this approach could be easily influenced by an adversary by encrypting the packet payload, thereby hiding any malicious activities. Therefore, the addition of encryption to the traffic payload protects and hides the malicious activities, resulting in an inability to perform DPI (Dainotti et al., 2012). The inability to perform DPI on an encrypted payload can be attributed to a different output produced each time since a new symmetric key is generated for each session established. For example, in the Transport Layer Security (TLS), which leverages encryption to protect communications, a handshake occurs first, during which a new symmetric key is generated and securely shared between client and server (Dierks & Rescorla, 2008).

As a result, this would allow an adversary to influence the machine learning classifier to cause a misclassification of malicious traffic as benign. This misclassification of an encrypted payload is caused by the fact of the payload n-gram representation features learned during training, not matching the extracted features at test (detection) time. Previously, encryption of the packet payload in of itself could have been an indicator of malicious activity or a signature for traffic classification. However, Internet traffic is increasingly becoming encrypted, as of 2016 approximately 30 percent of the top page search results on Google used HTTPS (SSL/TLS) (Meyers, 2016). According to Google Transparency Report on HTTPS encryption in the Web, 95% of traffic across Google's infrastructure is encrypted and 75% of Windows based Chrome users browsed to HTTPS encrypted websites as of June 2018 (Google, 2018). The trend of Internet encrypted traffic is on the rise and will become widespread in the future.

Network Flow Features

An alternative traffic analysis mechanism is to use derived characteristics of the packet or network flow of traffic. In this instance traffic analysis is performed at a flow level which contains a sequence of packets which may be a bi-directional (client and server) or unidirectional (single sided) conversation. There exist several characteristics of a network flow such as the unique connection tuple (Source IP, Destination IP, Source Port, and Destination Port), inter-arrival packet times, unique TCP flags set, protocols used, non-conforming protocol use, frequency of communication, packet or protocol sizes, sequences of packet or protocol sizes exchanged, and domain names leveraged. While, these are a few examples of characteristics, the possibilities of different cyber security features are endless.

Appendix A presents even further cyber security feature examples, which demonstrate 52 features from (Muehlstein, Zion, Bahumi, Kirshenboim, Dubin, Dvir, & Pele, 2017), 19 features from (Anderson, Paul, & McGrew, 2016), 3 features from (Wright, Monrose, & Masson, 2006), and 2 features from (Herrmann, Wendolsky, & Federrath, 2009). The examples in Appendix A is merely a brief taxonomy of cyber security features from four different studies, but still displays a large number of features. Hence, the number of cyber security feature possibilities is massive.

Additional features for input to the machine learning classifier could be extracted and represented from these characteristics such as the mean and standard deviation could be taken over the timing and packet sizes over the traffic flows. Additionally, signatures of non-encrypted payloads carried by standard network methods can also be checked against signatures of known malicious payloads.

Another example is the use of data mining approaches such as the term frequency and inverse document frequency to represent the frequency of TLS record sizes within a conversation (De Lucia, 2018). In this case the characteristic is the term frequency of the TLS record sizes, which then forms the feature vector for each conversation. This single characteristic maps to a medium sized feature space of 32,000 unique possibilities, which results in sparse vectors since not all record sizes are present in every conversation. However, the sequence of TLS record sizes could be represented in a multitude of different ways to create features. For example, a possible alternative representation could be the total number of bytes, weighted average, and standard deviation sent in each direction. For an attacker to perturb their traffic flow to be misclassified as another type of traffic flow (malicious vs benign), they would need to modify the sequence of TLS record sizes being exchanged in each direction to match the pattern of another type of traffic.

Yet, another example is the attribution of TLS encrypted malware to a specific malware family (Anderson, Paul, & McGrew, 2016). Attribution using traffic analysis is performed using 19 different features such as identical TLS parameter use, sequence of packet lengths and times, network flow data, byte distribution, the TLS handshake list of offered cipher-suites, list of advertised extensions, and the public key length (Anderson et al., 2016). These features were also used to differentiate benign from malicious TLS clients (Anderson et al., 2016).

In this traffic analysis method, there are many features directly taken from the characteristics and some which are derived. Again, these characteristics could be represented in many ways to form the features which will be input into the machine learning classifier. For an attacker to cause misclassification of their traffic, they would need to modify many different characteristics. As an example, an adversary could modify the list of cipher-suites offered and extensions supported to match that of another traffic flow. However, the adversary may need to perturb several features to accurately cause misclassification.

4. AML CYBER SECURITY

In AML cyber security traffic, the adversary will perturb the malicious network application (i.e. malware, bot-net communication) traffic to appear as benign. For example, an adversary will perturb their Nmap network scanning traffic to appear as benign to a network scanning detector (machine learning classifier), resulting in a misclassification. However, just as in the image classification, there are constraints which are levied on the perturbation performed by the adversary.

There are many constraints within the cyber security area. Some example constraints include adherence to the respective networking (i.e., TCP, IP, TLS) protocol widely known standard documents (i.e. RFCs), implementation of offered services (i.e. TLS cipher suites offered in a Client Hello message), allowing the successful transmission of the message contents of a bot-net communication, and not negatively impacting the goal of malware contained within the network traffic. The constraints can change based on the objective and implementations chosen by the adversary (malware, bot-net traffic, or TLS client). For example, an adversary performing perturbation of the network traffic must be done within the bounds of the specific network protocol being leveraged (i.e., non-normal window sizes, improper TCP flags set).

Additionally, there is indirectly a human element for a constraint. Traffic analysis by a machine learning classifier may be also augmented with a human analyst. Therefore, the perturbations of the malicious traffic must be performed in a method which would not be noticeable by an experienced network analyst.

AML Perturbation

To cause misclassification, one of the fundamental components which an adversary must perturb is the features which are leveraged by the targeted SVM cyber security classifier. As discussed earlier, an adversary would need to perturb their malicious network traffic to mimic the features of a legitimate traffic flow, to hide their malicious activities. For example, a bot-net developer would need to perturb the bot-net traffic to look like either another bot-net (misattribution) or look like legitimate application traffic. We are assuming the adversary will leverage encryption, which implies that DPI is unusable, resulting in the need to use traffic analysis features. The next two examples are based on the network flow feature examples discussed in section 3.

Recall the first example network flow features discussed was the use of the TLS record sizes as a feature. The adversary would only need to perturb the single feature of the TLS record sizes. The TLS record sizes of the adversary's malicious traffic would need to be perturbed to mimic the sequence and distribution of TLS record sizes from a legitimate network traffic flow. However, this may have a cascading effect in producing a larger number of packets and increase of latency and inter-arrival times. For example, this increase could be attributed to a larger TLS record size resulting in longer processing times at the end nodes and transmission time of the message or malware to be sent. Much thought must be given by the adversary, as to the effects caused by the perturbation. However, this cascading effect could also be a benefit to the defense against AML. The attacker would also have the constraint of having to perturb the TLS record sizes, while still achieving a malicious goal.

Recall the second example network flow features discussed was the list of cipher suites offered, packet lengths, and timing. The adversary would need to perturb many more features of the malicious network traffic to mimic another legitimate network flow. For example, the attacker would need to perturb the list of cipher suites offered, the packet lengths, and the timing among many other features. The difficulty and cost, in terms of time, of mimicking another traffic flow, increases linearly as the number of disparate features increase. Each of the features is a disparate characteristic which must be manipulated to cause misclassification. Additional characteristic perturbations increase adversary implementation time to achieve misclassification. Additionally, perturbing a single feature may have a detrimental unintended effect on another feature.

