Compliance Graph Analysis Using Epidemiology Modeling

Noah L. Schrick
Tandy School of Computer Science
The University of Tulsa
Tulsa, USA
noah-schrick@utulsa.edu

Peter J. Hawrylak
Tandy School of Computer Science
The University of Tulsa
Tulsa, USA
peter-hawrylak@utulsa.edu

***Abstract*—**

***Index Terms*—Compliance Graph; Attack Graph; Epidemiology Modeling; Risk Assessment;**

I. Introduction

CG Intro

Epidemiology modeling is frequently used to model the behavior and dynamics of a disease across a population. Most commonly, a SIR model (Susceptible - Infected - Recovered model) is used to model how compartments of a population change with regard to a disease. The model is able to capture how the "susceptible" subset of the population (the subset that is able to become infected) changes over time, as well as how the "infected" subset (the subset that is currently infected with the disease) and "recovered" subset (the subset that has been infected, but has since recovered) change over time. By examining the epidemic curves, it is possible to plan and prepare possible responses to mitigate the impact, as well as measure or predict an epidemic point. These models help provide insight on how to prepare, prevent, or mitigate challenges brought on by a disease outbreak. Epidemiology modeling has been widely investigated over recent years, and various works have been presented for modeling secondary infections through a model [1], additionally using multiple regression [2] or ridge regression [3], and including new representations for social and community interactions [4].

This work discusses an epidemiology modeling approach for compliance graph analysis. This work leverages the ability of epidemiology models to measure the behavior of compliance and regulation violations in a system or set of system and determine the risk they may have. This work also leverages the predictive power of epidemiology models to present mitigation or preparation strategies for future violations. By using an epidemiology approach, it is possible to refine constraint-bound correction schemes for the predicted or expected violations that systems may face.

II. Related Works

The application of epidemiology modeling is not limited to modeling the dynamics of a disease. Various works have been performed to apply epidemiology modeling specifically to computer and information systems. The authors of [5] were able to analyze the robustness of a computer network using an epidemiology model, the authors of [6] were able to determine configurations for intrusion-detection systems, the authors of [7] were able to model the spread of computer viruses, and the authors of [8] were able to model the spread of malware over a network. Epidemiology modeling is also able to predict and measure social or relationship conflicts that may occur from third-party information system or information technology providers. The authors of [9] were able to predict how these conflicts may propagate during a third-party provider implementation using an SIR model. Epidemiology models continue to see analytical improvements and usage across fields. The authors of [10] successfully demonstrated the use of a time-series data analytics approach, and the authors of [11] present a data-based model that is capable of integrating with, and crawling across, internet sites to minimize the manual overhead of implementing closed-form epidemiology expressions. The authors of [12] present a wide array of techniques that work in conjunction with the fields of artificial intelligence, big data, cognition, and analytical epidemiology to showcase promising results with causal probability theory.

III. Approach to Multi-Strain Epidemiology Modeling

A single-strain approach was originally used for analyzing the compliance and violation trends of environments. In a single-strain approach, all violations are grouped together as a single, generic "violation", all warnings are grouped together as a single, generic "warning", and all compliant systems, states, policies, and behaviors are grouped together as a single, generic "in-good-standing" label. This approach is useful for examining the system at a high-level and for identifying concerning overall trends regarding compliance or

regulation standings. However, this single-strain approach does not provide insight on the severity of violations, and does not provide distinct information on types of violations. This form of analysis is unable to discern which regulations are met or not met, and is unable to distinguish where the cause of concern may lie. In real systems, various violations may occur, each having a different level of severity. In addition, real systems are prone to violating multiple regulations simultaneously. In order to capture and examine violations at a more granular level, a multi-strain epidemiology model will be required. The multi-strain approach functions by treating each violation as a unique strain of infection, which allows for the system to be modelled across all strains with relevant violation information properly captured.

A. Alternative Model Details

Various works have been presented for a multi-strain epidemiology model. Specifically to computer and information systems, the authors of [13] present a SI_1I_2SD model (Susceptible, Infected-with-Strain-1, Infected-with-Strain-2, Susceptible, Deceased) for modeling malware over Internet of Things Wireless Sensor Networks. This model is able to capture two different strains of malware that each have a dependency on the other. However, for compliance graph analysis, there can, and likely will, be a greater number of violations present. As a result, limiting an epidemiology model to two strains is not fully realistic. The authors of [14] constructed a multi-strain PDE-SIS (Partial Differential Equation Susceptible - Infected - Susceptible) model that resembles the following and is capable of modeling multiple strains:

$$\begin{cases} \partial_t S = d_{\mathrm{S}}\Delta S + \sum_{i=1}^{k}\gamma_i(x)I_i - \frac{S\sum_{i=1}^{k}\beta_i(x)I_i}{S+\sum_{j=1}^{k}I_j} & x\in\Omega,\ t>0,\\ \partial_t I_i = d_i\Delta I_i - \gamma_i(x)I_i + \beta_i(x)\frac{SI_i}{S+\sum_{j=1}^{k}I_j} & x\in\Omega,\ t>0,\\ d_{\mathrm{S}}\frac{\partial S}{\partial\vec{n}} = d_i\frac{\partial I_i}{\partial\vec{n}} = 0 & x\in\partial\Omega,\ t>0,\\ N = \int_\Omega[S(0,x)+\sum_{i=1}^{k}I_i(0,x)]\mathrm{d}x > 0 \end{cases}$$

