APPLICABILITY OF ROUGH SET THEORY FOR ANALYSIS OF PHISHING THREATS

Rohit Ramesh, Sanjiban Sekhar Roy, Anvesha Sinha

School of Computing Sciences and Engineering, VIT Vellore

***ABSTRACT*:** Phishing has for long now been a social nuisance. It includes various engineering schemes which are used to obtain discrete and important information from its prey by using a trusted component in electronic communication systems. Given its importance, it is a necessity that we study or analyze this scheme in order to generate a solution. Rough set theory from the past three decades has helped us study various systems with incomplete information domains. It has helped us gain invaluable information by in depth analysis of incomplete information systems. Hence, in this paper we analyze and derive results from a Phishing data set, comprising of domains all over the world and also generic domains using rough set theory. We will then scrutinize some of our results and put forth our conclusion.
***KEYWORDS*:** Phishing; Rough Set Theory; Domain Name.

I. INTRODUCTION

Phishing may be defined as an embezzlement scheme wherein various social engineering schemes are used to obtain discrete information from its prey having a motive to cause losses [OSO11]. This scheme maybe implemented through various channels such as e-mails, false pages etc. Unsolicited or unwarranted emails are not only a social nuisance but may also be extremely dangerous with respect to exploitation of personal warranted data [ARD12]. Various hijacking algorithms are used and convince the user to respond to the trap [***06]. Various technical subterfuge techniques are used to plant CRIMEWARES directly into user forms enabling direct withdrawal of user data. Phishing may also include diverting traffic into corrupted infrastructures through various proxies controlled by phishes [***06].

The Rough set theory can be considered one of the most important conceptions of modern day science. Introduced in the eighties by Pawlak [***06], it is an extension of the set theory solicited for the study and in depth understanding of intelligent systems comprising of incomplete or insufficient information [Paw82a, Paw84, Paw85, Paw82b].The usefulness and scope of the theory has been proved to be unparalleled. This theory may be considered as complementary to other generalizations established in set theory [Paw02]. Rough set theory has found applications in wide spread domains such as environment decisions, banking decisions, healthcare and various other fields [JST05]. It is founded on the assumption that every ascendable object in the universe has some information or data associated with it. These objects that are characterized by the same information are indiscernible. The relation generated so (discernibility relation) forms the mathematical structure of the rough set theory [JST05]. We have used the rough set exploration system, a freeware to derive the results [S+12].

1.1. Basic mathematical overview of Rough Set

The basic mathematical sketch can be defined as [Paw02]. Let S denote a non-empty and finite set, and let $P \subseteq S \times S$ be an equivalence relation on S. The pair constituted by par=(S, P) is known as the approximation space. The equivalence relation P effectively partitions the set S into disjoint sets. Let there be an empty set $\varnothing$ such that the equivalence classes of P and $\varnothing$ are called the atomic sets in the approximation space par=(S,P). The family of all the composing set (including $\varnothing$) is denoted by Co(par) and forms a Boolean algebra. The union of elementary sets is called as a composed set. Let $X \subseteq S$ be an arbitrary set, the available information (equivalence classes of P) will not be enough to give an accurate representation of X. Hence, in such a situation, one may classify X with respect to a pair of lower and higher approximations given by:

$$par\downarrow(X) = \{x \mid [x]_p \subseteq X\} \quad (1)$$

$$par\uparrow(X) = \{x \mid [x]_p \cap X \neq \varnothing\} \quad (2)$$

The lower approximation can be defined as the union of all elementary sets which are subsets of X. The upper approximation may be defined as the union of all elementary sets which have a non-empty intersection with X.

