I.J. Modern Education and Computer Science, **2014, 8, 25-33**
Published Online August 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2014.08.04

Malware Propagation on Social Time Varying Networks: A Comparative Study of Machine Learning Frameworks

A.A. Ojugo, E. Ben-Iwhiwhu, and O. Kekeje
Department of Mathematics/Computer Science, Federal University of Petroleum Resources Effurun, Nigeria
ojugo.arnold@fupre.edu.ng, askofbenese@hotmail.com, debbykekeje@gmail.com

M.O. Yerokun and I.J.B. Iyawa
Department of Computer Sci., Federal College of education (Technical) Asaba, Delta State, Nigeria
agapenexus@hotmail.com, iyawaben@hotmail.com

***Abstract*—**Significant research into the logarithmic analysis of complex networks yields solution to help minimize virus spread and propagation over networks. This task of virus propagation is been a recurring subject, and design of complex models will yield modeling solutions used in a number of events not limited to and include propagation, dataflow, network immunization, resource management, service distribution, adoption of viral marketing etc. Stochastic models are successfully used to predict the virus propagation processes and its effects on networks. The study employs SI-models for independent cascade and the dynamic models with Enron dataset (of e-mail addresses) and presents comparative result using varied machine models. Study samples 25,000 emails of Enron dataset with Entropy and Information Gain computed to address issues of blocking targeting and extent of virus spread on graphs. Study addressed the problem of the expected spread immunization and the expected epidemic spread minimization; but not the epidemic threshold (for space constraint).

***Index Terms*—**Stochastic, immunize, network, graph, SIS, SIR.

I. Introduction

Social networks are dynamic and their normal operation is continually threatened by unethical users often referred to as hackers. These use the spread of harmful and malicious data or program called malware to wreak havoc on network users. Today, Internet has become a high target for the spread of such malwares – as hackers do damage globally, much more easily and faster. Thus, its early detection is imperative to minimize damage caused by it (Desai, 2008).

Malwares are known to attach copies of itself, alters the behaviour as well as modifies attributes of its host machine's files without the user's knowledge (Szor, 2005). Malwares also can sometimes, modify their codes as they infect, to include an evolved copy (Dawkins, 1989). Malwares are grouped into simple, encrypted, polymorphics and metamorphic viruses. Malware are considered malicious software if they consist of codes, scripts, active contents and other software – designed to disrupt or deny operations, gather data that tends to loss of privacy or exploitation, gain unauthorized access to system resources, and other abusive behaviour (Singhal and Raul, 2012). Thus, software codes are considered a malware based on the perceived intentions of its creator rather than any particular feats. Thus, malwares are

Antivirus (AVs) software is designed to detect, prevent and remove all malware such as viruses, worms, Trojans, spyware and adware. AVs detection mechanism are broadly grouped into: (a) signature-based scans for signature, and to evade it – virus makers create new virus strings that can alter their structure while keeping its functionality via code obfuscation method, and (b) code emulation creates sandbox so that files are executed in it while scanning for virus. If virus is detected, it is no threat – as it is running in controlled environment that limit damage to host machine (Singhal and Raul, 2012).

AVs can often impair system performance as any incorrect decision may lead to security breach as it runs at the kernel of the operating system. If an antivirus uses heuristics, its success depends on the right balance between positives and negatives. Malware are no longer execs. Macros can present security risk and antivirus heavily relies on signature-detection. Some malwares evades signature detection effective (Filiolel, 2005) via code obfuscation and encryption methods. Studies show the best AVs can never yield a perfect detection. This is because all scanners can yield a false positive result and identify benign files as malware (Bishop, 2006).

II. Network Topologies

Social networks are used for spread of data – making it easier for users to disseminate useful data as well as viruses. The problem of virus propagation has been a recurring subject and ongoing research notes that every harmful data spread over such networks are considered as malware or viruses as can be interchangeably used; while the process of impeding the spread of such harmful data (malware) over such social network is referred to as network immunization. This aims to prevent the spread of such malwares, protect such networks from virus attacks and control data and sensitive information leakages – while at same time noting that our resources such as vaccination, AVs and influences are quite costly and limited in their capability to discover such malware. With such AVs and vaccinations, users aim to achieve the best effect; while still allocating the least resources possible (Ojugo et al 2013a).

An adversary can wreak more havoc being aware of the propagation model used. In simplest form, a social network is seen as a complex graph. Thus, the propagation model has as input graph G = (V,E), a state vector $S_v^{(t)}$for each node vertex v ∈ V at t, and a parameter vector P. Based on states of all interacting nodes, it outputs new state vector $S_v^{(t+1)}$for each node at t+1 (Giakkoupis et al, 2010). Models are applied to synthetic dataset with graph types as (Mitchell, 1997; Giakkoupis et al, 2010; Kermach and McKendrick, 1927; Pastor-Satorras and Vespignani, 2002; Watts, 1999; Newmann, 2003):

a. **Scale-Free Graphs:** Probability that node *x* in network is of degree k is proportional to $k^{-\gamma}$ with $\gamma > 1$. Scale free graph are modeled as by Barabasi and Albert (1999). It inserts nodes sequentially with each node linked to an existing one chosen with a probability proportional to its current degree in a tree-fashion with grandparent-parent-children-grandchildren structure and it builds graph with exponent $\gamma = 3$ denoted with G_{sf}. Each node in the graph can be autonomous but must be connected to an existing one. Thus, two nodes are connected together on the graph via physical link between two corresponding autonomous systems. Such is referred to as Autonomous Systems. Another scale free graph consists of undirected edges between nodes, also termed Co-Author graphs.
b. **Small World Graphs** are those with small characteristics path length L (average shortest path between any pair of vertices) and large clustering coefficient C (the average fraction of pairs of neighbours of a node also connected to each other). We generate small-world graphs using the generating model proposed in Watts (1999). We use G_{swL} to denote small world graph with path length feat; while G_{swC} to denote those of large clustering coefficient.

