Application of Routine Activity Theory to Cyber Intrusion Location and Time

Kevin Bock, Sydnee Shannon, Yazdan Movahedi, and Michel Cukier
University of Maryland
College Park, Maryland, 20742
kbock@umd.edu, sydnee12@umd.edu, ymovahed@umd.edu, mcukier@umd.edu

***Abstract*—Routine Activity Theory (RAT) is used by criminologists to explain the situational factors that influence crime in the physical world. RAT states that crime is most likely when a motivated offender, a vulnerable victim, and a lack of capable guardianship converge. We hypothesize that the time of cybercriminal actions will align with the principles of RAT. We analyzed data from over 20,000 intrusions on a large set of target computers over a period of four years. A statistically significant pattern is found in the time of intrusions in the local timezone of the victim hosts and native timezone of the attacker; intrusions geolocated to China demonstrate a stronger statistically significant pattern. The results suggest that RAT does apply to cyberspace, and further conclusions and policy implications are discussed.**

Keywords— Routine activity theory, cyber intrusion, attack time, geolocation, honeypots, attack patterns.

I. Introduction

Although criminologists have long employed theoretical frameworks that analyze crime in the physical world, little work has been done to explore how these theories apply to cyber intrusions. Our research aims to address this gap by exploring the applicability of Routine Activity Theory (RAT) to cybercriminals in an effort to learn more about attacker behavior. More specifically, we examine the time of day cyber intrusions were recorded over a period of four years to look for periods of higher intrusion volume.

RAT is a subset of Crime Opportunity Theory developed by Felson et. al [1]. The premise of RAT is that the convergence of a motivated offender and a vulnerable victim in the absence of a capable guardian increases the likelihood of crime occurring. Criminologists have used RAT to explain the situational factors that influence crime and, by extension, patterns of criminal activity. Several studies confirm that the frequency of victimization is particularly high during times when motivated offenders and vulnerable targets are likely to meet. For example, analysis of victimization surveys has shown that the risk of violence against youth varies greatly by time of day [2], [4], [3], [5]. This variation in victimization can be attributed to the fluctuations in capability of guardians and the vulnerability of the victims [6]. RAT is commonly used to explain why shootings occur most often late at night or early in the morning, as the guardianship of the police is typically at its lowest [7]; why most burglaries occur during the work hours, as the victims are most vulnerable [8]; and why most juvenile crimes occur immediately in the mid-afternoon (after school), because the attackers are most available [9], [10].

As it pertains to cybercrime, RAT has been successfully applied to account for traditional forms of online offenses, such as harassment and stalking [11]. The theory was also used to explain online victimization patterns [12], and again later to describe cybercrime at a macro level [13]. RAT was also tied to patterns of online fraud victimizations [14], [15]. Chon described the application of RAT to cyberspace as a "pragmatic conceptualization [that] has been widely accepted and dogmatically adopted" [16]. All of these studies support the idea that RAT is applicable in the context of cyberspace [17], [18].

Additional research has been conducted on the applicability of RAT to various system attacks. Lin conducted a small-scale study into attack patterns on mail servers [19]. Although Lin admits that a relative lack of data prohibited any authoritative analysis, he found that most attacks occurred in the evening, and the fewest in the early morning. Maimon examined the daily patterns of attack attempts on a university network by analyzing intrusion detection systems data across three different blocks of time and found that almost 60% of attacks in 2007 occurred between 9AM and 5PM [20]. While the empirical study in [20] focuses on attack attempts based on data from intrusion detection systems, this paper uses a much richer dataset of successful attacks, i.e., cyber intrusions.

This paper extends the application of RAT to the domain of cyber intrusions, an area in which sound criminological understanding is crucial, particularly as cyber threats become increasingly dangerous and widespread [21]. We assume a causal relationship between the fluctuations of the three metrics defined by RAT and the daily pattern of crime [20]. We explore the daily patterns of cyber intrusions into a set of target computers, i.e., honeypots. A honeypot is an information system resource used to divert attackers away from critical resources as well as a tool to study an attacker's methods [22]. Additionally, using the Internet Protocol (IP) addresses of the computers used to launch the intrusions, we perform a coarse-grained geolocation to derive the approximate time zones of the attacker. IP Geolocation is the process by which physical locality information is derived from an IP address [23]; such a practice has already been shown to be sufficiently accurate at the country or city level [24].

This paper shows that RAT could be applied to cyberspace. Thus policies could be derived, e.g., an allocation of resources based on the likelihood of facing a successful attack. The rest of

the paper is structured as follows. Section II presents the hypothesis. Section III details our method and Section IV our results. Section V discusses the obtained results. Section VI lists the limitations. Finally, Section VII concludes the paper.

II. Hypothesis

The definitions of attackers, targets, and guardians in cyberspace differ from those in the physical world. Unlike in a physical setting, the attacker might not be a person; rather an 'attacker' is the machine associated with the IP address used to connect to the victim host. Importantly, this is not necessarily a physical individual, or even the IP address from which the attack originated, it is just the last machine used to propagate the attack towards the victim host. Guardianship in cyberspace also operates differently. Many systems implement a firewall or some type of intrusion detection: a "guardian" that operates effectively at all times, regardless of the load on the system and the time of day [25]. In addition, many systems have a dedicated system administrator team for observing network traffic and looking for anomalies: a guardian that could be considered weaker at certain points in the day [26]. Attackers could consider the system weaker at night, in the absence of a watchful system administrator, or during the height of the system load, in which the heightened traffic could obscure an ongoing attack.