As an example, if the two features are the TLS record sizes and the number of cipher suites offered, it will require disparate perturbations to the malicious traffic. An adversary would not only need to mimic the TLS record size sequences, but also the offered cipher suites. To mimic the cipher suites, the adversary would need to not only add it to the list, but also implement these cipher suites in the malicious client software. The additional implementation time to achieve these perturbations, indirectly increases the cost to the adversary.

Adversary Knowledge

Recall the Black-box, Grey-box, and White-box model for adversarial knowledge as discussed in section 2. All three of these models hold for the machine learning based cyber security of traffic analysis. However, there are some differences in the Black-box case, which will be expanded on. Recall in the Black-box case the adversary can only view the input and the output classes. However, in the cyber security traffic analysis, only the input is observed and known by the adversary and the output is not known or observed.

For example, let's assume the traffic analysis is being employed in a passive IDS in an enterprise environment. Traditionally, a passive IDS will raise and write alerts to the log file or notify an administrator for identified malicious network traffic. Therefore, the result is only known to the network administrator and not by the adversary.

To augment this example, let's now assume it is an active IDS within an enterprise environment. In this case, the IDS will act on the identified malicious network traffic, perhaps by blocking it. Again, there is no direct notification to the adversary of the output of the IDS machine learning classification. However, the adversary may be able to infer the classification output, since the adversary will notice their traffic being blocked, since the attack will fail or expected results are not received. The adversary can then infer that their network traffic was classified as malicious. Although, this observation of an attack failing or not receiving expected results may be indicative of some other problem that occurred, while the adversarial network traffic was in fact classified as benign.

In section 4 the discussion of perturbation of network traffic features is based on a Grey-box perspective, where the adversary is aware of the features which are being input into the traffic analysis machine learning classifier. Therefore, the adversary understands which network characteristics of their network flow must be perturbed to cause misclassification. However, the adversary may not know which subset of the features best represent another legitimate traffic class. Additionally, the adversary may not have an awareness of the representation of the network traffic characteristics. Lastly, the perturbation of certain features may cause an inadvertent change to another feature which may nullify the perturbation causing the adversary's network traffic to be correctly classified.

In the Black-box case of perturbation and AML, the features are unknown to the adversary. Therefore, the adversary is not aware of which features should be perturbed to mimic legitimate network traffic. In most cases the adversary would not be able to directly view the output of the traffic analysis machine learning classifier. However, the features could be vastly complex to be inferred even if the adversary were able to view the output. Therefore, in the black-box case, where features are unknown, the vulnerability to traffic misclassification is significantly reduced. As Appendix A, displays a large number of cyber security features from just a few different studies, the massive number of possibilities can be overwhelming to an adversary. Hence, the combination of as few as several different features themselves could be a defense against AML, since the adversary does not know which features to perturb.

5. AML CYBER SECURITY EXAMPLE

Background and Dataset

Earlier we discussed the ability of an adversary to conduct an AML attack in the context of a cyber security network detection classifier. We will now discuss our approach of AML conducted on a network scanning detector classifier and dataset consisting of network flow features originating from benign and malicious (Nmap network scanning) hosts. A notorious network scanning software tool leveraged by attackers is Nmap.

Normally attackers conduct network scanning in the initial phases of an attack to better understand the network and the ports open on a host. The attacker can then perform additional probes to uncover a specific software package and version listening on an open port. The discovery of a specific software package and version will assist the attacker in identifying a vulnerability to leverage in an attack.

The targeted SVM network scanning detector classifier was reconstructed based on the descriptions and features described in (Venkatesan, Sugrim, Izmailov, & Chiang, 2018). We implement the SVM classifier in the python programming language and scikit-learn. Additionally, the dataset leveraged was produced by the same authors (Venkatesan et al., 2018). The initial set of 11 different network flow features was reduced by feature selection to 3, consisting of the percentage of unsuccessful TCP connections, UDP, and ICMP connections (Venkatesan et al., 2018). The detector is trained using these 3 features which are extracted from network flows of benign (no scanning activity) and scanning (Nmap scanning) hosts.

Attacker Goal and Assumptions

The objective of the attacker is to hide (evade detection) the presence of the network scanning activity taking place on a network. Thus, the attacker will need to cause misclassification of a host's traffic flow as benign opposed to scanning. The attacker will need to perform several steps in order to cause a misclassification by the network scanning detector classifier.

The attack will be carried out from a grey-box perspective. In this scenario, the attacker does not have access to the trained target scanning classifier (including hyper parameters) and training dataset. We assume the attacker has knowledge of the specific 3 features being used in the target classifier and has access to a dataset of benign network flows (i.e. contains no scanning activity). The attacker may already have access to the target network being monitored by the network scanning classifier and can passively collect benign network flows. A benign network flow dataset can also be built offline by an attacker.

Approach

The steps for an attacker to achieve the objective of misclassification (AML) of network scanning as benign traffic will be further described. A prerequisite for an attacker to perform AML is the ability to collect benign network flows and generate a network flows for network scanning activity. The resulting network flows are processed and analyzed to create a labeled (i.e. Benign and Scanning) dataset. Each network flow correlates to a sample in the dataset consisting of the 3 feature values (percentage of unsuccessful TCP connections, UDP, and ICMP connections) required by the network scanning detector classifier.

Using the newly created dataset, the attacker uses the nearest neighbor algorithm to identify the benign sample which is closest to each scanning sample and records the 3 feature values. These feature values are used as a baseline to compute the amount of TCP traffic which must be generated to cause misclassification. The additional TCP traffic will cause the 3 feature values of the scanning sample to decrease and mimic a benign sample. Lastly, based on the proceeding calculations, the attacker must generate additional TCP traffic on the actual host during scanning activities to cause misclassification by the target network scanning classifier.

Results

Experimentation was conducted using the network scanning detector classifier and AML method previously discussed. The dataset was split into 80% and 20% for training and testing respectively. The test dataset consisted of 40 scanning and 45 benign samples. Before introducing the AML attack, the baseline accuracy of the network scanning detector classifier was 100%.

	Accuracy
Baseline	100 %
AML	76 %

Table 18- Baseline vs AML accuracy

The collection of benign network flows was simulated by using benign samples in the test dataset. A total of 20 scanning samples to be perturbed were also selected from the test dataset. During the AML attack, 20 of the 40 scanning samples were perturbed using the method previously described. All 20 of the perturbed scanning samples were misclassified as benign. As a result, the classification accuracy of the network scanning detector reduced to 76% as seen in table 1.

It is expected that all perturbed samples would be misclassified, since the AML attack is mimicking benign sample feature values. Thereby rendering the network scanner detector ineffective. Therefore, a defense against this type of AML attack is required to continue successful detection of network scanning activity.

We propose the addition of features and ensemble techniques as a defense to this AML attack. The addition of features will make an attack incomprehensible as the number of characteristics for an attacker to perturb would grow. While, an ensemble would allow the combination of weak learners to form a stronger learner.

A proposed ensemble learner is composed of a network scanner detector as previously discussed and an anomaly detector for an abnormal amount of traffic originating from a host. Additionally, the introduction of a feature which has a direct relationship with existing features. An introduction of a feature in the anomaly detector of average amount of traffic for a host would cause an increase as an attacker conducts the AML attack. The generation of additional TCP traffic during the AML attack would cause an anomaly detection.