Graphs of G_{swL} are influenced by α, which intuitively determines the probability of two nodes being connected given a number of their common neighbours. It controls to what extent the graph has small and densely connected components. As α nears infinity, G_{swL} becomes a random graph. Conversely, graphs of G_{swC} are influenced by q, which determines the probability of an edge in the lattice being rewired to connect to a random node in the graph. Thus, initialized on a ring lattice, each node is of degree k. Small values of q entails graph G has high clustering coefficient and large average path length; while large values q creates random graphs. For q-values close to 0.01, the generated graphs are small-world graphs. Note that G_{swL}, G_{swC} and G_{sf} are quite distinct graphs.

III. Susceptible-Infect Models

There are two major susceptible-infect (SI) models: SI-Remove (SIR – for which a node can be in state: (a) susceptible: if the node has no virus but becomes infected if it is exposed to it, (b) infected: if the node has the virus and can pass it on to others, and (c) removed: if node had the virus but has been recovered or virus dies. Node is permanently immunized and can no longer propagation as a particular node cannot be infected twice) and SI-Susceptible (SIS – a node can be cured but not immunized. Thus, it can be infected again. Such node switches between susceptible and immunized.

Giakkoupis et al (2010) and Lahiri and Cebrian (2010) A graph holds these definitions as true:

a. Network is a directed or undirected graph for propagation of virus. A node is represented as v ∈V; and edge (u,v) ∈ E represents interactions between two individuals or nodes in the system. It also assume that the graph is drawn from a specific family (algorithm consider all possible graphs). For G = (V,E) as a dynamic network, E is set of edges that are time-stamped, $(u,v)_t \in E$ are interactions at $t \in Z^+$. In a typical SI setting, set of nodes are initialized as *activated*. The propagation process proceeds in discrete time-steps such that at each time-step, an activated vertex may come in contact with *inactive*vertices. This continues till a stop criterion is satisfied or there are no more inactive vertices.
b. The virus propagation model that determines how the virus is spread on the network.
c. Immunization model aims to minimize viruses spread and an immunized node cannot transfer/receive a virus. It is conceptually removed from graph. Cost of immunization model is, number of nodes immunized.
d. Adversary with knowledge of propagation model, plants *d* copies of malware in network so as to maximize speed of its spread is denoted as F_d. An adaptive adversary is one who has knowledge of choices made by the immunization algorithm; while a randomized adversary places copies of virus, uniformly at random on the network.

A. *Independent Cascade (SIR) Model*

It is a discrete-time special case SIR for which at t = 0, an adversary inserts *d* virus copies to some nodes on graph. If node *x* is infected at t, it can infect any neighbour *y* currently uninfected with the probability P_{xy} – so that *x* succeeds in infecting *y* at t+1; Else, *x* tries again in the future (even if *y* gets infected by another neighbour). This process continues and stops after *n*-steps if no more infections are possible. The model also requires a node to stay infected exactly once as proven by Kempe et al

(2003). A graph of size M, has M_d subset of nodes and d virus copies placed on the network. With propagation complete, $S(M_d,G)$ is expected number of infected nodes. Expectation exceeds all random choices made by propagation model. Eq. 1 is maximum expected number of infected nodes and maximum exceeds all possible initial virus placements. Equation 1 is given as thus:

$$S_d(G) = \max_{M_d} S(M_j, G) \quad (1)$$

Set $A_d = arg \max_{M_d} S(M_d, G)$ corresponds to choices made by an adaptive adversary. $S_d(G)$ is epidemic spread in G and a similar definition of epidemic spread of randomize adversary defines the expected epidemic spread minimization as in Eq. 2 in which case, expectation takes over all possible positions of the d viruses placed on the network and given by:

$$S'_d(G) = \mathrm{E}_{\mathrm{M_d}}[S(M_d, G)] \quad (2)$$

B. Dynamic Propagation (SIS) Model

In SIS, viruses are seen as dynamic birth-death process that evolves overtime. It continues to either propagate or eventually die. An infected node x spreads virus to node y in time t with infection rate of $\frac{\beta}{\delta}$ and probability β. At same time, an infected node may recover with probability δ. With adjacency matrix T, $\lambda_1(T)$ is largest eigen-value of T. The condition $\frac{\beta}{\delta} < \frac{1}{\lambda_1(T)}$ holds true as epidemic threshold and is sufficient for quick recovery, easily proven (Ganesh et al, 2005; Wang et al, 2003.