We hypothesize that the data will align with principles of RAT and that the time of intrusions will show a correlation with the attacker's view of the relative weakness of the guardian. According to Yar, cyber criminals view the vulnerability and number of victims as relatively constant across time, as the Internet is so large that there are presumably always vulnerable targets to attack [27]. As such, only the perceived capability of the system's guardian should influence the likelihood of an attack.

It is important to note that patterns of intrusions potentially could undermine the application of RAT; a near uniform distribution of intrusions would illustrate that attacker behavior is not influenced by fluctuations in victim vulnerability, guardian capability, or attacker availability. As such, we define our null hypotheses as follows:

Null Hypothesis: A uniform distribution of intrusions will be observed throughout the hours of the day.

With respect to local (EST) time, we hypothesize that attackers will view guardians as least capable at local night, when oversight from system administrators and other users is at its lowest, and thus attack at this time. As such, we expect to see a defined pattern of attacks in which the most intrusions are seen in this block of time. Consequently, we expect to see the fewest number of intrusions during the local day. Importantly, we affirm our assumptions with Yar [27] that guardianship is a more important factor than victim availability in determining attack volume, and as a consequence, test an opposite hypothesis as did Maimon [20].

Hypothesis 1: In the local (EST) time zone, the lowest number of intrusions will occur during the local daytime hours.

We next examine time of attack relative to the time zone of the source of the attack. We define native time as the time seen by the attacker in the time zone to which the IP address of the attack geolocates. We hypothesize that an attacker would have the greatest opportunity to attack during their native evening and nighttime. This expectation is in alignment with patterns of traditional crime [7], [8], [9], [10], operating under the assumption that the average cybercriminal is otherwise occupied during working day hours, and would have more time to explore and exploit systems in the evening and night. As such, we expect that the frequency of intrusions will be higher during these times based on the native time of the attacker.

Hypothesis 2: In the native time of the attack, the highest number of intrusions will occur during the evening and night hours.

III. Method

A. Data Processing

Data used for this paper was provided by a set of target high interaction honeypots constructed by Sobesto in 2011 [28]. The honeynet testbed, a large network of honeypots, was connected to the Internet on September 8th, 2011, and the framework remained stable and identical to its original state for the duration of its lifetime [28]. In this framework, a deployment refers to the lifetime of a honeypot, which begins when an attacker first connects. Deployments last for 30 days after the initial connection, at which time the honeypot is erased and reset to await the next attack. A session is created whenever an attacker gains entry to a honeypot, whether by brute force attack or successful login to SSH. Many sessions could be created within each individual honeypot deployment [28].

It is important to note that the only cyber attacks considered in this study were SSH intrusions into a Linux-based system. There are many different avenues for cyber attacks, and SSH is only one of them. As such, attackers that specialize or even participate in other forms of attacks, including web-based exploitation, DDOS attacks, or others, were outside of the scope of our analysis. Furthermore, as the honeynet system was entirely Linux based, intrusions designated for Windows, OS X, or other operating systems were out of scope.

Intrusion data was queried and delivered on April 29th, 2015, and no intrusions after that date were considered. Numerous experiments were run on this framework, but only the intrusions logged to the control groups of each experiment were included in our analysis, as these honeypots were identically configured throughout the duration of the framework. The delivered dataset contained one entry for each session opened on a honeypot. Each entry has a number of relevant fields, including the deployment ID of the honeypot, the IP address of the attacker, the time of day (EST) the intrusion occurred, and the country to which the framework geolocated the IP address. In total, 20,773 intrusions were considered in the scope of our analysis over a period of 1,256 days: 38 intrusions in 2011, 1,253 in 2012, 4,859 in 2013, 14,181 in 2014, and 427 in 2015. Note that the data collection periods in 2011 and 2015 are incomplete since the collection started on September 8th, 2011 and the dataset was provided on April 29th, 2015.

To examine the effect of RAT, the local (EST) and native time of the intrusion were analyzed. To determine the time of day of the intrusion from the attacker's time zone, the recorded IP address was geocoded to a longitude and latitude using

MaxMind's GeoLiteCities database [29]. In total, 15 IP addresses could not be geolocated, and as such they were disregarded from analysis, leaving 20,758 intrusions total. Drawing from the methodology of Wang [24], we use a public service [30] to determine the time zone from this longitude and latitude, with which the native time of the attacker was calculated. The local and calculated native times were analyzed in different aggregate blocks of time. The number of intrusions was analyzed by blocks of two hours, four hours, and six hours to observe a larger pattern.

B. Statistical Analysis

In this section, we discuss the various approaches we used to analyze our dataset. We first assessed whether we could apply a one-way ANOVA on the blocks of time. Since this method requires normally distributed data, we applied the Anderson-Darling test for normality. The number of intrusions was examined by hour block per day, but as most days did not have intrusions in every block of time, a high number of zeros was present in the dataset (68% of data). Thus the dataset is highly skewed and we anticipated it would violate the normality assumption, which was confirmed by the p-values lower than 0.001 with the Anderson-Darling test. We then applied several transformations on the dataset so that the transformed data would not violate the normality assumption. Applied transformations included shift-log transformation and Box-Cox transformation. For all the applied transformations, the Anderson-Darling test gave p-values lower than 0.001, which means that the transformed data also violate the normality assumption.