For any subsets, the following properties are satisfied by the lower approximations-

$$par\downarrow(\varnothing) = \varnothing \quad (1)$$

$$par\downarrow(X) \subseteq X \quad (2)$$

$$X \subseteq Y \Rightarrow par\downarrow(X) \subseteq par\downarrow(Y) \quad (3)$$

$$par\downarrow(X \cap Y) = par\downarrow(X) \cap par\downarrow(Y) \quad (4)$$

$$par\downarrow(X \cap Y) \supseteq par\downarrow(X) \subseteq par\downarrow(Y) \quad (5)$$

$$par\downarrow(U) = U \quad (6)$$

$$X \subseteq par\downarrow(par\uparrow(X)) \quad (7)$$

$$par\downarrow(X) \subseteq par\downarrow(par\downarrow(X)) \quad (8)$$

$$par\uparrow(X) \subseteq par\downarrow(par\uparrow(X)) \quad (9)$$

$$par\downarrow(X \bigcup Y) \Rightarrow par\downarrow(X) \bigcup par\downarrow(Y) \quad (10)$$

and the upper approximations satisfy the above conditions-

$$par\uparrow(\varnothing) = \varnothing \quad (1)$$

$$X \subseteq Y \Rightarrow par\uparrow(X) \subseteq par\uparrow(Y) \quad (2)$$

$$par\uparrow(X) = \sim par\downarrow(\sim X) \quad (3)$$

$$par\uparrow(par\downarrow(X)) \subseteq par\downarrow(X) \quad (4)$$

$$par\uparrow(par\uparrow(X)) \subseteq par\uparrow(X) \quad (5)$$

$$par\uparrow(X \bigcup Y) = par\uparrow(X) \bigcup par\uparrow(Y) \quad (6)$$

$$par\uparrow(\varnothing) = \varnothing \quad (7)$$

$$\mathbf{par\uparrow(X \cap Y) \subseteq par\uparrow(X) \cap par\uparrow(Y)} \quad (8)$$

$$par\uparrow(par\downarrow(X)) \subseteq X \quad (9)$$

$$X \subseteq par\uparrow(X)$$

Based on the lower and upper approximations of the set X ⊆ S, the Universe S can be divide into three regions which are disjoint in nature namely, the positive region POS(X), the negative region NEG(X), and the boundary region BON(X) given by [Paw02]:

POS(X) = $par\downarrow(X)$

NEG(X) = S - $par\uparrow(X)$

BON(X) = $par\uparrow(X) - par\downarrow(X)$

II. DATA SET

Table 1. Phishing Data Set [RA12]

OBJECTS 200

TLD	TLD location	No.of unique phishing attacks	unique domain names used	Domains in registry	Score (phishing domains/10000 domains)	Score (attacks per10000 domains)	average uptime	median uptime	no.of malicious domains registered	malicious registrations per 10000 domains
in	India	1690	1351	1674552	8.1	10.1	23:27:07	7:57:25	474	2.8
cn	China	156	120	3502064	0.3	0.4	24:05:44	12:55:60	11	0.0
br	Brazil	4039	3207	2959495	10.8	13.6	22:02:59	6:18:34	24	0.1
us	UnitedStates	626	303	1784000	1.7	3.5	15:14:37	2:54:22	20	0.1
uk	UnitedKingdom	1433	1190	10131000	1.2	1.4	33:54:15	10:42:50	28	0.0
sg	Singapore	89	66	140107	4.7	6.4	55:29:05	17:18:39	0	0.0
ru	RussianFed	1304	829	3860995	2.1	3.4	39:31:53	13:02:11	8	0.0
net	genericTLD	6518	3515	15097524	2.3	4.3	22:26:22	5:19:49	208	0.1
mx	Mexico	328	248	568577	4.4	5.8	31:28:35	11:35:45	1	0.0
fr	France	703	502	2339564	2.1	3.0	31:15:25	13:08:43	2	0.0
gov	USgovernment	303	5000	1784000	6.0	6.0	23:44:24	4:53:26	0	0.0
eu	EuropeanUnion	347	276	3592000	0.8	1.0	20:22:28	7:30:20	7	0.0
es	Spain	381	288	1548844	1.9	2.5	29:40:55	12:10:00	2	0.0
ae	UnitedArabEmirates	29	21	94000	2.2	3.1	54:09:35	6:52:32	0	0.0
aero	sponsoredTLD	3	3	7980	3.8	3.8	23:16:24	12:08:09	0	0.0
au	Australia	1383	1101	1731128	6.4	8.0	24:58:02	5:03:41	2	0.0
ca	Canada	632	521	1926000	2.7	3.3	28:42:33	5:32:05	5	0.0
de	Germany	849	573	15069393	0.4	0.6	32:35:26	12:36:49	6	0.0
info	genericTLD	1764	1514	8153167	1.9	2.2	12:32:24	4:01:44	231	0.3
tr	Turkey	170	138	302008	4.6	5.6	30:33:06	10:58:56	2	0.1