IV. The Immunization Problem

Typical challenges in SI propagation model are as follows and from the immunization problem perspective, we have:

1. **Targeting:** Which vertices are targeted as initiators by an adversary to result in max extent of spread (Cohen et al, 2003)? This is a hard NP to solve optimally, regardless of propagation model used (Kempe et al, 2003). This is also referred to as the Epidemic Immunization problem so that given the graph G, a number of d virus copies and a number k, we immunize k nodes in G so that expected spread $S_d(G')$in immunized graph is minimized. This hard NP complete has the role of an adversary played by the influence-maximization model of Kempe et al (2003), whose proof is omitted due to space constraint. This is addressed with Eq. 1 above.
2. **Extent:** With G, subset of initially activated node vertices and propagation model, how many vertices are expected to be activated after a specific time/period? It is also referred to as the Expected Minimization problem in which we have the graph G, a number of d virus copies and a number k. We aim to immunize k nodes in G such that the expected spread $S'_d(G')$in the immunized graph is minimized. It is a hard NP-complete task that attempts to immunize G with random strategy for influence spread and closely related to the sum-of-squares partition task as studied in Aspnes et al (2005) as addressed by Eq. 2.
3. **Blocking:** Which vertices are targeted for immunization to minimize the expected number of activated vertices (Singhal and Raul, 2012; Dezso and Barabasi, 2002)? This is also called Epidemic Threshold problem in that given G, a number of d virus copies and an infection rate of $\frac{\beta}{\delta}$, we immunize the minimum number of k nodes so that $\frac{\beta}{\delta} < \frac{1}{\lambda_1(T)}$ holds true. Thus, the epidemic spread $S'_d(G')$in the graph is minimal. The task attempts to immunize G with influence spread while seeking the minimal number of nodes that can be immunized.

The study aims to compute the epidemic threshold using the various stochastic (machine learning) models to unveil the feats and corresponding underlying dataset probabilities.

V. Machine Learning Frameworks

Machine learning as a branch of artificial intelligence is a scientific discipline that deals with development and design of algorithms that allows machines (computers) to evolve its behaviour based on empirical data such as sensors data and databases. A learner takes advantage of data to capture its characteristics of interest of their underlying and unknown probability distribution. Such data may illustrate relationships between observed variables. Major focus on machine learning is to automatically learn to recognize complex patterns and make intelligent decisions from it (Singhal and Raul, 2012).

A. Dataset

The Enron e-mail dataset is one of the largest, corporate e-mail environment dataset available and is naturally represents a dynamic social network. Each vertex in network is an e-mail address and a directed timestamped edge representing e-mail sent between two addresses. Lahiri and Cebrain (2010) parsed and obtained the headers of all e-mails of about 1,326,771 timestamped individual e-mails between 84,716 addresses, with 215,841 unique timestamps as non-uniformly covering a period of approximately 4years.

We sampled a subset of 25000 addresses representing about 30% for the graphs G_{sf}, G_{swL} and G_{swC} families. In all cases, we used p = 0.25, q = 0.009 and α = 6 to gener ate the graphs. These result in models' graph having low average path length and high clustering coefficient. There exists the relationship between parameters (p, q and α) and the clustering coefficient as studied in (Watts, 1999). α starts with value 1 till it reaches 6 or more. The clustering coefficient drop as α increase and for small values of q, high clustering coefficient is observed while clustering coefficient drops as q tends to 1.

B. *Hybrid Genetic Algorithm Particle Swarm Optimization (HGAPSO)*

This hybrid combines GA and PSO with the framework and workings as thus:

GA is inspired by Darwinian evolution (survival of fittest) and consists of a pool of solutions for natural selection to a specific task. Each individual is solution for which an optimal is found via 4-operators namely: initialize, select, crossover and mutation (Ojugo et al, 2012). Individuals with genes close to its optimal is said to be fit, and the fitness function determines how close an individual is to the optimal solution.

PSO attempts to predict motion as it investigates collective intelligence cum socio-cognitive of swarms as well as specify a model of randomly initialized solutions propagated in space of n-dimensional vectors towards an optimal result, over number of moves based on large amount of data about the domain, assimilated and shared by the swarm. In a bid to generate and select particles (solution) adapted to its environment via its set objectives and employed constraints, the desirable traits evolves and remains in the swarms' composition (as results generated) over traits with weaker undesirable feats (Ojugo et al, 2013). PSO is continuous and thus, is modified to handle discrete design variables.

We first proceed with GA, which achieves its fitness as it finds solution to network. Its dynamic, non-linear model can be made linear so as to resolve it analytically. The dynamic nature of graph as social network makes them impossible to resolve analytically using non-linearity (if considered as a multiple copies model). Let v^t be an n-dimensional vector of states at t-steps and $v_d^t$is number of virus copies at node *x* at t-steps. Initialized at t = 0, $v_d^0$is number of *d* copies planted by an adversary. At t+1, the model evolves for (all) nodes *x,y,z* in the network, and for each v_d^tcopies of virus planted at node *x*, virus is propagated to node *y* with probability β. Virus dies with probability 1 – δ, and if Δ = βT + diag(1 – δ,…, 1 – δ) is true, $v^t$is the expected state of system at time t. Model is completely linear if $\Delta v^t = \Delta v^{t+1}$ proven as in (Giakkoupis et al, 2010; Kleinberg, 2007;Hethcotee, 1989).