We then analyzed the non-transformed data using non-parametric approaches. Non-parametric tests such as the Kruskal-Wallis test do not require data to be normally distributed and assume identical distributions for different groups. Moreover, these tests consider different distributions for groups with non-equal standard deviations [31]. Since, the difference within variances of variables in each category was not significant (i.e., variances of the number of intrusions in each of the time blocks), we could apply the Kruskal-Wallis test. In addition, due to the large numbers of comparisons, using the Bonferroni correction was not possible since it would lead to very conservative results. To correct the p-values for pairwise comparisons, we applied a post-hoc Kruskal-Wallis test [32] ("pgirmess" package in R).

Finally, to confirm that our results from the Kruskal-Wallis test are consistent due to the large and positive skewness as well as overdispersion of our dataset ("overdispersion" is a situation in which the variability of individual counts may exceed the value expected from the model [33]), we used the negative binomial regression model to compare the mean number of intrusions between time blocks. Negative binomial regression models offer techniques to handle zero-inflated variables with large positive skews. We identified a time block as the dependent variable in the regression model and considered other time blocks as independent variables. If the coefficient of an independent time block is not significant, the expected value of both the dependent and independent blocks are not significantly different.

In all the performed tests, the null hypothesis states that the mean values of the time blocks are not significantly different from each other. Note that in the rest of the paper, "statistically significant" means that our analysis is significant at 95%.

IV. Results

A. Descriptive Statistics

From September 8th, 2011 to April 29th, 2015, 20,758 intrusions were observed. The first three rows of Table I show the number and percentage of intrusions for the top three countries. A number of factors could potentially impact our results since they are based on the location of the attack origin. More precisely, we will address the issue of botnets and cloud hosting platforms ("CHP").

First, we address the issue of botnets. We define a connection as originating from a botnet if it opened a session in a honeypot deployment that saw more than 60 total sessions. Once a deployment was created, in order for a second unique user to create a session, it would have to have either been given the system's login credentials, or gain entry by brute force attack into the honeypot. However, we make the assumption that it is unlikely for over 60 attackers to compromise the same deployment and thus, in these cases, it is assumed that the credentials were shared. Such credential sharing across a multitude of IP addresses is indicative of either a botnet, a distributed network of end-hosts that are capable of information and command dissemination [34]; or of a tool capable of IP address randomization [35]. The number 60 was chosen as the cutoff for the bot classification, as it is the point at which the number of sessions per deployment levels off. Figure 1 shows the graphical distribution of session per deployment. Note that to emphasize the curve, the x-axis is cutoff after 750 deployments. The black dot marks 60 sessions per deployment.

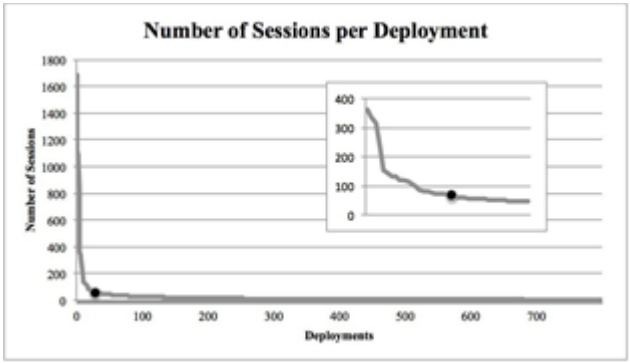

Fig. 1. Number of Sessions per Deployment

The fourth and fifth columns of Table I show the number and percentage of intrusions when intrusions due to botnets were removed. About 33% of the total number of intrusions were due to botnets, but this ratio varies by country. For the US, only 24% of the intrusions were not from botnets. For China, this percentage is 90%. For Romania, it reached 99%. For the other countries, 90% of the intrusions did not come from botnets.

Second, we address the issue of intrusions originating from cloud hosting platforms. The goal is to reduce the impact of cloud hosting platforms on geolocation. Many of these platforms host many of their IP addresses in the United States [29], thereby

masking the original location of the attacker and rendering geolocation inaccurate. As such, datasets were generated for analysis with and without the IP addresses that fell within the IP ranges of Amazon's EC2 Web Services, Microsoft's Azure Cloud Computing Services, and OVH Hosting Public Cloud - three highly popular cloud hosting services. Note that although the DigitalOcean's Cloud, Linode Computing Cloud, and the Aliyun Cloud Engine are also popular hosting services, they do not publish their IP ranges, so intrusions originating from their platforms could not be identified.

The sixth and seventh columns of Table I show the number and percentage of intrusions when intrusions due to cloud hosting platforms were removed. About 33% of the number of intrusions were due to attacks from cloud hosting platforms. For the US, only 18% of the intrusions were not launched from cloud hosting platforms. As all of the IP ranges associated with the cloud hosting platforms were geolocated to the United States, for all other countries, no intrusions were associated with cloud hosting platforms.

B. Local Time Results

In this section, we discuss the results of the intrusions over time observed on the target computers, analyzed in twelve two-hour blocks. The mean number of intrusions in (16:00-18:00), (20:00-22:00), (22:00-0:00), and (0:00-2:00) blocks are the lowest, and no significant difference exists between the means of them ($p=0.08$). Next, we test if there exists a significant difference between the means of any other block. Compared to the 8:00-10:00 block, we find that no other block shows a significant deviation from this mean ($p=0.07$). Repeating this analysis for the mean number of intrusions observed in the (16:00-18:00), (0:00-2:00) and (18:00-20:00) blocks similarly shows no significant deviation amongst them. We rely on this pattern of statistical tests for each subsequent analysis.