6. CONCLUSION

Summary

Adversarial Influence of Machine Learning (AML) has become the forefront of the security of machine learning but has largely been applied to image classification, which has been established for many years. It is imperative to understand the effects of AML transferability to cyber security in network traffic analysis. Features are a fundamental component of the machine learning classification process. Therefore, the features of the cyber security area must be well understood.

We believe features play a crucial role in the classifier and in developing resiliency. It is important to look at these vulnerabilities from a grey and black box perspective. Even though in the grey-box perspective an adversary will be aware of the features leveraged by the classifier, they will still need to know the subset of features which are representative of their traffic flow. Additionally, a larger number of features to perturb will result in an increased cost to the adversary and in some cases may not be feasible. Lastly, from a black-box perspective, many disparate features themselves may be a sufficient defense against AML.

Future Work

We propose to conduct further exploration with several disparate features and an SVM for cyber security network detection classifier. Further exploration is expected to reveal the differences in the variety of fundamental feature distributions within a cyber security machine learning implementation in comparison to the image domain. As discussed earlier, the fundamental feature in image classification is the pixel intensity. An adversary need only perturb pixel values in an intelligent manner to achieve misclassification. Whereas in a cyber security machine learning classifier, the adversary would need to perturb disparate features of the network flow to achieve misclassification.

During our experimentation, we will perturb features of the network traffic flow, to achieve misclassification and evaluate the importance of features in an adversarial environment. The proposed experimentation will be evaluated using a representative cyber security network detection machine learning classifier. Lastly, we propose the development of defensive algorithms to protect against misclassification will include the use of ensemble machine learning methods which leverage a variety of disparate features for classification.

7. REFERENCES

Anderson, B., Paul, S., & McGrew, D. (2016). Deciphering Malware’s use of TLS (without Decryption). Journal of Computer Virology and Hacking Techniques, 1-17.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., & Roli, F. (2013, September). Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 387-402). Springer, Berlin, Heidelberg.

Chang, R. J., Harang, R. E., & Payer, G. S. (2013). Extremely lightweight intrusion detection (ELIDe) (No. ARL-CR-0730). ARMY RESEARCH LAB ADELPHI MD COMPUTATIONAL AND INFORMATION SCIENCES DIRECTORATE.

Dainotti, A., Pescape, A., & Claffy, K. C. (2012). Issues and future directions in traffic classification. IEEE network, 26(1)

De Lucia, M. J., & Cotton, C. (2018, May). Identifying and detecting applications within TLS traffic. In Cyber Sensing 2018 (Vol. 10630, p. 106300U). International Society for Optics and Photonics.

De Lucia, M. J., & Cotton, C. (2018, Nov). Importance of Features in Adversarial Machine Learning for Cyber Security. In *2018 Proceedings of the Conference on Information Systems Applied Research*, Norfolk, VA. ISSN: 2167-1508.

Dierks, T., & Rescorla, E. (2008). The transport layer security (TLS) protocol version 1.2 (No. RFC 5246) <https://tools.ietf.org/html/rfc5246>
(1 March 2018).

Google. “Transparency Report, “HTTPS Encryption on the Web.” Retrieved July 13, 2018 from <https://transparencyreport.google.com/https/overview>

Herrmann, D., Wendolsky, R., & Federrath, H. (2009, November). Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier. In *Proceedings of the 2009 ACM workshop on Cloud computing security* (pp. 31-42). ACM.

Kim, H., Claffy, K. C., Fomenkov, M., Barman, D., Faloutsos, M., & Lee, K. (2008, December). Internet traffic classification demystified: myths, caveats, and the best practices. In *Proceedings of the 2008 ACM CoNEXT conference* (p. 11). ACM.

Laskov, P., & Lippmann, R. (2010). Machine learning in adversarial environments. *Machine Learning*, 81(2), pp. 115-119.

Meyers, P. J. “HTTPS Tops 30%: How Google Is Winning the Long War.” Moz, (5 July 2016), Retrieved March 6, 2018 from <https://moz.com/blog/https tops-30-how-google-is-winning-the-long-war>

Muehlstein, J., Zion, Y., Bahumi, M., Kirshenboim, I., Dubin, R., Dvir, A., & Pele, O. (2017, January). Analyzing HTTPS encrypted traffic to identify user's operating system, browser and application. *In Consumer Communications & Networking Conference*

(CCNC), 2017 14th IEEE Annual (pp. 1-6). IEEE.

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., & Roli, F. (2017, November). Towards poisoning of deep learning algorithms with back-gradient optimization. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 27-38). ACM.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In Security and Privacy (EuroS&P), 2016 IEEE European Symposium on (pp. 372-387). IEEE.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (pp. 506-519). ACM.

Susarla, D., & Ozdemir, S. Feature Engineering Made Easy. Packt Publishing, 2018.

Venkatesan, S., Sugrim, S., Izmailov, R., & Chiang, C.-Y. J. (2018). On Detecting Manifestation of Adversary Characteristics. In Proceedings of the MILCOM 2018, Los Angeles, CA (pp.431-437). IEEE.

Wright, C. V., Monrose, F., & Masson, G. M. (2006). On inferring application protocol behaviors in encrypted network traffic. *Journal of Machine Learning Research*, 7(Dec), 2745-2769.

Zheng, A., & Casari, A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly, 2018.

Appendix A: Cyber Features Taxonomy

# Forward packets	Max throughput of backward peaks
# Forward total bytes	Backward min peak throughput
Min forward interarrival time difference	Backward STD peak throughput
Max forward interarrival time difference	Forward number of bursts
Mean forward interarrival time difference	Backward number of bursts
STD forward inter arrival time difference	Forward min peak throughput
Mean forward packets	Mean throughput of forward peaks
STD forward packets	Forward STD peak throughput
# Backward packets	Mean backward peak inter arrival time diff
# Backward total bytes	Minimum backward peak inter arrival time diff
Min backward interarrival time difference	Maximum backward peak inter arrival time diff
Max backward interarrival time difference	STD backward peak inter arrival time diff
Mean backward interarrival time difference	Mean forward peak inter arrival time diff
STD backward inter arrival time difference	Minimum forward peak inter arrival time diff
Mean backward packets	Maximum forward peak inter arrival time diff
STD backward packets	STD forward peak inter arrival time diff
Mean forward TTL value	# Keep alive packets
Minimum forward packet	TCP Maximum Segment Size
Minimum backward packet	Forward SSL Version
Maximum forward packet	Mean throughput of backward peaks
# Total packets	Forward peak MAX throughput
Minimum packet size	SSL session ID len
Maximum packet size	# SSL cipher methods
Mean packet size	# SSL extension count
Packet size variance	# SSL compression methods
TCP initial window size	TCP window scaling factor

(Muehlstein, Zion, Bahumi, Kirshenboim, Dubin, Dvir, & Pele, 2017)

Inbound bytes	Sequence of packet inter arrival times
Outbound bytes	Byte distribution of packet payload
Inbound packets	TLS version
Outbound packets	Order list of offered cipher suites
Source port	List of supported TLS extensions
Destination port	Selected cipher suite
Total duration of flow in seconds	Selected TLS extensions
Sequence of Packet lengths	Client public key length
Sequence of TLS record lengths	Sequence of TLS record times
Sequence of TLS record types	