For PSOGA, model starts of with PSO particle position and velocity initialization with other operators as thus (Lahiri and Cebrian, 2010):

a. Particle Position/Velocity: A node refers to a particle point in vector space that changes position between iterations based on velocity updates. Node positions and velocities are randomly initialized (active and inactive) with lower and upper bounds of the design variables (X_{min} and X_{max}) as in Eq. 3 and Eq. 4 respectively – since the adversary places *d* virus copies at points for maximum propagation. Fitness value is computed for the swarm as thus: (a) its global best P_i^g, (b) neighbourhood best $P_i^{t-g}$and (c) particle best position P_i^t. Velocity update uses current motion V_i^t to search for V_i^{t+1}. To avoid local maxima entrapment, ensure good coverage and convergence time, PSO uses a uniformly distributed rand() with 3-factors: (a) inertia ϖ, (b) particle confidence ϕ1 and (c) swarm influence ϕ2 to find optima (see Eq. 5 and Eq. 6). $V_i^{t+1}$is particle velocity at t+1, ϖ+$V_i^t$is current motion, $X_i^t$is particle position, ϕ1 is set between [1.5,2] and ϕ2 is set between [2,2.5]. The weighted sum fitness function is adopted, for the study:

$$X_i^o = X_{min} + rand(X_{max} - X_{min}) \tag{3}$$

$$V_i^t = \frac{X_{min}+rand(X_{max}-X_{min})}{\nabla t} = \frac{Position}{Time} \tag{4}$$

$$V_i^{t+1} = \omega + V_i^t + \phi_1 rand() \frac{(P_i^t - X_i^t)}{\nabla t} + \phi_2 rand() \frac{(P_i^t - X_i^t)}{\nabla t} \tag{5}$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} * \nabla t \left| f\left(P_i^g\right) - f\left(P_i^{t-g}\right) \right| \in \tag{6}$$

b. Crossover point is randomly, uniformly selected between [1,β]. Two new state vectors are created by swapping $S_v^{(t)}$ and $S_u^{(t)}$ defined by all position including and after index C. Two vector states strings are denoted as st_1 and st_2respectively.

c. Objective score of each new state vector is then evaluated according to the fitness function f(x). if any of them have a greater fitness value that either of their parent node, the corresponding parent nodes state vector string is replaced by its offspring for the next iteration, achieved via Eq. 7 and Eq. 8 respectively:

$$S_u^{(t+1)} = \operatorname{argmax}_{x \in \{S_u^{(t)}, S_u^{(t+1)}, st_1, st_2\}} f(x) \tag{7}$$

$$S_v^{(t+1)} = \operatorname{argmax}_{x \in \{S_v^{(t)}, S_v^{(t+1)}, st_1, st_2\}} f(x) \tag{8}$$

In case of ties in fitness score (original versus new string) vector, original state vector remains. It is close to GA with spatially distributed population (Min et al, 2006; Payne and Eppstein, 2006) as GA's selection is replaced with real network data that dictates sequence of mating. Propagation in GASDM occurs as state vectors modified via crossover and is subsequently adopted based on its fitness value.

d. With fitness function computed, nodes can assume values even outside their lower/upper bounds constraint – due to their current position and computed velocity. As function of the rapid growing vector velocity – causing inactive nodes to become active for propagation via divergence; instead of node convergence to help in immunization and virus propagation minimization. To avoid this, nodes that violate their bounds are artificially brought back to its k-nearest via Eq. 9 (to avoid velocity explosion and handle functional constraints via a linear exterior penalty). This is the epidemic threshold value for immunization. Thus, with infection rate $\frac{\beta}{\delta}$, $\frac{\beta}{\delta} < \frac{1}{\lambda_1(T)}$ holds as logarithmic expected time for virus to die (function of number of nodes in G that an adversary targets for maximum virus spread). Some graphs hold $\frac{\beta}{\delta} > \frac{1}{\lambda_1(T)}$ as recovery rate so that the expected time for virus to die is exponential (Aspnes et al, 2005).

$$f(x) = \varphi(X) + \sum_{i=1}^{N_{max}} X_i * \max[0, g_i(x)] \qquad (9)$$

Solutions are updated via the model's crossover and fitness (objective) function computed with these in mind: (a) velocity range normalized between [0-1], dividing it by maximum range of the particles, (b) each position randomly determines if crossover is needed as determined by the velocity, and (c) If required, the position will set to the value of same position in nbest by swapping the values.

C. Decision Tree (Random Forest Algorithm)

Random Forest Model as a decision tree predictor has each node trained on partially, independently instance subsets of complete training dataset. The predicted output of classified instance is the most frequent class output of node (Szor, 2005; McGraw and Morrisett, 2002; Mitchell, 1997).

It uses hill-climbing to search a space for optima. Once a peak is found, it restarts with another randomly chosen starting point (as such peak may not be the only one that exists). Its merit is simplicity with functions with too many maxima. Each random trial done in isolation helps immunize the nodes and overall shape of the domain is transparent to an adversary – because, as random search progresses, it continues to allocate its trials evenly over the space and evaluates as many points in the both regions found with low- and high-fitness values. Its choice is in selecting feats and attributes in graph to test is via information gain at each step while it grows the graph. The algorithm as Mitchell (1997), Ojugo et al, (2012b) is thus:

```
DT (Examples, Target_Attribute, Attributes)
//Example are dataset, Attributes are other feats tested by model.
//Target_Attribute is attributes with values to be predicted,
//Return is a decision that correctly detects a given Examples.
Create a Root node of Graph
If Examples are positive, Return single_node Graph Root with label=+
If Examples are negative, Return single_node Graph Root with label = -
If Attribute is empty, Return single_node Graph Root, with label = most
   common value of Target_Attribute in Examples
Otherwise Begin
    a.  A ← the attribute from attributes that best* classifies Examples
    b.  The decision attribute for Root ← A
    c.  For each possible value $v_i$, of A,
        Add new graph branch below Root, corresponding to test A =
        $v_i$
        Let Examples $v_i$ be subset of Examples that have value $v_i$ for A
        If Examples is $v_i$ empty
          Then below this new branch, add leaf with label = most
        common value of Target_Attribute in Examples
Else below this new branch, add the subtree
              IDA(Examples $v_i$, Target_Attribute, Attributes – {A})
End: Return Root
```