We then expanded our analysis to four hour blocks and six hour blocks in an attempt to make observations at a macro level; similarly, we found statistically significant differences in both cases ($p<0.01$). We also considered the impact of botnets in our analysis. When removing intrusions attributed to botnets, we still found statistically significant differences in the two hour blocks, four hour blocks and six hour blocks ($p<0.01$). We then analyzed the impact of intrusions launched from cloud hosting platforms. As all of the attacks from the cloud hosting platforms (CHP) were geolocated to the United States, the number of intrusions linked to cloud hosting platforms artificially inflated the number of attacks linked to the United States. With these intrusions removed, we still found significant differences within the two-hour blocks, four hours blocks, and six hours block ($p<0.01$). As CHP artificially inflated attacks originating from the United States, China is actually the country from which the largest number of intrusions originated, and we wanted to confirm that the intrusions geolocated to China did not alter our results. When considering the intrusions without China, we also found significant differences in the two-hour blocks, four-hour blocks, and six-hour block ($p<0.01$). The results of the number and percentage of intrusions overall as measured in two hour blocks are shown in Table II with sessions from botnets and cloud hosting platforms removed. ("CHP" in columns six and seven refers to Cloud Hosting Platforms.)

The mean number of observed intrusions between (16:00-18:00), (20:00-22:00), (22:00-0:00), and (0:00-2:00) are the lowest, and no significant difference was found between them. In addition, the mean number of intrusions in these blocks is shown to be significantly different from the mean number of intrusions observed during the rest of the day. In considering the rest of the day, no significant difference is found between the mean numbers of intrusions recorded to any other block. Overall, the lowest number of intrusions is observed between 16:00 and 2:00. The distribution of intrusions over time from botnets or cloud hosting platforms is high but nearly constant. Indeed, in removing the data associated with botnets or cloud hosting platforms from the analysis, the pattern of daily intrusions becomes more defined. The mean number of intrusions observed each two hour block between 16:00 and 2:00 is statistically significantly lower, and the means of the (16:00-18:00), (18:00-20:00), and (0:00-2:00) blocks are greater than the number of intrusions observed within (20:00-0:00). In addition, the mean number of intrusions in these blocks is statistically significantly lower than the mean number of intrusions seen during the rest of the day; however, no other block in the rest of day is not significantly different from any other. These results are the same after excluding intrusions linked to botnets or cloud hosting platforms.

Given the high number of intrusions linked to an IP address in China, we analyzed the dataset when removing the intrusions linked to China and when only considering intrusions originating from China. Columns 2-5 of Table IV show the number and percentage of intrusions after removing intrusions that originated from China and isolating intrusions linked to China. Although the pattern is less defined after removing attacks from China, a statistically significant pattern emerges nevertheless ($p<0.01$); the fewest attacks were observed between 22:00 and 2:00, and no significant difference was found between any of the blocks in this range. No statistically significant difference was found between the numbers of intrusions into any other block. These results are the same before excluding intrusions from China.

When focusing only on intrusions originating from China, the lowest number of intrusions was recorded between 16:00 and 0:00, and no statistically significant difference is found between the mean number of intrusions recorded in any of these blocks. The number of intrusions received between 4:00 to 8:00 is the highest, and the means are not significantly different between the blocks. Additionally, no statistically significant difference is found amongst the rest of the blocks. We observe three distinct groups of time blocks with significance. These analyzes confirm the existence of a defined pattern in the number of intrusions linked to China. Interestingly, the patterns observed in the exclusively China dataset mirror the results found in the entire dataset, indicating that the pattern of intrusions originating from China had a strong influence on the overall pattern of intrusions seen.

C. Native Time Results

In this section we discuss the results of the intrusion times based on the native time of the attacker. The mean number of intrusions over time was analyzed in twelve two hour blocks. We found that there were statistically significant differences

among the time blocks ($p<0.01$). We expanded our analysis to four hour blocks and six hour blocks, and similarly found statistically significant differences in both cases ($p<0.01$, $p=0.04$). We also considered the impact of botnets in our analysis. When removing intrusions attributed to botnets, we still found statistically significant differences in the two hour blocks ($p<0.01$), four hours blocks ($p<0.01$) and six hours blocks ($p=0.03$). We then analyzed the impact of intrusions launched from cloud hosting platforms, and still found statistically significant differences in the two hour blocks, four hours blocks and six hours blocks ($p<0.01$). As aforementioned, it is important to confirm that the volume of intrusions that originated in China did not alter our results. When considering the intrusions without China, we also found statistically significant differences in the two hour blocks ($p<0.01$), four hours blocks ($p<0.01$), but not six hours block ($p=0.06$). The dataset with China removed examined across six hour blocks was hence the only analysis with insignificant results; this is likely because the periods of lower attack volume are split evenly across two blocks. The results of the number and percentage of intrusions overall for two hour blocks are shown in Table III without the effects of botnets and cloud hosting platforms. (”CHP” in columns six and seven refers to Cloud Hosting Platforms.)

The mean number of intrusions received during (4:00-6:00), (6:00-8:00), (8:00-10:00) and (12:00-14:00) are statistically significantly lower than the rest of the blocks, and no statistically significant difference is found between them. No significant difference is found between the mean number of intrusions recorded during any other blocks of time.