(Anderson, Paul, & McGrew, 2016)

TCP packet size	Packet direction	Inter arrival time

(Wright, Monrose, & Masson, 2006)

IP packet size	Packet direction

(Herrmann, Wendolsky, & Federrath, 2009)

Standardizing Public Utility Data: A Case Study of a Rural Mid-Size Utility

Edgar Hassler
hassleree@appstate.edu

Joseph Cazier
cazierja@appstate.edu

Jamie Russell
russellja@appstate.edu

Thomas Mueller
muellerts@appstate.edu

Daniel Paprocki
paprockidj@appstate.edu

Center for Analytics Research and Education
Appalachian State University
Boone, NC 28608

Abstract

Energy is important to our daily lives. Energy data is important to utilities to meet operational goals and have better relationships with their customers. However, to get the most value out of their data they need to combine it with data from other secondary sources and apply advanced analytics techniques. This can only be done effectively and efficiently if utilities standardize their data in a way that allows the merger of their data with that of other sources to take place seamlessly. This article uses a case study to illustrate why small to midsize utilities should adopt a data standard, discusses some of the challenges in choosing and adopting a standard and concludes the process of moving to a data standard in our case. Results will be interesting to utilities and any industry contemplating taking advantage of the opportunities afforded by big data.

Keywords: Data Standardization, Energy Analytics, Sustainability, Big Data, Analytics

1. INTRODUCTION

Energy is ubiquitous to our modern lives. Consequently, energy data is also becoming ubiquitous in our modern digital era. Using this data to improve customer service, predict and prevent power outages, optimize power generation and persuade users to modify their energy consumption patterns can significantly improve the efficiency, profitability and sustainability of regional power providers. This is especially true when operational utility data is combined with other secondary data sources as described below.

During the 1990's, there was a wave of business innovation focused on what was branded *Business Intelligence* (BI). While definitions vary, the primary focus of BI was to gather and make sense of a company's *internal* data from their operations (Negash, 2014). This led to many operational improvements and efficiencies.

Today we are evolving into the next wave of data. Companies are integrating external or secondary data from outside of their core business to create even more value (McAfee & Brynjolfsson, 2012). This has come to be known as the era of *Big Data*, where data from multiple sources is combined, aggregated and analyzed in new and more advanced ways. Often by adding data from multiple sources there is a synergistic or exponential increase in the benefits that can accrue for the organization (Barton & Court, 2012).

Examples of secondary data in the utility industry might include merging property tax records. Such records can give the utility information such as the size of the home, heating source, tax value, age and other information that can be analyzed to help understand power consumption patterns. Likewise, weather data, a key driver of electricity consumption, can be merged into the data set to better understand how it impacts electricity use, used to predict future energy usage, and simulate what-if scenarios. Additional information like demographic data can be acquired and merged into the company operational data to better segment and reach out to consumers with individualized messaging via a consumer preferred media channel based on age, income, occupation, educational level and other factors. All of this can lead to a smarter, data driven utility.

To take advantage of the promise offered by big data and analytics, companies need to first be able to acquire the data by successfully merging their operational data with other relevant secondary data for analysis. To do this, there needs to be a way to connect their data with relevant data from outside the organization. This is usually done in the form of a primary key that uniquely identifies a record in each of the internal and external data sets.

However, the merge can only take place if the keys match. Since the most common identifier across most utility and external sources of data is the utility service address, this is generally the best key to use. However, in many data sets, addresses do not follow a set standard with many possible variations in how they are recorded and stored. Additionally, the data are fraught with errors as clerks often try to decipher hard to read handwriting or manage typos in a consistent manner. There are many "right" ways to write and store an address. This is the fundamental problem, as different formats (or lack of formats) won't match on a secondary data set merge.

To solve this problem, utilities can adopt and consistently follow a data standard. By having their data recorded in a consistent methodological way, they can merge their data with outside sources using the same format or convert the data to that format with a conversion code block. However, to do so efficiently, the data needs to be stored in a standard consistent format for transformation and merging of the data.

In this applied research paper, we utilize a case study to discuss some of the benefits data science can bring to the utility industry, illuminate the role of adopting, migrating to and enforcing a data standard in the process and share our experience with a rural public utility as they work to transform their data and adopt a data standard.

2. LITERATURE REVIEW

Public utilities makeup to 94.3% of all utilities in the U.S. and serve 68.3% of all energy consumers (Public Power, 2017). Public power utilities are governed by elected and appointed boards, which include mayors, city council members and citizens for the common good, rather than by large investors groups. The mission of a publicly utility is to optimize benefits for local customers, usually in the form affordable energy rates (California Energy Commission, 2017).

Some of the current and potential benefits derived from the application of analytics to a utility's big data merged with secondary data sources include: benefits for the utility, consumers and society as a whole. We will briefly discuss each of these in the subsections below.

For the Utility

As with all businesses, utilities provide a service for which consumers are willing to pay. That service is the reliable and affordable delivery of power on demand. At its most simplistic, the connections between the utility and the consumer are the power cables and the mail service by which the bill is delivered and the check is received. The records or data which have been typically recorded consist of a monthly meter reading. The widespread adoption of smart meters and other monitoring equipment is rapidly changing this traditional model bringing with it great potential benefits and also some real challenges.

The adoption of smart meters and other monitoring equipment generates near real-time information which has the potential to dramatically improve utility operations in many areas including: successfully integrating intermittent renewable energy and other distributed energy resources, predicting and quickly respond to outages, increasing overall grid efficiency, improving load forecasting, providing rapid anomaly detection, improving demand side management, and, in the future, successfully integrating the growing demand for electric vehicle charging (Aria and Bae, 2016; Katz, 2018; Schuelke-Leech et al, 2015; Wen et al, 2018; Zhou et al, 2016). Another benefit of successfully interpreting this data is the ability to change the communication paradigm from a monthly bill mailing to real-time, individualized, interactive consumer messaging.

The benefits above come with the challenges of managing a huge increase in data volume compounded by integration of newer technologies into older existing infrastructure, and the fact that these increased data volumes are being overlaid upon legacy databases and other dated information systems (Katz, 2018; Schuelke-Leech et al, 2015; Zhou et al, 2016). Suggestions for managing these challenges include data compression, the creation of comprehensive data frameworks, and making sure that utility data systems are aligned with the best practices of modern data analytics and data science (Akhavan-Hejazi and Mohsenian-Rad, 2018; Munshi and Mohamed, 2017; Wen et al, 2018).

The potential benefits of big data for utilities are numerous. Even though the specific mechanisms for enacting those benefits are not fully developed, the informational underpinning for these systems will be based on the best practices of data analytics.

For the Consumer

Targeted, clear and well-presented messaging from power providers, to consumers, will build an advantageous relationship. Energy consumers are partial to information that allows them to make informed decisions on utilitarian, psychological and social implications within their communities (Hartmann & Apaolaza-Ibáñez, 2012). Social media has become an essential tool in promoting the corporate brand and communicating directly with the consumer. The purpose, many times, is to share a pro-environment attitude and to encourage customers to take action (Ballew, Omoto, & Winter, 2015). Social media consumers hold reduced risk aversion, sustain higher brand loyalty, and are overall more satisfied with the brand (power company) when social media information is available (Reisenwitz, 2013).