Entropy characterizes all impurity of an arbitrary collection of nodes on G, which contains both active (infected) and inactive (uninfected) nodes as a Boolean classification.

$$Entropy(E) \equiv -p_{\oplus} log_2 \, p_{\oplus} - p_{\ominus} log_2 \, p_{\ominus} \qquad (10)$$

The sample consists of n=25000 e-mail address from which we have normal and infected nodes to form the graph network. The inactive (p+) = 20000, infected (active/p-) nodes where adversary plants viruses p- = 5000. To compute Entropy, we have that:

$$E \equiv -\frac{20000}{25000} log_2 \frac{20000}{25000} - \frac{5000}{25000} log_2 \frac{5000}{25000}$$
$$E \equiv [-(0.8) log_2 \, (0.8)] - [(0.2) log_2 \, (0.2)]$$
$$= 0.0775 + 0.1398 = 0.22$$

Information Gain is the expected reduction in entropy caused by partitioning graph according to its attributes (infected and uninfected) nodes. IG is info about target function value, given the value of another attribute A. IG of attribute (A) is:

$$Gain(E, A) \equiv (E) - \sum_{v \in Values(A)} \frac{|E_v|}{|E|} Entropy(E_v) \qquad (11)$$

Values(A) is set of all possible values of Attribute A, E_v is E subset of attributes A with value v. Our second is the expected entropy after partitioning with attribute A (sum of all entropies of each subset E_v weighted by fraction of Examples $\frac{E_v}{E}$ of E_v).

$$Gain(E, A) \\ \equiv E - \sum_{v \in \{inactive, immunized\}} \frac{|E_v|}{|E|} Entropy(E_v)$$

$$IG \equiv 0.220 - \left\{\frac{8000}{25000} * 0.811\right\} - \left\{\frac{23000}{25000} * 0.921\right\}$$
$$= 0.220 - \{-0.587\} \equiv 0.220 + 0.587 \equiv 0.807$$

Thus, we have an Epidemic threshold value so as to enable us scale 81% of nodes as most likely to be infected before complete immunization is achieved. However, IG is further corrected via Eq. 12:

$$Gain(E, A) \equiv Gain(E, A) \pm \left[\frac{\sum_{i=o}^{n} Gain(X_i)}{n}\right] \qquad (12)$$

D. Na ïve Bayesian Model

Bayesian model describes probability distribution of a set of nodes on graph by specifying a set of conditional independent assumptions along with set of conditional probabilities. Thus, allows stating conditional assumptions that simply just applies to subset of nodes on the network by providing an intermediate and more tractable solution. It applies to each instance that assumptions of each graph attribute values are conditionally independent of the target value. Thus, the assumptions is that given target value of an instance, the probability of observing the interactions between nodes in the graph is the product of their probabilities from the individual attributes (Szor, 2005; Alpaydin, 2010; Mitchell, 1997, Harrington, 2012).

Naïve Bayes is used for two reasons: (a) it computes explicit probabilities for hypotheses and outperforms other methods in this regard, and (b) it provides useful insight into understanding other algorithms that do not explicitly manipulate probabilities. Mitchell (1997) Adopting Bayesian model for the study is based on feats as thus:

a. Each observed training incrementally increase or decrease the estimates probabilities that a hypothesis is correct as a means of providing flexible learning rather than outright elimination of such hypothesis as inconsistent.
b. Prior knowledge (via apriori probability of each hypothesis and corresponding probability distribution over observed data for each of possible hypothesis) can be obtained with observed data to determine the final probability of a hypothesis.
c. Bayesian method accommodates hypothesis that makes probabilistic predictions.
d. New instances can be classified by combining predictions of multiple hypotheses, weighted by their probabilities.
e. Cases where Bayesian proves computationally intractable, it however yield a standard for optimal decision against which other practical methods can be measured.

We assume every active node that has been infected and/or immunized is equally probable a priori. Any such probable hypothesis as maximum a posteriori:

$$h_{MAP} = \text{argmax}_{h\in H} P(D|h)P(h) \qquad (13)$$

With sample n=25000 e-mail address. Normal (inactive/p+) = 20000, infected (activated/p-) nodes where adversary plants viruses p- = 5000. We compute probabilities of the maximum likelihood both from Entropy and correlation (of the correctly classified versus false negative classification) as thus:

$$P(\text{infected/p-}) = 0.2, \quad P(\neg\text{infected}) = 0.8$$
$$P(\oplus|\text{infected}) = 0.81, \quad P(\ominus \mid \text{infected}) = 0.19$$
$$P(\oplus|\neg\text{infected}) = 0.07, \quad P(\ominus \mid\neg\text{infected}) = 0.93$$

Then, we have that:

$$P(\oplus|\text{infected})\ P(\text{infected}) = (0.81)(0.2) = 0.162$$
$$P(\oplus|\neg\text{infected})\ P(\neg\text{infected}) = (0.07)(0.8) = 0.056$$

$h_{MAP} =$ (¬infected)yields maximum a priori likelihood of propagation. The exact probabilities of maximum spread and propagation before complete immunization, is determined by normalizing these probabilities to sum up to 1; And this yields maximum propagation (epidemic threshold) given by:

$$P(infected|\oplus) \equiv \frac{0.162}{0.162+0.056} = 0.74$$

Thus, with 5000-infected nodes at initialization, the epidemic threshold is 74%, at which time network will be completely immunized or the virus dies out.