Similarly to our findings in local time, the distribution of intrusions over time from botnets or cloud hosting platforms is high but nearly constant. Indeed, in removing the data associated with botnets or cloud hosting platforms from the analysis, the pattern of daily intrusions becomes more defined. Each block between 4:00 and 10:00 saw the lowest number of intrusions, and no significant difference is found between them. The number of intrusions observed between 12:00 and 14:00 is higher than these blocks, but saw fewer intrusions than the rest of the blocks of time did. No statistically significant difference was found between the number of intrusions received in any other block.

Given the high number of intrusions linked to an IP address in China, we analyzed the dataset when removing the intrusions linked to China and when only considering intrusions associated with China. Columns 6-9 of Table IV show the number and percentage of intrusions when removing intrusions where the attack originated in China and when only considering intrusions linked to China. Although the pattern is less defined, a statistically significant pattern emerges nevertheless ($p<0.01$). When removing intrusions linked to China, we observe the least number of intrusions between 4:00 and 10:00. It is important to note that again no statistically significant difference is found in the number of intrusions recorded in each block between 4:00 and 10:00, and between any other block. Interestingly, this pattern is very similar to the pattern identified after removing the intrusions linked to botnets or cloud hosting platforms.

We now focus only on intrusions associated with China. As China is 12-13 hours ahead of EST (depending on Daylight Savings Time), we expect to see the aforementioned pattern discovered in the local analysis approximately shifted by 12 hours. Indeed, the lowest number of intrusions is recorded between 4:00 and 14:00, and no statistically significant difference is found between these blocks of time. We also observe spikes in the number of intrusions from 16:00 to 18:00 and from 20:00 to 22:00. No statistically significant difference is found between the means of any other block of time.

V. DISCUSSION

A. Discussion of Local Results

We observe statistically significant patterns in which the majority of intrusions occur from 2AM to 4PM local time. In examining intrusions originating from China in isolation, peak number of attacks were observed from 12AM to 4PM. As we expected to see the fewest number of intrusions during the daytime hours, we reject our first hypothesis. However, given that a uniform distribution of intrusions was not observed, we reject our null hypothesis as well. These results still display patterns supported by RAT, and our results support conclusions drawn by Maimon [20].

In alignment with RAT, there are two potential interpretations of these results. First, attackers may view guardianship as weakest during the height of the system load; rather than avoiding peak times of usage, attackers instead attempt to ”blend in” with heightened levels of traffic. Attackers may believe that system guardians would be otherwise occupied or overwhelmed by the daily traffic to notice an ongoing attack. In this case, guardian strength is a major determining factor in the decision to attack. Such an interpretation can be linked to similar spikes in theft rates in the wake of natural disasters, as typical guardians (the police) are otherwise preoccupied with relief efforts to effectively act as guardians [36].

Although we initially made the assumption, supported by Yar in 2006 [27], that attackers would view victim availability as near constant, this assumption may not be accurate. Instead, attackers may expect victims to be most available during the day hours, regardless of the relative strength of the guardian, and thus prefer to attack during this time. This would show that variability in guardianship is equal to or less important than victim availability in determining attack volume. Such an interpretation is supported by Maimon [20].

In examining the data from intrusions geolocated to a nation other than China, we observe a statistically significant pattern in which the highest number of intrusions were observed between the hours of 2AM to 10PM. However, such a wide range indicates that no singular defined period of heightened intrusion volume exists; rather we identify only a period of substantially decreased intrusions (from 10PM to 2AM). As such, although these results can be attributed to weakened guardianship or higher victim availability, the large duration of sustained heightened attack volume prohibit finer analysis of the motivating factors behind such a distribution.

TABLE I. NUMBER OF INTRUSIONS BY COUNTRY

Country	Number of Intrusions	Percent of Intrusions	Number of Intrusions w/o Bots	Percent of Total w/o Bots	Number of Intrusion w/o CHP	Percent of Total w/o CHP	Number of Intrusions w/o Bots and w/o CHP	Percent of Total Intrusions w/o Bots and w/o CHP
United States	8021	38.6%	1939	9.3%	1439	6.9%	1422	6.8%
China	5825	28.0%	5222	25.2%	5825	28.0%	5222	25.2%
Romania	1492	7.2%	1472	7.1%	1492	7.2%	1472	7.1%
Other	5420	26.1%	4896	23.6%	5400	26.0%	4878	23.5%
Total	20758	100%	13529	65.5%	14156	68.2%	12994	62.6%

TABLE II. LOCAL ANALYSIS

Local Hour Blocks	Number of Intrusions	Percent of Intrusions	Number of Intrusions w/o Bots	Percent of Intrusions w/o Bots	Number of Intrusions w/o CHP	Percent of Intrusions w/o CHP
0:00 - 1:59	1538	7.41%	963	7.12%	978	6.91%
2:00 - 3:59	1714	8.26%	1140	8.43%	1177	8.31%
4:00 - 5:59	1982	9.55%	1333	9.85%	1404	9.92%
6:00 - 7:59	1942	9.36%	1306	9.65%	1400	9.89%
8:00 - 9:59	1791	8.63%	1181	8.73%	1264	8.93%
10:00 - 11:59	1968	9.48%	1403	10.37%	1444	10.2%
12:00 - 13:59	1980	9.54%	1358	10.04%	1438	10.16%
14:00 - 15:59	1934	9.32%	1321	9.76%	1369	9.67%
16:00-17:59	1531	7.38%	895	6.62%	967	6.83%
18:00 - 19:59	1610	7.76%	1021	7.55%	1064	7.52%
20:00 - 21:59	1389	6.69%	809	5.98%	829	5.86%
22:00 - 23:59	1379	6.64%	799	5.91%	822	5.81%
Total	20758	100%	13529	65.2%	14156	68.2%