An essential component in energy consumption decision making is sustainable development. It has become a crucial topic for consumers. There is heightened interest in waste management, greenhouse gas emissions and renewable energy. Social media technologies have attracted attention as being a viable tool to encourage sustainable actions (Sogari, Pucci, Aquilani, & Zanni, 2017). Many times, consumers who engage social media are more aware of sustainability practices, then make buying and consumption decisions based on social media information. Studies indicate millennial consumers hold a strong affinity with social messaging and appreciate companies that help in acquiring information (Dabija, Bejan, & Tipi, 2018; Hartmann and Apaolaza-Ibáñez, 2012).

For Society

Big data and cutting-edge analytics have the opportunity to play an important role in improving the way the energy sector interacts with society as a whole. As the world becomes more focused on utilizing resources in an efficient manner, it is imperative that the energy sector continues to innovate to meet these societal demands. By using ever improving computing power and data analytics, the energy sector can begin to have a focused conversation with their clientele regarding energy consumption and conservation. Utility companies can use analytics to communicate with customers regarding upcoming high electric demand events (peak power) that put a strain on the environment and people’s checkbook. This can help society understand when electricity generation is most expensive and most carbon intensive, and how they might conserve during those periods (Kenworthy, 2016).

This benefits everyone in a community, as the impact of power production on the environment is lessened, both the utility company and the clientele benefit from reduced peak power costs, and there is a general increased community exposure to these important issues (Feldman et. al., 2015). The educational benefits can cascade down into other sectors of one’s life, changing the way one makes choices that may impact the planet. This not only includes energy, but also sustainable consumption of water, and food. There is a great potential for society to benefit from our energy providers taking advantage of these exciting new technologies.

Next, we discuss some of the challenges to adopting big data techniques and how to overcome them.

Challenges to Adopting Big Data in Public Power

While these benefits are real, many challenges remain in place before they can be fully realized, especially among our small rural utilities and cooperative municipalities which make up the majority of utilities in the U.S. (Public Power, 2017). There is a lot of data being collected, especially as we start to enter the age of smart meters and appliances. However most of these public utilities use the data primarily for their billing system and internal operational efficiencies. While important, this leaves the true potential of big data for customer engagement, sustainability and total business optimization largely untapped.

Analytical tools continue to be developed and provide value for many firms in other sectors, or the better funded large private utilities, but many public utilities continue to be left behind. The primary reasons include:

- *Non-Standard Data* - Each public utility has evolved to keep data in its own way, with their own system targeted to their business environment and influenced by their history and focus. Some controlled data entry, like customer addresses, tightly. Most do not.

- *Legacy Rules and Regulations* - While rules and regulations differ depending on the utility size, ownership structure and location, they can be quite influential on a regulated public utility’s use, or lack of use, of their data for activities beyond billing. A detailed analysis of rules and regulations is beyond the scope of this paper, but we acknowledge it here.

- *Lack of knowledge* - Most public utility managers and engineers are smart, well trained hard-working people. However, the discipline of Data Science did not exist as a discipline or course of study when most of them finished school, thus they need training and guidance to understand the benefits of applying analytics to their data to improve their business and customer relationships. While we give some examples of how this can be used in this paper, the primary purpose here is to discuss the importance of a data standard.

Data Standardization

Although all these reasons are important barriers to taking advantage of the promise of big data and analytics in this digital age, the primary focus of this paper is on the issue of data standardization. The rest of this paper explores how data standardization can create value for the consumer, public utilities (especial small to mid-size ones) and society; discusses some of the challenges to standardizing data; and shares a case study for a small rural utility.

3. MATERIALS AND METHOD

A Primer on Merging Data

Pieces of information about a customer (i.e. name, address, phone, etc.) are usually stored together as a unit called a record. As like records are collected into a set it becomes necessary to have a unique identifier for each record. This unique identifier is known as a primary key in database terms (Elmasri & Navathe, 2000). The primary key may be a simple account number or other identifier that is assigned to only one customer.

In order to relate different record sets to each other, the primary key from one record is included as a piece of information in other related records (Elmasri & Navathe, 2000). For example, a customer's name and contact details may be stored in one record set, while their meter reading with the date and time of the reading are stored in a different record set. This avoids duplication of customer details in each record of a reading and improves efficiency. To ensure that each reading is related to the correct customer, we simply include the associated customer identifier in each meter reading.

To recombine the details of records from two different record sets one simply matches the primary key from a record with the records containing the same key in the second set. This is known as a *join* in database terms (Elmasri & Navathe, 2000).

The ability to join two different record sets that do not share a key is thus challenging, but pivotal to linking data from different sources to create extended data records for each consumer.

Standardizing Public Utility Data

By standardizing the way that small utilities record addresses of their meters, the ability to merge secondary data, such as demographics and property records, becomes much more achievable. If they have a public standard, such as using the postal record standard (USPS Pub 28) or a 911 standard, the new data can be transformed into this same format and merged with the existing data.

Challenges to Data Standardization

With all the potential benefits of using big data and analytics to improve operational efficiencies, reduce their environmental impact and build a more engaging relationships with their customers, it is important to address the barriers of data standardization. However, some of the technical and operational barriers to achieving this can be formidable, as illustrated below.

Resource Constraints

For most small organizations, few resources can be dedicated to cleaning data and performing analytics. Data analytics can be very costly in several ways, not just financially. To retroactively apply the standard to addresses already in the system is very time-consuming and can take hundreds, even thousands of hours. Many smaller utilities do not have the time, capital, or man-power to put towards retroactively cleaning addresses to conduct analysis. This presents a technical and resource issue that many companies do not have the ability to overcome on their own.

Customer Variability

Utility firms do not cater to only one type of customer, there are several different types to consider when determining a standard format for the input of addressing data. Energy meters are attached to many different types of structures, including residential, commercial, educational, governmental, and infrastructure. Each structural type represents a distinct set of shared characteristics and is charged a rate aligning with the purpose of the structure. Houses and certain apartment complexes are considered to be residential. Local private businesses are the commercial structures. The buildings located on a university campus are educational structures. Meters classified as infrastructure are structures that use power but are not buildings such as street lights, well pumps, or large signs covered with lighting. Each of these types of structures has its own unique identifiers that need to be clear and available in the addressing data. Additionally, there are subsets within some of the categories, such as single-family homes, and duplexes. It is non-trivial finding an addressing standard that accommodates all the different categories. Utilities need to start with a base addressing standard, and then expand to accommodate all possibilities, including residential, commercial, other building types, and infrastructure addresses.

Multiple Types of Merges

An additional challenge is that one address does not always mean one meter. For example, the property records received from county officials are optimized for tax collection. They generally contain one record for each tax structure, not utility meter. Thus, there are four common types of matches when merging the datasets:

1. One meter address to one property record (e.g. single family home)

2. Multiple meter addresses to one property record (e.g. apartment building)

3. One meter address to multiple property records (e.g. home with detached garage or barn)

4. Multiple meter addresses to multiple property records (e.g. shopping mall complex)

In cases 2 - 4 this creates additional challenges because one cannot discern exactly to which structure a meter is attached.

Standards of Available Data to Merge

Many towns and cities utilize the E911 addressing standards, which assigns a standardized address for emergency organizations to easily locate a building or structure. These standards are often self-determined but follow addressing standards that have been developed over years of trial and error. This standard is based primarily on the guidelines laid out in the United State postal code, which is universal throughout the nation. However, because each municipality has their own digressions from the absolutes of the standard outlined in the postal code, E911 addressing standards are not necessarily the same from one town to the next. This poses a problem because each municipality may impose a slightly different set of addressing standards than the set of standards that have been chosen for one application.