VI. Discussion And Findings

From sampled subset of 25000 addresses (30% of dataset, with p = 0.25, q = 0.009 and α = 6 to generate graphs), the results with low path length and high clustering coefficient is as in table 1). There exists a relationship: that α starts with value 1 and tends upwards till it reaches 6. As α increase for small values of q, a high clustering coefficient is observed while clustering coefficient drops as q tends to or approaches 1. The epidemic threshold (ET) displays rate at which virus dies out (recovery rate) or that for which model completely immunizes the graph. The Expected Spread Immunization (ESI) was computed as 91% for both SIR and SIS models employed. Also, the Expected Epidemic Spread Minimization (EESM) was computed as 97% for both SIR and SIS models in use. Other findings are as thus:

a. GAPSO was run 25times with time varying between 21seconds and 4 minutes. Its convergence time depends on how close initial population is to the solution as well as on mutation applied to the individuals in the pool. With dataset (25000 addresses), correctly classified instance is 23567 (94.3%), incorrectly classified instances is 1433 (5.7%). The Epidemic Threshold was computed as 85% with overall accuracy of about 91%.
b. RFA was run 25times with time varying between 11seconds and 1minute. Its convergence time depends on how fast each random trial was completed as well as the random search with its continued allocation of trials evenly over its search space cum evaluation of as many points in both regions found with low- and high-fitness values. Its choice is in selecting feats and attributes in graph to test, also contributes to this convergence time in computing the fitness value for the training dataset. With dataset (25000 addresses), correctly classified set of 96.7%, incorrectly classified instances of3.3%. ET is computed as 81% with an overall accuracy of about 97%.
c. Bayesian model was run 25times with time varying between 11seconds and 1minute. Its convergence time depends on how fast each random trial was completed as well as the random search with its
d. continued allocation of trials evenly over its search space cum evaluation of as many points in both regions found with low- and high-fitness values. Its choice is in selecting feats and attributes in graph to test, also contributes to this convergence time in computing the fitness value for the training dataset.

With same dataset (25000 addresses), correctly classified set 95.9%, incorrectly classified instances of4.1%. ET is74% with an overall accuracy of about 96%.

A. Model Implementation Tradeoffs

Trade-offs in modelling often fallunder these classifications and groupings as thus (Ojugo et al, 2013c):

a. **Result Presentation:** Modelersand researchers quite often display flawed and unfounded results – with the aim to validate their new and/or modified model rather than re-test the limitations, biasness, insufficiency and inabilities of existing ones. This is because negative results are less valuable and most of such models aim to curb the non-linearity and dynamism in the phenomena they are predicting alongside discovering feats and underlying properties of the historic datasets used, to train, cross validate and test such models.
b. **Efficiency:**modelersand researchers can often use figure to show how well their prediction is quite in agreement with observed values (even with their limited dataset used for training the model that is often times squeezed). Some plot for observed and predicted values are often not easily distinguishable – as such modelers do not even provide numerical data to support their claim for their system (though their model is in 'good agreement' with observed values). Some measure of goodness does not provide the relevant information.
c. **Insufficient Testing:** Validation simply compare predicted versus observed values. Many studies suffer from inadequate dataset. If a model aims to predict and simulate a dynamic event or phenomena, such ability should not be demonstrated with unfounded results with limited dataset, displaying (often) misleading results and inconclusive and unclear contributions. Model must be adequately tested, with materials and methods for such experiments laid bare so that such predictions can be repeated if need be to validate the usefulness and authenticity of such models.
d. **Validation:** is not an undertaking to be carried out by a researcher or research group; but rather, a scientific dialogue. Improper model applications and ambiguous results often impede such dialogue. This study aims to greatly minimize confusion in propagation model as well as further mathematical epidemiology.

B. Rationale for Choice of Algorithms

The comparisons are as follows:

a. **Stochastic Model:** are mostly inspired by evolution laws and biological population cum behaviours. They are heuristics that search a domain space for optimal solution to a task. They use hill-climbing method that are flexible, adaptive to changing states and suited for real-time application. GA guarantees high global convergence to optimal point for multimodal tasks. It initializes with a random population, allocates increasing trials to regions of the space found with high fitness and finds optimal in time. Its demerit is that they are not good with linear systems in that if the optimal is in a small region surrounded by regions of low fitness – the function becomes difficult to optimize.
b. **Gradient/Greedy Search:**A number of different methods for optimizing well-behaved continuous functions have been developed which rely on using information about the gradient of the function to guide the direction of search. If the derivative of the function cannot be computed, because it is discontinuous, for example, these methods often fail. Such methods are generally referred to as *hill-climbing*. They can perform well on functions with only one peak (*unimodal*functions). But on functions with many peaks, (multimodal functions), they suffer from the problem that the first peak found will be climbed, and this may not be the highest peak. Having reached the top of a local maximum, no further progress can be made. Conversely, the iterative search is a combined random and gradient search that employs an *iterated hill-climbing* search. Once one peak is located, the process restarts with another, randomly chosen point. Its merit of simplicity. RFA/Bayesian are chosen for these reasons: (a) Instances are represented in graph as attributes value pairs, (b) the target function has discrete output values as it assigns Boolean classifications to each network, (c) disjunctive description may be required, and (d) training sample data may contain errors and may contain missing attributes values.