TABLE III. NATIVE ANALYSIS

Native Hour Blocks	Number of Intrusions	Percent of Intrusions	Number of Intrusions w/o Bots	Percent of Intrusions w/o Bots	Number of Intrusions w/o CHP	Percent of Intrusions w/o CHP
0:00 - 1:59	1879	9.05%	1249	9.23%	1311	9.26%
2:00 - 3:59	1785	8.6%	1128	8.34%	1202	8.49%
4:00 - 5:59	1415	6.82%	778	5.75%	813	5.74%
6:00 - 7:59	1361	6.56%	760	5.62%	831	5.87%
8:00 - 9:59	1444	6.96%	914	6.76%	938	6.63%
10:00 - 11:59	1717	8.27%	1160	8.57%	1193	8.43%
12:00 - 13:59	1618	7.79%	1075	7.95%	1100	7.77%
14:00 - 15:59	2043	9.84%	1450	10.72%	1483	10.48%
16:00 - 17:59	1784	8.59%	1196	8.84%	1272	8.99%
18:00 - 19:59	1966	9.47%	1318	9.74%	1390	9.82%
20:00 - 21:59	1878	9.05%	1232	9.11%	1308	9.24%
22:00 - 23:59	1868	9.0%	1269	9.38%	1315	9.29%
Total	20758	100%	13529	65.2%	14156	68.2%

TABLE IV. CHINA ANALYSIS

Hour Blocks	Number of Intrusions Local China	Percent of Intrusions Local China	Number of Intrusions Local w/o China	Percent of Intrusions Local w/o China	Number of Intrusions Native China	Percent of Intrusions Native China	Number of Intrusions Native w/o China	Percent of Intrusions Native w/o China
0:00 - 1:59	566	9.72%	972	6.51%	566	9.72%	1313	8.79%
2:00 - 3:59	528	9.06%	1186	7.94%	542	9.30%	1243	8.32%
4:00 - 5:59	641	11.00%	1341	8.98%	357	6.13%	1058	7.08%
6:00 - 7:59	633	10.87%	1309	8.77%	354	6.08%	1007	6.74%
8:00 - 9:59	517	8.88%	1274	8.53%	355	6.09%	1089	7.29%
10:00 - 11:59	554	9.51%	1414	9.47%	306	5.25%	1411	9.45%
12:00 - 13:59	571	9.80%	1409	9.44%	353	6.06%	1265	8.47%
14:00 - 15:59	466	8.00%	1468	9.83%	683	11.73%	1360	9.11%
16:00 - 17:59	321	5.51%	1210	8.10%	588	10.09%	1196	8.01%
18:00 - 19:59	381	6.54%	1229	8.23%	676	11.61%	1290	8.64%
20:00 - 21:59	338	5.80%	1051	7.04%	509	8.74%	1369	9.17%
22:00 - 23:59	309	5.30%	1070	7.17%	536	9.20%	1332	8.92%
Total	5825	100%	14933	100%	5825	100%	14933	100%

B. Discussion of Native Results

We identified statistically significant patterns, in which the lowest number of intrusions is launched from 4AM to 2PM native time. Intrusions originating from China showed a more defined but similar significant daily pattern, with the lowest number of intrusions between 4AM and 2PM, and significant spikes from 4PM to 10PM. As we expected to see the highest number of attacks during the evening hours, we accept our second hypothesis and reject our null hypothesis.

These results closely align with the patterns of physical crime, and lend themselves well to interpretation from RAT. Most intrusions were launched during non-work hours, when individuals would presumably have the most opportunity to commit an intrusion. We affirm this conclusion with afore-mentioned patterns of physical crime [7], [8], [9], [10]; the assumption is that most attackers are otherwise employed or occupied during traditional workday hours. These results demonstrate the applicability of RAT to cyberspace in the native time of the attacker.

We also observe a statistically significant pattern of attack in the dataset with attacks originating from China removed. The highest number of intrusions were observed between the hours of 10AM to 4AM. However, as identified in the local analysis, such a wide range indicates the lack of a defined period of heightened intrusion. As such, although the period of substantially decreased intrusions (4AM to 10AM) can be attributed to attacker availability, the large number of hours of heightened attack volume again prohibit finer analysis.

C. Implication of Results

Throughout our methodology, we conduct analysis in the local time and the native time separately to isolate specific patterns of intrusion from the victim and the attacker. However, it is crucial to note that local and native times are entirely interrelated, and the juxtaposition of the local and native analysis sheds valuable insight into which aspects of RAT influence attacker behavior the most. For example, should analysis in the local time demonstrate a stronger statistically significant pattern than in native time, we can deduce that attackers around the world are tailoring their time of attack specifically to avoid or concentrate on a specific range of local times. In this case, RAT would predict that victim availability or strength of guardianship are most important factors in the decision to commit a cyber intrusion. However, should patterns uncovered in the native time analysis prove more defined than in local time, we can assume that attackers around the world are attacking in times most convenient for them, and that attacker availability has the greatest influence on the occurrence of cybercrime.