The above data set has been taken from [RA12]. It consists of various domains all throughout the world which also include some specific generic domains also. It has attributes specifying the number of unique phishing attacks, unique domain names used, average uptime, median uptime, the number of domains in registry, scores which denote the number of attacks per thousand domains and the phishing domains per thousand domains. It also contains the number of malicious domains registered and malicious registrations per thousand domains.

III. EXPERIMENTAL RESULTS

3.1. Reduct Calculation

The positive region (explained above) above contains all instances of U that can be segmented to instances

of U/Q (where Q belongs to the resultant attribute) using the information in attributes P where P belongs to the result generating attributes. Considering this notion of the positive region, the rough set degree of dependency of a set of attributes Q on a set of attributes P is denoted by –For P, Q ⊂ A, (where A is a non-empty finite set of attributes Such that a: U → Va and Va is the values attribute a might take).It is noted that Q relies on P on a particular degree D ($0 \leq k \leq 1$) denoted P →D Q if and only if, D = $|POS_P(Q)|/|U|$. The reduction of attributes can be achieved by comparing the equivalence statements Obtained from the attributes [Yao98]. The Attributes are eliminated so that the reduced set of values contains the same predictive correctness of the decision attribute as the original template. Hence, a reduct might be defined as a subset of the minimal cardinality of an attribute conglomerate [Baz96]. The reducts generated by the data set taken for reference are-

Table 2. Reducts

Reducts
{ TLD, "TLD location" }
{ "no.of unique phishing attacks" }
{ TLD, "domains in registry" }
{ TLD, "malicious registraions per 10,000 domains" }
{ TLD, "no.of malicious domains registered" }
{ "TLD location", "domains in registry" }
{ "TLD location", "malicious registraions per 10,000 domains" }
{ "TLD location", "no.of malicious domains registered" }
{ "unique domain names used" }
{ "score(attacks per 10000 domains)" }
{ "domains in registry", "malicious registraions per 10,000 domains" }
{ "domains in registry", "no.of malicious domains registered" }
{ "average uptime" }
{ "median uptime" }

3.2. Dynamic Reduct Calculation

Those well versed with the rough set theory, will know that dynamic reducts are most useful while classifying unseen cases. The dynamic results obtained are considered to be the most stable or reliable reducts obtained. If A is a decision table taken for reference, then we may consider another table B such that its universe is a subset of the universe of A to be a sub table of A. Let S (A) denote all the sub tables of A. Let G ⊆ S (A), then the dynamic reduct DR (A, G) can be denoted as-

REDUCT (A, d) ∩ $\bigcap_{B \in G}$ REDUCT (B, d). Any element of the set DR (A, G) can be called a dynamic reduct of A [Baz96]. The dynamic reducts obtained from the data set are-

Table 3. Dynamic Reducts

Reducts
{ "no.of unique phishing attacks" }
{ "unique domain names used" }
{ "score(attacks per 10000 domains)" }
{ "average uptime" }
{ "median uptime" }
{ TLD, "malicious registraions per 10,000 domains" }
{ TLD, "no.of malicious domains registered" }
{ "domains in registry", "malicious registraions per 10,000 domains" }
{ "domains in registry", "no.of malicious domains registered" }
{ "TLD location", "malicious registraions per 10,000 domains" }
{ "TLD location", "no.of malicious domains registered" }
{ TLD, "TLD location" }
{ TLD, "domains in registry" }
{ "TLD location", "domains in registry" }