VII. Conclusion

Mathematical models have been successfully used today to determine epidemic spread of malware. Numerous recent studies on mathematical epidemiology focuses on the analytic epidemic thresholds for time-varying propagation models as applied on different families of network – seeking insight into the nature of such epidemic, its threshold as well as to unveil if such propagation continues or eventually, dies out (Bougna et al, 2003; Barthelemy et al, 2005; Barabasi and Albert, 1999; Dawkins, 1993).

Models serve as educational, predictive tools to compile all existing knowledge and information about a task, serve as a vehicle to communicate hypotheses, a means to investigate parameters crucial in estimation and help us better understand a problem domain. Its development, sensitivity and failure analysis helps reflect on the theories and functioning of nature systems. Simple models may not provide enough new data, whereas very complex models may not be fully understood. A model's use and application as an intellectual tool requires less accurate numerical agreement between prediction and observations. Rather, it requires feedback mechanisms, as more important – since only models that are understandable and manageable, can be fully explored. A balance between complexity and simplicity is crucial for studying the relevant processes and still, to understand how the model works.

Our comparative (machine learning and stochastic) spread and propagation models provides a framework that is best suited for large organizations with enterprise gateway level to act as central antivirus engine to supplement AVs, present at the end-users' computers. It will be employed to easily detect malwares and act as a knowledgebase to help detect newer forms. While a costly model requiring costly infrastructure, it can help protect valuable data in an enterprise from security threats and prevent immense financial damage. Its only demerit is that it require large processing power and thus, cannot be adopted by a home users (Singhal and Raul, 2012; Gao et al, 2011' Ojugo et al, 2013d). Studies on mathematical epidemiology as successfully used in malware detection, is focused on analytic epidemic thresholds for varying spread models and families of graphs – seeking insight into the nature of such epidemic, its threshold and to unveil if such epidemic will continue to spread or die.

References

[1] E. Alpaydin, *Introduction to Machine Learning*, McGraw Hill publications, ISBN: 0070428077, New Jersey, 2010.

[2] J. Aspnes, K. Chang and A. Yampolsky, *Inoculation Strategies for Victims of Viruses and the Sum-of-Squares Partition Problem*. In *SODA*, 2005.

[3] A. L. Barabasi and R. Albert, *Emergence of scaling in random network*. Science, 286, p23, 1999.

[4] M. Barthelemy, A. Barrat, R. Pastor-Satorras and A. Vespignani, "Dynamical patterns of epidemic outbreaks in complex heterogeneous networks". Journal of Theoretical Biology, p54, 2005.

[5] C.M. Bishop, *Pattern Recognition and Machine Learning*, ISBN-13: 978-0387-31073-2, Springer Science and Business Media, LLC, 2006.

[6] M. Boguna, R. Pastor-Satorras and A. Vespignani, *Epidemic Spreading in Complex Networks with Degree Correlations*. Statistical Mechanics of Complex Networks, p36, 2003.

[7] R. Cohen, S. Havlin and D. Ben-Avraham, *Efficient Immunization Strategies for Computer Networks and Populations*. Phys. Rev Letters, p232, 2003.

[8] R. Dawkins, *The Selfish Gene* (2nd edition) Oxford University Press, 1989.

[9] R. Dawkins, *Viruses of the Mind* in B. Dahlbom (Ed.) *Dennett and his Critics: Demystifying the Mind*, Blackwell, USA, p12, 1993.

[10] P. Desai, *Towards an Undetectable Computer Virus*, Masters Thesis, Department of Computer Science, San Jose State University, 2008.

[11] Z. Dezso and A. Barabasi, *Halting Virus in Scale-free Networks*. Phys. Rev E66, p67, 2002.

[12] E. Filiolel, *Computer Viruses: Theory to Applications*, Springer, ISBN: 2287-23939-1, 2005.

[13] A. Ganesh, L. Massouli and D. Towsley, "The Effect of Network Topology on the Spread of Epidemics". In *IEEE INFOCOM*, 2005.

[14] C. Gao, J. Liu and N. Zhong, *Network Immunization and Virus Propagation in Emails Network: Experiment and Evaluation Analysis*, Knowledge and Information Systems, 27(2), p253-279, 2011.

[15] G. Giakkoupis, A. Gionis, E. Terzi and P. Tsaparas *Models and Algorithms for Network Immunization*, Engr. Letters, 243, p89, 2010.

[16] P. Harrington, *Machine Learning in Action*, Manning publications, ISBN: 9781617290183, New York, 2012.

[17] D. Kempe, J. Kleinberg and E. Tardos, *Maximizing the Spread of Influence through a Social Network*. In *SIGKDD*, 2003.

[18] W. Kermack and A. McKendrick, *A Contribution to the Mathematical Theory of Epidemics*. Proceedings Royal Society London, 1927.

[19] J. Kleinberg, *Cascading Behavior in Networks: Algorithmic Economic issues*. Algorithmic game theory, NY, 2007.

[20] M. Lahiri, A.S. Maiya, R. Sulo, K. Habiba and T. Y. Berger-Wolf, "The Impact of Structural Changes on Predictions of Diffusion in Networks". In IEEE ICDM Workshops, p939, 2008.

[21] M. Lahiri and M. Cebrain, *The Genetic Algorithm as a General Diffusion Model for Social Networks*, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2010.

[22] T. M. Mitchell, *Machine Learning*, McGraw Hill publications, ISBN: 0070428077, New Jersey, 1997.