4. CASE STUDY FOR A RURAL UTILITY

We worked with a small rural utility in the Southeastern part us the U.S. with around 10,000 meters to help identify a data standard and transition their legacy data to that standard. The goal was to prepare them to integrate additional secondary data into their system, so they could segment their customer base and personalize their outreach to consumers around important issues including peak power, pre-pay options, and energy assistance programs.

In the sections that follow we discuss some of the challenges faced and how we addressed them.

Merge Goal and Approach

As a first step, the utility wanted to merge publicly available government data from property tax records with their billing records to better understand power consumption patterns as they relate to home characteristics. The county exported a spreadsheet (Excel) from their database containing information such as address, year built, tax value, square feet and heating source. The task was to then merge this with 10 years of monthly energy readings for each residence. ~8,000 relevant meter ID's and addresses were selected for this analysis. A sample of the type of data from each original source is shown in Figure 1 (in Appendix A).

Recall from the previous discussion on databases that in order to merge (join) two different record sets there must be a shared key between them. The most immediate key to merge secondary data to the example utility company meter locations is the address. However, each meter location has two fields, each containing a portion of the information required to create a standardized address. To create a standard address, each field must be broken down into granular pieces and the pieces then recombined to form a complete address.

Constructing a Standard

Once information from the two address fields are broken down into distinct pieces, a standard for each field needs to be in place. As a starting point, USPS Publication 28 - which outlines preferred ways of representing all aspects of an address - was chosen. Because Publication 28 standards are meant to accurately represent a location for the purpose of parcel delivery, it provides a widely recognizable method of storing address information. Other possibilities include the E911 standard, geocodes or other positioning systems.

The final standard for address data consists of 12 individual fields. The fields were derived through an iterative process in which fields were added to the USPS Publication 28 base. A complete list of fields, along with their definitions, can be found in Appendix B.

Data Conversion

To illustrate how an address is broken up and cleaned, Figure 2 (see appendix A) represents the raw data. From this point, programming scripts are used to separate out all of the pertinent information.

There are several pieces of information located in each field of the raw data. The field titled Service Address contains a place name (Nathans Walk), Secondary Unit Indicator (A3), City (Local City), State (NC), and Zip (77777). The Line and Pole field contains a utility grid identifier (D16), Street Number (872), Predirectional (W), Street Name (EMPEROR), Address Suffix (ST), and Secondary Indicator (APT). Figure 3 (see Appendix A) shows the address fields properly segmented.

The individual fields may then be viewed as subcomponents of a larger, composite key that will be used to merge the datasets.

Data Filtering

After determining a standard, data management software was used to retroactively apply the standard to all addresses in the current dataset. The addresses were then filtered to uncover records missing critical components of the address such as street number, street name or street suffix. Records for which this information was unrecoverable were removed from the dataset as they cannot be matched. Essentially, they do not have a key.

Logical Steps of the Merge with Secondary Data

The following is a summary of the steps taken to prepare and merge the data, including an approximate percent of merged observations after each step.

Step 1: Upper case the "Service Address", "Line and Pole", and county address fields.

- This ensures that two fields with equivalent content are not prevented from matching due to a difference in casing.
- Nothing merges at this point, because there are no identically defined fields to match with in the meter address data set.

Step 2: Begin creation of distinct address fields by parsing (breaking up) the "Line and Pole" and "Service Address" fields.

- The "Line and Pole" field begins with a grid identification number, which is not part of an address. This is first segmented out into its own field and stored for future use but will not be useful for the merge with secondary data.
- After removing the grid identification number, the beginning of each string now contains a numeric value indicating the street number.

This numeric value is separated from the character values that follow and used to fill the address number field. The remaining characters were moved into a temporary address street column to be further broken apart.

- Nothing merges at this point, because the data is not fully segmented.

Step 3: Create Street Name and Suffix fields then standardize Street Suffix to Publication 28 addressing standards.

- A new field for Street Suffix is formed, using the example utility meter address data. First, to properly format the suffix field, the main dataset is segmented into smaller datasets based on patterns displayed in the line-and-pole address that indicate a proper suffix. For example, all variations of the suffix ROAD are removed to form a new dataset, standardized to RD, and then reintegrated back into the master list. A similar process is followed for all other major suffixes.

- This step is completed for both the utility and county datasets and the fields for address number, street name, and street suffix will be utilized as keys to merge the meter data with secondary data from the property records.

- After performing these operations, approximately 50% of the meter locations merged with a property record from the county tax records.

Step 4: Correct and clean data entry errors found in the Street Name fields.

- Many observations have small errors in the street name that prevent the two data sources from merging properly. Therefore, it is appropriate to look at the observations that do not merge, which isolates some of the errors. Then manually, one must sort through the street names and document errors. For example, a street name like "George Critcher", might be abbreviated as "GEO CRITCH." This type of error or misspellings should be isolated, and then changed with programming scripts.

- These operations increase the overall merge yield, while standardizing the data at the same.

- After performing these operations, approximately 83% of the meter locations merged with a county property record.

Step 5: Manually search for parcel outlines in the tax record website to match parcels to addresses.

- Each meter is located on a parcel of land, which is recorded in the county tax records. Therefore, searching each meter address in google maps shows the parcel outline - providing a method of visually matching a meter address with a parcel ID. The parcel id is manually recorded for each unmatched meter, and then the tax records for these remaining meter addresses are merged by that parcel ID.

- After performing these operations, 87.6% of the meter locations merged with a property record from the county tax records.

While this process does not completely merge all of the records, it does allow us to perform analytics on the majority of the meter locations. It was decided to have our analytics team stop at this point as merging additional records would require a person to physically visit each meter to confirm the address. The utility plans to have some of their meter readers complete this task in the near future and report back with the data.

Discussions with the utility database administrator has revealed that a project is underway to geolocate each meter. As shapefiles are available for each parcel, it is hoped that an additional step of matching meter locations inside a parcel-shape can be added to complete the merging and standardization of the datasets when geolocations become available.

Figure 4 provides a visually summary of the types of merges found in the final merged dataset.

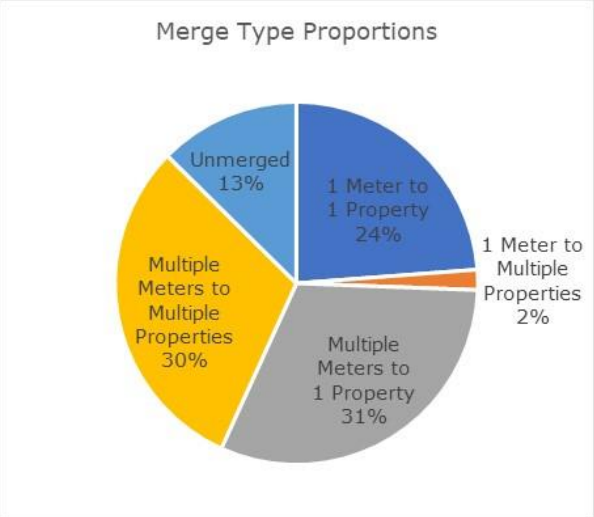


Figure 4. Merge types in final dataset

5. DISCUSSION

When merging enhancement data into an existing dataset, considerable work must be accomplished before analytics can begin - especially when relying on address data as a key for the merge. With address information, it is important to define a standard, convert all data to the standard through dissection and iterative cleaning, and filter the data for cases that must be handled manually.