3.3. Rule Extraction

Rough set theory has proved to be an important tool in extracting rules from information tables. It is a wide domain which consists of attribute reduction mechanisms and attributes value reduction mechanisms. Rule extraction has often proved to be a problem as it is an N-P hard problem [Yao10]. The basics to rule extraction lies in information systems which can be defined as S = (U,A,V, f) ,where the universe is denoted by U, a finite set comprising of N objects U = $\{x_1, x_2, x_3, \ldots . x_n\}$, A be a set of finite attributes divided into disjoint sets or A = C ∪ D , where C is a set of conditional attributes and D is a set of the various decision attributes.
$V = \bigcup_{q \in A} V_q$ (where V q is a domain of the attribute q), f :U × A→V is the information function (called information function) such that the function f(x,q) ∈ *V*q for every attribute q ∈ A and x ∈ U [X+07, DTH06]. The following rules can be extracted from the reference phishing data-

Table 4. Decision Rules

Decision rules
("TLD location"=in)=>("score (phishing domains/10000 domains)"={0[180]})
("no.of unique phishing attacks"=0)=>("score (phishing domains/10000 domains)"={0[180]})
("unique domain names used"=0)=>("score (phishing domains/10000 domains)"={0[180]})
("domains in registry"=0)=>("score (phishing domains/10000 domains)"={0[180]})
(TLD=in)&("malicious registraions per 10,000 domains"=0)=>("score (phishing domains/10000 domains)"={0[180]})
("score(attacks per 10000 domains)"=0)=>("score (phishing domains/10000 domains)"={0[180]})
("average uptime"=in)=>("score (phishing domains/10000 domains)"={0[180]})
("median uptime"=in)=>("score (phishing domains/10000 domains)"={0[180]})
(TLD=in)&("no.of malicious domains registered"=0)=>("score (phishing domains/10000 domains)"={0[180]})

After discretization of the table [SW92], we obtain the following cuts from the information set-

Table 5. Cuts

Attribute	Size	Description
TLD	0	*
"TLD location"	0	*
"no.of unique phis...	3	1.5; 163.0; 503.5
"unique domain na...	1	1145.5
"domains in regist...	2	2132780.0; 914208...
"malicious registr...	0	*
"score(attacks per...	2	3.45; 5.7
"average uptime"	0	*
"median uptime"	0	*
"no.of malicious d...	0	*

3.4 Graphical Analysis

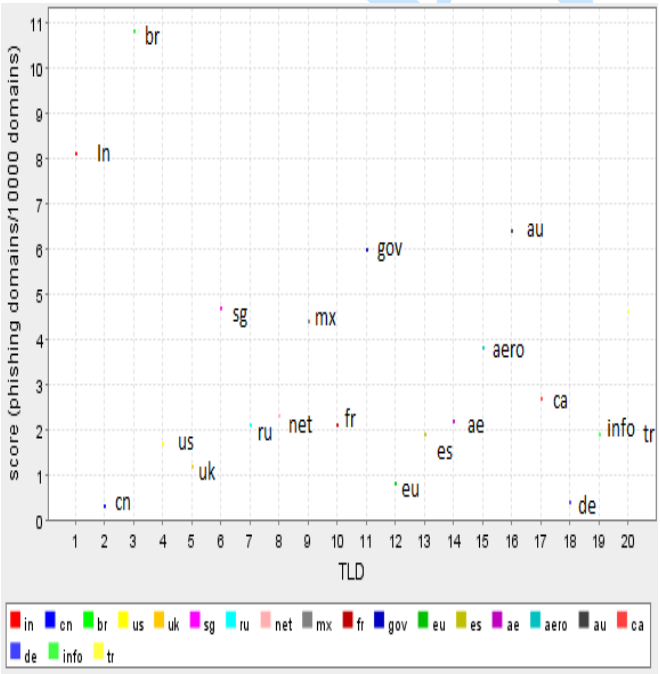

Figure 1. TLD vs. Score (phishing domains/1000 domains)

With the available decision attribute information we also generate a graph comparing the different scores which is the number of phishing domains per 10000 domains. After analyzing the graph, we deduce that Brazil or br has the highest phishing domains per 10000 domains while domains like Canada (cn), Denmark (de) and European Union (eu) are on the lower side. We also notice that generic domains are generally on the lower side and India (in) is ranked second in the number of phishing domains per 10000 domains.