Interestingly, in comparing the results from local and native analysis of the entire dataset, both display patterns of comparable levels of statistical significance with relatively equal length of hours of attack minima. This holds true in examining the entire dataset without the intrusions originating in cloud hosting platforms, without the intrusions due to botnets, and without the intrusions where the attack originated in China. As such, for these datasets, we cannot determine which aspect of RAT has the strongest influence on the occurrence of cyber intrusion. However, in examining the subset of attacks originating from China, we observe a notable spike in intrusions from 4PM to 10PM; such a tight peak is not observed with significance in the local time. RAT would thus predict that actors who launched attacks geolocated to China are most influenced by attacker availability in deciding when to attack. The difference in accuracy between native time statistical significance between China and the rest of the dataset is discussed at greater length in the limitations section.

Our results have important implications for cybersecurity policy, and particularly underscore the importance of improved guardianship in the cyber world. As asserted by Yar in 2006, the capability of offenders cannot be controlled and it is near impossible to guarantee a vulnerability-free system [27]. Since most attacks occurred during the day, increasing the number of Information Technology (IT) security specialists during these hours to protect and monitor the system could help mollify the damages done by cyber intrusions.

VI. LIMITATIONS

A number of limitations are present in this study. First, the geolocation of an attacker's IP address does not necessarily correspond to the actual location of the individual or machine that initiated the attack. Hence, all results discussed pertain only to the last machine that connected to the target computers, rather than an individual person. Additionally, it is possible that attacks received are from a bot, script, or automated tool. However, every intruding IP address originated from an attacker, whether through a proxy or direct connection, by means of a person explicitly typing the commands or the person who launched said attacking tool. Each of these attack types can easily be tailored to act only during certain times of the day, and each were configured or deployed by an attacker for an expressly malicious purpose. As such, any incoming IP address is assumed to be a deliberate attacker, regardless if the attack originated from a script, a person, or a tool.

Geolocation services are incapable of achieving perfect accuracy in geocoding each IP address to a specific latitude and longitude, which can impact the calculated native time of attack. A 2011 analysis of MaxMind GeoLite Services, the geolocation service used in this study, showed that GeoLiteCities is the most accurate service (on average) as compared to competitive options, with an average accuracy rating of 95.8% [37]. Furthermore, MaxMind reports that their geolocation databases are accurate up to 92% within 250km of the reported city, and at least 85% accurate within 100km of the reported city. When geolocated at the country level, MaxMind reports almost 99% accuracy [38]. Most native time analysis was geocoded to a longitude and latitude instead of a country; however, China only has one official time zone, and therefore we can rely on the provided country level geolocation to greatly improve accuracy.

The data used spans over 4 years, which means that IP addresses could have been reassigned since. Unfortunately, the provided data only geolocated the IP address to the country level, and since many countries span more than one time zone, all of the IP addresses had to be re-geolocated during processing. Only 7.6% of the calculated locations differed from the country

they were geolocated to in the initial data collection, suggesting that the time zones of many of the IP addresses did not drastically change in the interim. Inaccuracies in the geolocation could have negatively impacted the native time analysis, weakening the significance of the daily patterns discovered. Moreover, such inaccuracy could explain why the pattern from China was distinctly more defined than the rest of the native times. China has only one timezone, and as the original dataset performed a country-level IP geolocation at the time of intrusion, native times calculated based on intrusions to China are inherently more accurate, as the rest of the intrusions had to be geolocated to a timezone despite potential IP reallocation.

We acknowledge Yar's concerns about the difficulty of translating RAT's reliance on temporal convergence to cyberspace [39]. However, due to the nature of cyber intrusions, temporal convergence is guaranteed, as in order for a successful intrusion to occur, an attacker must converge with a victim system at the same time on the same host.

VII. CONCLUSIONS

In this paper, we proposed an application of Routine Activity Theory (RAT) to cyber intrusions. Our hypothesis postulated that attackers would choose to attack during the time of day during which they would have the most opportunity to attack, when the guardian of the victim system was least capable, and when the victim system was most vulnerable. We predicted that this would occur in a similar pattern with normal crime, and that more intrusions would occur at night, when the attackers would view guardianship of the systems as the lowest and vulnerability the highest.

We analyzed intrusion data recorded on targeted systems over a four year period. We showed a statistically significant attack pattern at the granularity of two hour blocks of time, and made advances in exploring these patterns of attacks from the time zone of the attacker. We identified that the mean number of intrusions is statistically significantly lowest between the hours of 4PM to 2AM in the local time zone, and 2AM to 4PM in the native time zone. We also made recommendations concerning the specific hours in which systems are most at risk, and when improved guardianship is most needed. In addition, we showed statistically significant patterns in the time of attacks based on the native time zone of the attacker, shedding valuable insight into cybercriminal behavior. In isolating attacks originating from China, we discovered patterns of attack with strong significance.

ACKNOWLEDGMENT

This research is supported by NSF award #1223634.