In the course of our investigation, we found that using a more granular address standard required additional preparation and parsing work in the beginning, however it simplified many of the cleaning tasks and, more importantly, we have found it is now easier to reassemble the pieces in new formats to complete additional merges.

Despite only merging 88% of the records, the utility in this case was very satisfied with the results. The resultant dataset is sufficient to complete the next phase of the project, which is segmenting their customers to better understand power consumption patterns as they relate to home characteristics. Based on this knowledge, better communication and incentive programs can be designed and targeted where they will have the most impact on reducing peak demand.

There is also immediate value in standardizing data before analytics even begin. The utility has increased confidence in billing records, it facilitates standard operating procedures, and assists efforts regarding legal and regulatory compliance.

Smart meters are changing the landscape for utilities. Data standards becomes even more important as the volume, velocity, and variety of data increase. The increasing need for real-time analysis means there is reduced time for extended data cleanup efforts.

This means that it is important to maintain data within standards once they are implemented. To do so may include staff training programs to improve data entry and curation. It may also include improved checks and enforcement within processing systems.

6. CONCLUSION

Data standards are an important part of the business intelligence and data analytics framework. They make the merging of additional data sources possible and facilitate the transfer of data to modern tools, thus enhancing decision-making.

In future work, we are excited to expand the merging of data to additional sources such as demographics and studies regarding how and what to communicate with consumers to reduce peak demand. With additional utilities collaborating in the pooling and merging of data, big data, indeed clean big data, that can lead to peak demand reductions is within the grasp of even small rural utilities.

7. ACKNOWLEDGEMENTS

Special thanks to our graduate students Dalton Stout and Lucas Stinson for their help on this project.

8. REFERENCES

Akhavan-Hejazi, H., & Mohsenian-Rad, H. (2018). Power systems big data analytics: An assessment of paradigm shift barriers and prospects. *Energy Reports*, *4*, 91–100. https://doi.org/10.1016/j.egyr.2017.11.002

Arghandeh, R., & Zhou, Y. (2017). *Big Data Application in Power Systems*. Elsevier.

Arias, M. B., & Bae, S. (2016). Electric vehicle charging demand forecasting model based on big data technologies. *Applied Energy*, *183*, 327–339. https://doi.org/10.1016/j.apenergy.2016.08.080

Ballew, M. T., Omoto, A. M., & Winter, P. L. (2015). Using Web 2.0 and Social Media Technologies to Foster Proenvironmental Action. *Sustainability*, *7*(8), 10620–10648. https://doi.org/10.3390/su70810620

Barton, D., & Court, D. (2012). Making Advanced Analytics Work For You. *Harvard Busimess Review*, *90*(10), 78.

Dabija, D.-C., Bejan, B. M., & Tipi, N. (2018). Generation X versus Millennials communication behaviour on social media when purchasing food versus tourist services. https://doi.org/10.15240/tul/001/2018-1-013

Differences Between Publicly and Investor-Owned Utilities. (n.d.). Retrieved July 14, 2018, from http://www.energy.ca.gov/pou_reporting/background/difference_pou_iou.html

Electricity and the Environment - Energy Explained. (2014, November 22). Retrieved July 14, 2018, from https://www.eia.gov/energyexplained/index.php?page=electricity_environment

Elmasri, R., & Navathe, S. B. (2000). *Fundamentals of Database Systems* (3rd ed.). Reading, MA: Addison-Wesley.

Feldman, B., Tanner, M., & Rose, C. (2015). *PEAK DEMAND REDUCTION STRATEGY* (p. 59). Advanced Energy Economy. Retrieved from http://info.aee.net/hubfs/PDF/aee-peak-demand-reduction-strategy.pdf?t=1446657847375

Hartmann, P., & Apaolaza-Ibáñez, V. (2012). Consumer attitude and purchase intention toward green energy brands: The roles of psychological benefits and environmental concern. *Journal of Business Research*, *65*(9), 1254–1263. https://doi.org/10.1016/j.jbusres.2011.11.001

Kenworthy, B. (n.d.). Real-time Data Analytics. Retrieved July 14, 2018, from https://www.elp.com/articles/powergrid_international/print/volume-21/issue-3/features/real-time-data-analytics.html

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, *90*(10), 60–68.

Munshi, A. A., & Mohamed, Y. A.-R. I. (2017). Big data framework for analytics in smart grids. *Electric Power Systems Research*, *151*, 369–380. https://doi.org/10.1016/j.epsr.2017.06.006

Negash, S. (2004). BUSINESS INTELLIGENCE. *Communications of the Association for Information Systems*, *13*, 20.

Publication 28 - Postal Addressing Standards. (n.d.). U.S. Postal Service.

Reisenwitz, T. H. (2013). A comparison of the social media consumer and the non-social media consumer. *International Journal of Internet Marketing and Advertising*, *8*(1), 19–31. https://doi.org/10.1504/IJIMA.2013.056587

Schuelke-Leech, B.-A., Barry, B., Muratori, M., & Yurkovich, B. J. (2015). Big Data issues and opportunities for electric utilities. *Renewable and Sustainable Energy Reviews*, *52*, 937–947. https://doi.org/10.1016/j.rser.2015.07.128

Sogari, G., Pucci, T., Aquilani, B., & Zanni, L. (2017). Millennial Generation and Environmental Sustainability: The Role of Social Media in the Consumer Purchasing Behavior for Wine. *Sustainability*, *9*(10), 1911. https://doi.org/10.3390/su9101911

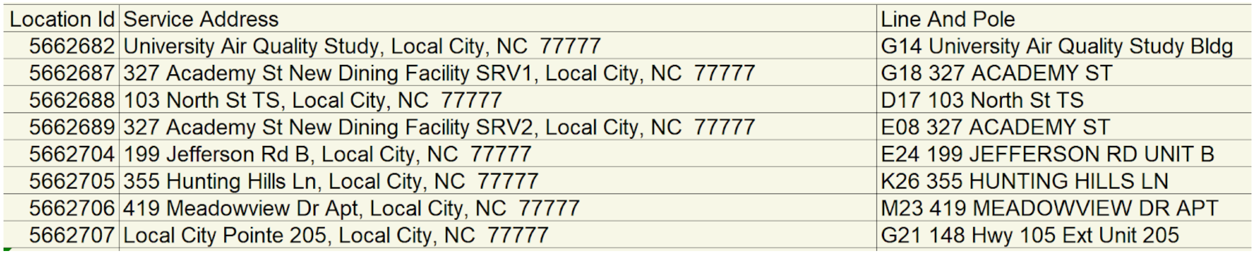
Stats and Facts. (n.d.). Retrieved July 14, 2018, from https://www.publicpower.org/public-power/stats-and-facts

Wen, L., Zhou, K., Yang, S., & Li, L. (2018). Compression of smart meter big data: A survey. *Renewable and Sustainable Energy Reviews*, *91*, 59–69. https://doi.org/10.1016/j.rser.2018.03.088

Zhou, K., Fu, C., & Yang, S. (2016). Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews*, *56*, 215–225. https://doi.org/10.1016/j.rser.2015.11.050