CONCLUSION

We would like to conclude that we have completely analyzed the phishing information data set and successfully derived the reducts, cut sets, rules and dynamic reducts from the data set. We have also analyzed the attributes in the data set and have identified the decision attribute. With this information we have derived a graph TLD vs. score (Phishing domains per 10000 domains) and have found out that Brazil ranks first followed by India according to the y-ordinate and countries like Canada, Denmark lie on the lower side with respect to the y-ordinate.

ACKNOWLEDGMENT

We would like to express our most sincere gratitude to the higher ranking management of VIT University for their most kind help and constant encouragement towards our work.

REFERENCES

[ARD12] **R. M. Amin, J. J.C.H. Ryan, J. R. van Dorp** - *Detecting Targeted Malicious Email*, 1540-7993/12©2012 IEEE.

[Baz96] **J. G. Bazan** - *Dynamic Reducts and Statistical Inference*, 1996.

[DTH06] **R. Dhamija, J. D. Tygar, M. Hearst** - *Why Phishing Works*, CHI 2006 Proceedings, Security, April 22-27, 2006.

[JST05] **R. Jensen, Q. Shen, A. Tuson** - *Finding Rough Set Reducts with SAT*, RSFDGrC 2005, Springer-verlag Berlin Heidelberg, LNAI 3641, pp. 194-203, 2005.

[OSO11] **C.K. Olivo, A. O. Santin, L. S. Oliveira** - *Obtaining the threat model for e-mail Phishing*, Appl. Soft Comput. J. (2011), doi: 10.1016/ j.asoc.2011.06.016.

[Paw02] **Z. Pawlak** - *Rough set theory and its applications*, Journal of telecommunications and information technology, 3/2002.

[Paw82a] **Z. Pawlak** - *Rough sets*, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.

[Paw82b] **Z. Pawlak** - *Rough sets: a new approach to vagueness*, in: Fuzzy Logic for the Management of Uncertainty, edited by L.A. Zadeh And J. Kacprzyk, Eds., John Wiley & Sons, New York, pp. 105-118, 1982.

[Paw84] **Z. Pawlak** - *Rough classification*, International Journal of Man-Machine Studies, 20, pp.469-483, 1984.

[Paw85] **Z. Pawlak** - *Rough sets and fuzzy sets*, Fuzzy Sets and Systems, 17, pp. 99-102, 1985.

[RA12] **R. Rasmussen, G. Aaron** - *APWG Global Phishing Survey: Trends and Domain Name Use 1H2012*, APWG, October, 2012.

[SW92] **D. Slezak, J. Wroblewski** - *Classification Algorithms Based on Linear Combinations of Features*, Springer verlag Berlin Heidelberg, PKDD'99.LNAI 1704, PP. 548-553, 1992.

[S+12] **Sanjiban Sekhar Roy, Rohit Ramesh, Anvesha Sinha, Anupam Gupta** - *Cancer data investigation using variable precision rough set with flexible classification*, Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology ACM, 2012.

[X+07] **E. Xu, Shaocheng Tong, Liangshan Shao, Yongjun Li, Dianke Jiao** - *Rough Set Research on Rule Extraction in Information Table*, Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) 0-7695-2874-0/07 IEEE, 2007.

[Yao10] **Y. Yao** - *Notes on Rough Set Approximations and Associated Measures*, Journal of Zhejiang Ocean University (Natural Science), Vol29, No. 5, pp. 399-410, 2010.

[Yao98] **Y. Y. Yao** - *Generalized rough set models*, in: Rough Sets in Knowledge Discovery, Physica-Verlag, Heidelberg, pp. 286-318, 1998.

[YWL97] **Y. Y. Yao, S. K. M. Wong, T. Y. Lin** - *A Review of Rough Set Models, rough sets and data mining*, Kluwer Academic Publishers, 1997.

[***06] **Anti-Phishing Working Group** - *Phishing Activity Trends Report*, July, 2006.