REFERENCES

[1] Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American sociological review, 588-608.
[2] Snyder, H.N. (1999). Juvenile offenders and victims: 1999 national report. Washington, D.C.: National Center for Juvenile Justice.
[3] Snyder, H. N. and Sickmund, M. (2006), Juvenile Offenders and Victims: 2006 National Report. Washington, DC: US Department of Justice, Office of Justice Programs, Office of Juveniles Justice and Delinquency Prevention
[4] Wiebe, D. J., Meeker, J. W., & Vila, B. (1999). Hourly trends of gang crime incidents, 1995-1998. Fact Sheet. Office of Justice Programs. Office of Juvenile Justice and Delinquency Prevention. Washington, DC: US Department of Justice.
[5] Felson, M. and Boba, R. (2010), Crime and Everyday Life, 4th edn. Sage Publications
[6] Roman, Caterina Gouvis. (2002) "Analytical Strategy." Schools as Generators of Crime: Routine Activities and the Sociology of Place. N.p.: U.S. Department of Justice, 2002. 72-75. Print.
[7] Kuang, Cliff. (2011)"When Do Criminals Prowl The Streets?" Trulia, 14 July 2011.
[8] Catalano, S. M. (2010). *Victimization during household burglary*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.
[9] Kvaraceus, W. (1945). Juvenile Delinquency and the School. New York: World Book Company
[10] Gottfredson, D. C., & Soulé, D. A. (2005). The timing of property crime, violent crime, and substance use among juveniles. *Journal of Research in Crime and Delinquency*, *42*(1), 110-120.
[11] Bossler, A. M. and Holt, T. J. (2009), 'On-Line Activities, Guardianship, and Malware Infection: An Examination of Routine Activities Theory', International Journal of Cyber Criminology, 3: 400–20.
[12] Hutchings, A., & Hayes, H. (2008). Routine activity theory and phishing victimisation: Who gets caught in the net. *Current Issues Crim. Just.*, *20*, 433.
[13] Kigerl, A. (2012) "Routine Activity Theory and the Determinants of High Cybercrime Countries", Social Science Computer Review.
[14] Pratt, T. C., Holtfreter, K. and Reisig, M. D. (2010), 'Routine Online Activity and Internet Fraud Targeting: Extending the Generality of Routine Activity Theory', Journal of Research in Crime and Delinquency, 47: 267–96.
[15] Nhan, J., Kinkade, P. and Burns, R. (2009), 'Finding a Pot of Gold at the End of an Internet Rainbow: Further Examination of Fraudulent Email Solicitation', International Journal of Cyber Criminology, 3: 452–75.
[16] Chon, Steven (2014). Routine Activity Theory and Cybercrime: What about Offender Resources?. ANU Cybercrime Observatory, p 2-6.
[17] Grabosky, P. N. (2001), 'Virtual Criminology: Old Wine in New Bottles?', Social and Legal Studies, 10: 243–9.
[18] Bolden, Micah-Sage. (2014). Theorizing Cybercrime: Applying Routine Activities Theory.
[19] Lin, Jon. (2014)."Visualizing SASL/POP3/IMAP Automated Dictionary Attacks." Tesuji, 14 July 2014.
[20] Maimon, D., Kamerdze, A., Cukier, M., & Sobesto, B. (2013). Daily trends and origin of computer-focused crimes against a large university computer network an application of the routine-activities and lifestyle Perspective. *British Journal of Criminology*, azs067.
[21] de Silva, E. (Ed.). (2015). National security and counterintelligence in the era of cyber espionage. IGI Global.
[22] Singh, A. N., & Joshi, R. C. (2011). A honeypot system for efficient capture and analysis of network attack traffic. In *Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on* (pp. 514-519). IEEE.
[23] Padmanabhan, V. N., & Subramanian, L. (2001, August). An investigation of geographic mapping techniques for Internet hosts. In *ACM SIGCOMM Computer Communication Review* (Vol. 31, No. 4, pp. 173-185). ACM.
[24] Wang, Y., Burgener, D., Flores, M., Kuzmanovic, A., & Huang, C. (2011, March). Towards Street-Level Client-Independent IP Geolocation. In *NSDI* (Vol. 11, pp. 27-27).
[25] Scarfone, K., & Hoffman, P. (2009). Guidelines on firewalls and firewall policy. *NIST Special Publication*, *800*, 41.
[26] Mitnick, K. D., & Simon, W. L. (2002). The art of deception: Controlling the human element of security. John Wiley & Sons.
[27] Yar, M. (2006). Cybercrime and Society. Sage Publications Ltd.
[28] Sobesto, B. (2015). Empirical studies based on honeypots for characterizing attackers behavior (Doctoral dissertation).
[29] MaxMind GeoLiteCities. https://dev.maxmind.com/geoip/legacy/geolite/
[30] Geonames. http://www.geonames.org/
[31] Dalgaard, P. (2008). Analysis of variance and the Kruskal–Wallis test. *Introductory Statistics with R*, 127-143.
[32] Siegel, S., & Castellan, N. J. (1981). J.(1988). Nonparametric Statistics for the Behavioral Sciences. *McGraw-HiU Book Company, New York*.
[33] Min, Y., & Agresti, A. (2002). Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, *1*(1), 7-33.
[34] Abu Rajab, M., Zarfoss, J., Monrose, F., & Terzis, A. (2006, October). A multifaceted approach to understanding the botnet phenomenon. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement* (pp. 41-52). ACM.
[35] inCloak. https://incloak.com/
[36] Lemieux, F. (2014). The impact of a natural disaster on altruistic behaviour and crime. *Disasters*, *38*(3), 483-499.
[37] Claffy, K., Fomenkov, M., & Huffaker, B. (2011). Geocompare: a comparison of public and commercial geolocation databases. CAIDA Tech Report. N.p., 2011.
[38] Maxmind. (2009). Accuracy of Country-Level IP Geocoding. Web. n.p. December 7, 2010.
[39] Yar, M. (2005). The Novelty of 'Cybercrime' An Assessment in Light of Routine Activity Theory. *European Journal of Criminology*, *2*(4), 407-427.