[23] M. E. J. Newman, *The Structure and Function of Complex Networks*. SIAM Reviews, 45(2), p167–256, 2003.

[24] A. Ojugo, A. Eboka, E. Okonta, R. Yoro and F. Aghware, "GA Rule-based Intrusion Detection System", Journal of Computing and Information Systems, 3(8), p1182, 2012a.

[25] A. A. Ojugo, M. Yerokun, A. Eboka and E. Ugboh, *Malware Propagation on Networks: Analysis, Propagation and Detection*, Technical-Report, Centre for High Performance and Dynamic Computing, TRON-02-2012-01, Federal Univ.of Petroleum Resource, Nigeria, p45, 2012b.

[26] A. A. Ojugo, *Virus Propagation on Time Varying Graphs*, Technical-Report, Centre for High Performance and Dynamic Computing,TRON-03-2013-01, Federal Univ. of Petroleum Resources, Nigeria, p24-37, 2013a.

[27] A. A. Ojugo, and R. Yoro, *Computational Intelligence in Stochastic Solution of Toroidal Queen*, Progress in Intelligence Computing Applications, 2(1), doi: 10.4156/pica.vol2.issue1.4, p46, 2013b.

[28] A. A. Ojugo, J. Emudianughe, R. E. Yoro, E. O. Okonta and A. O. Eboka, "Hybrid Artificial Neural Network Gravitational Search Algorithm for Rainfall Runoff", Progress in Intelligence Computing and Applications, 2(1), doi: 10.4156/pica.vol2.issue1.2, p22, 2013c.

[29] A. A. Ojugo, M. Yerokun, A. Eboka and E. Ugboh, *Virus Propagation on a Time Varying Network: Analysis and Detection*, Technical-Report, Centre for High Performance and Dynamic Computing, TRON-03-2013-12, Federal Univ. of Petroleum Resources, Nigeria, p234, 2013d.

[30] R. Pastor-Satorras and A. Vespignani, *Epidemics and Immunization in Scale-free Networks*. Handbook of Graphs and Networks: From the Genome to the Internet, 2002.

[31] P. Singhal and N. Raul, "Malware Detection Module using Machine Learning Algorithm to Assist Centralized Security in Enterprise Networks", Int. J. Network Security and Applications, 4(1), doi: 10.5121/ijnsa.2012.4106, p61, 2012.

[32] P. Szor, *The Art of Computer Virus Research and Defense*, Addison Wesley Symantec Press. ISBN-10: 0321304543, New Jersey, 2005.

[33] Y. Wang, D. Chakrabarti, C. Wang and C. Falousos, *Epidemic Spreading in Real Networks: An Eigenvalue view*

point. In *SRDS*, 2003.

[34] D. J. Watts, *Networks, Dynamics and the Small World Phenomenon*. American Journal of Sociology, 105, p234-245, 1999.

Authors' Profiles

Ojugo, Arnold Adimabua received BSc from Imo State Univ. Owerri in 2000, MSc from NnamdiAzikiwe Univ. Awka in 2005 and PhD from Ebonyi State Univ. Abakiliki in 2012 (Computer Sci.). Currently with Dept. of Math/Comp Sci, Federal Univ. of Petroleum Resources Effurun, Delta State, Nigeria. His research interests: Soft Intelligent Computing, Machine-Learning, Robotics Vision, Data Security/Forensics and Cloud Computing. Editor with Progress for Intelligent Computation and Application and Advancement for Scientific and Engineering Research.Member of: Nigerian Computer Society, Computer Professionals of Nigeria and International Association of Engineers

Ben-Iwhiwhu, Eseoghene received BSc from Federal University of Petroleum Resources, Effurun in 2012 in Computer Science. He his currently a graduate assistant with the Department of Mathematics/Computer Science at Federal University of Petroleum Resources, Effurun, Delta State, Nigeria. His research Interests include artificial intelligenece and robotics.

Kekeje, O. received her BSc from Benson Idahosa University, Edo state in 2008 in Computer Science. She iscurrently a gradudate assistant with the Department of Mathematics/Computer Science at Federal University of Petroleum Resources, Effurun, Delta State Nigeria. She is also currently pursing a master degree in computer science at the University of Port-Harcourt, Nigeria.Her research interests are in Distributed Computing, Data mining, and Concurrency models.

Yerokun, Oluwatoyin Mary received BSc from Univ. of Benin in 2000 and MSc from Nnamdi Azikiwe Univ. Awka in 2008 and currently a PhD student at Ebonyi State Uni. Abakiliki (in Computer Sci.). Currently lectures at Dept of Computer, Federal College of Education (Technical) Asaba, Delta State. Her research interests in: Software Evolution and Data Communications. She is a member of: International Association of Engineers (IAENG), Nigerian Computer Society (NCS) and Computer Professionals of Nigeria (CPN). Her details are: +2348034095720 / agapenexus@hotmail.co.uk.

Iyawa, Ifeyinwa Jane received BSc from Univ. of Lagos in 2000 and MSc from NnamdiAzikiwe Univ. Awka in 2005 and currently a PhD student at Ebonyi State Uni. Abakiliki (in Computer Sci.). Currently lectures at Dept of Computer, Federal College of Education (Technical) Asaba, Delta State. Her research interests in: Software Evolution and Data Communications. She is a member of: International Association of Engineers (IAENG), Nigerian Computer Society (NCS) and Computer Professionals of Nigeria (CPN). Her details are: +2348163165807 / iyawaben@hotmail.com