Appendix A

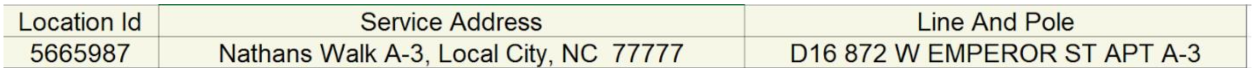
Example Figures



Location Id	Service Address	Line And Pole
5662682	University Air Quality Study, Local City, NC 77777	G14 University Air Quality Study Bldg
5662687	327 Academy St New Dining Facility SRV1, Local City, NC 77777	G18 327 ACADEMY ST
5662688	103 North St TS, Local City, NC 77777	D17 103 North St TS
5662689	327 Academy St New Dining Facility SRV2, Local City, NC 77777	E08 327 ACADEMY ST
5662704	199 Jefferson Rd B, Local City, NC 77777	E24 199 JEFFERSON RD UNIT B
5662705	355 Hunting Hills Ln, Local City, NC 77777	K26 355 HUNTING HILLS LN
5662706	419 Meadowview Dr Apt, Local City, NC 77777	M23 419 MEADOWVIEW DR APT
5662707	Local City Pointe 205, Local City, NC 77777	G21 148 Hwy 105 Ext Unit 205

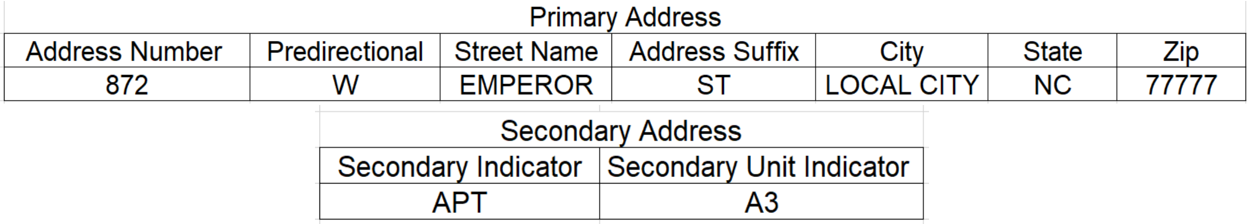
Figure 1. Sample Merge Data from Source Tax Records and Utility Address

Note in Figure 1 that the service address field has information regarding the city, state, zip code, and a place name (Local City Pointe), whereas, the Line and Pole field has the street number, street name, secondary unit, and the secondary unit indicator. The first step in the merger process is to split this address into the granular bits of information so they can be matched as a composite key where each sub-component is matched separately. After this standardization, additional information (i.e. property tax records) can be matched to the addresses.



Location Id	Service Address	Line And Pole
5665987	Nathans Walk A-3, Local City, NC 77777	D16 872 W EMPEROR ST APT A-3

Figure 2. Sample Non-Standard Addressing Data



Primary Address						
Address Number	Predirectional	Street Name	Address Suffix	City	State	Zip
872	W	EMPEROR	ST	LOCAL CITY	NC	77777

Secondary Address	
Secondary Indicator	Secondary Unit Indicator
APT	A3

Figure 3. Properly Segmented Address Fields

Appendix B

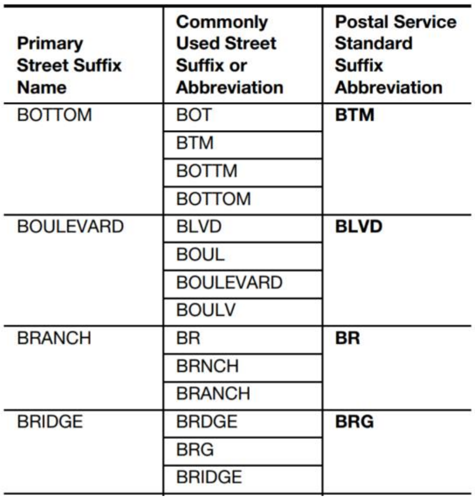
Final Addressing Standard

Street number - the number listed on the front door and/or mailbox of the property. This is positional descriptor along a street and is related to the cut in the street for the structures driveway. All addresses should contain a street number.

Predirectional Identifier - North, South, East, or West, shortened to only the first letter of the direction. This includes street such as "E King St" or "West King St." Not all addresses will have a predirectional identifier.

Street Name - the name of the road containing a curb cut for the site-in-question driveway. All addresses should contain a street name.

Street Suffix - the type of street the structure sits on, such as street, road, lane, etc. All the suffixes are abbreviated for entry into this field. All addresses should contain a street suffix.

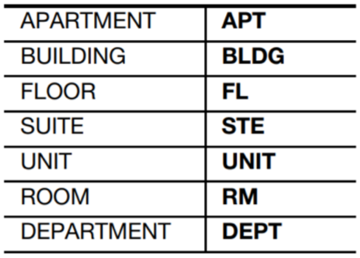


Primary Street Suffix Name	Commonly Used Street Suffix or Abbreviation	Postal Service Standard Suffix Abbreviation
BOTTOM	BOT	**BTM**
	BTM	
	BOTTM	
	BOTTOM	
BOULEVARD	BLVD	**BLVD**
	BOUL	
	BOULEVARD	
	BOULV	
BRANCH	BR	**BR**
	BRNCH	
	BRANCH	
BRIDGE	BRDGE	**BRG**
	BRG	
	BRIDGE	

Figure B-1. Suffix Abbreviations (Publication 28)

Postdirectional Identifier - North, South, East, or West, shortened to only the first letter of the direction. This includes addresses such as "US Hwy 421 S" or "US Hwy 421 N". Not all addresses will have a postdirectional identifier.

Secondary Unit Indicator - these are identifiers that indicate the metered address is part of a larger structure. Secondary unit indicators include apartment (apt), suite (ste), and unit. Not all addresses will have a secondary unit indicator.



APARTMENT	**APT**
BUILDING	**BLDG**
FLOOR	**FL**
SUITE	**STE**
UNIT	**UNIT**
ROOM	**RM**
DEPARTMENT	**DEPT**

Figure B-2. Secondary Unit Designators and Abbreviations (Publication 28)

Secondary Number - this number identifies the apartment or unit number of the metered address. In some cases, the secondary number can be a letter or a combination of numbers and letters. Not all addresses will have a secondary number.

Infrastructure Identifier - this contains information indicating that the meter is not attached to a residence. A structural identifier can indicate the meter is for a sign, well, traffic signal, etc. Not all addresses will have a structural identifier. We decided that this requires a separate field because this type of address is not a structure like most addresses. These observations should not have associated tax record information, as they are not buildings. In the future, we may decide to separate these items out to analyze energy consumption for items other than commercial or residential properties.

Place Name - this identifies the title of the collective in cases where the meter is assigned to one parcel or unit of a larger structure. Place names can be titles of apartment complexes, shopping centers, malls, etc. Not all addresses will have a place name. This information is important to separate out, as it contains unique identifying information.

City - the city that each address can be found in. Almost all of the example utility company meters are located in the city limits aside from a few, which are located just outside city limits in unincorporated territory. Not all addresses will list a city.

State - this indicates the state that each meter is located in. All the example utility company meters are located in North Carolina (NC). All addresses will list a state.

Zip Code - the postal code assigned by the USPS to each address. All of the addresses in the example utility company data share the same zip code except for the few on-campus addresses which have the university-specific zip code. All addresses will list a zip code.