Fitting compliance graph analysis to an SIS model (Susceptible, Infected, Susceptible) is achievable, but this model would lose many of the attributes, nuances, and representative details that are present in other models and real systems. Fitting each compliance graph node to only "Susceptible" or "Infected" compartments prevents modeling of exposure or recovery of a disease. As an alternative, the authors of [15] present a multi-strain model that would be favorable for compliance graph analysis. In this approach, the authors discuss a multi-strain SEIR (Susceptible - Exposed - Infected - Recovered) model that resembles the following:

$$\begin{cases} S_j(t) = P(t) - E_j(t) - I_j(t) - R_j(t),\\ \dot{E}_j(t) = (1-\mu(t))\beta_j S_j(t)I_j(t) - \sigma_j E_j(t),\\ \dot{I}_j(t) = \sigma_j E_j(t) - (\mu_j+\gamma_j)I_j(t),\\ \dot{R}_j(t) = \gamma_j I_j(t) - \delta_j R_j(t) \end{cases}$$

This approach includes an overall population P term for determining the Susceptible compartment, mitigation measures μ, recovery rate γ, and immunity loss δ. This approach provides greater insight than the previous models, but still lacks potentially useful information for compliance graph analysis. In this model, recovered individuals in the system do not eventually become susceptible again, and also do not face the risk of expiring from the disease. In real systems, once an asset is compliant with a regulation, it does not become immune to violating that regulation in the future; an asset is fully able to repeatedly fall out of compliance. In addition, an alternative approach to corrective recovery (namely for legacy systems, third-party vendors or applications, or other systems that carry excess risk) is to remove the system from the environment. In epidemiology modeling, this would be represented as a system expiring or compartmentalized as "Deceased". Though this model could provide additional parameters and compartments, it represents a solid foundational starting point for building a compatible model for compliance graph analysis, especially due to the detailed equilibrium, stability, and optimal control problem proofs. Section III-B discusses the details of the chosen model.

B. SEIRDS Model

For initial investigation and testing purposes, the selected epidemiology model for the single-strain compliance graph analysis (which has been adapted for the multi-strain approach) was a SEIRDS (Susceptible - Exposed - Infected - Recovered - Deceased) model. Unlike a SIR (Susceptible - Infected - Recovered) model, the SEIRDS model includes an exposed group to represent nodes that *will* become infected, as compared to the susceptible group which *can* become infected. The exposed group has additional parameters such as incubation time, which provide additional detail on when the group will transition to the infected group. This model also captures events where nodes in the infected group decease rather than recover. After recovery, there is a waning immunity period where the population can become susceptible again. A similar model can be seen in the works presented by the authors of [16]. This model better fits the information present in a compliance graph than the models described in Section III-A, and the contextualization of this model's compartments and parameters can be seen in Tables I and II. Figure 1 shows a block model representation of the SEIRDS model. Equation 1 displays the SEIRDS representation.

TABLE I: Compartment Descriptors for Compliance Graphs. Each compartment is contextualized to compliance graphs, with their meanings and brief explanations of how nodes may fit in compartments.

Compartment	Description	Contextualization to Compliance Graphs
S	Susceptible	All other nodes.
E	Exposed	Nodes flagged as at risk of compliance violation by an IDS, timeframe trigger, or other.
I	Infected	Nodes that are in violation.
R	Recovered	Nodes that were infected and were capable of automatic correction e.g., certificate renewal, scheduled maintenance, or other.
D	Deceased	Nodes removed from the compliance graph; Systems that were removed from the network through quarantine, DMZ, legacy removal, or other due to excess risk.

TABLE II: Parameter Descriptors for Compliance Graphs. Each parameter is contextualized to compliance graphs, with their meanings and brief explanations of how node relationships affect the parameters.

Parameter	Description	Contextualization to Compliance Graphs
β	Infection Rate	Probability of nodes falling out of compliance.
δ	Incubation Period	Once a node is at risk of falling out of compliance, how long it takes to actually violate a mandate or regulation.
γ_R	Recovery Rate	Probability of a system to correct its violation status.
γ_D	Death Rate	Probability of removal for a system in violation.
μ	Fatality Ratio	Natural rate at which any node may be removed from the network.
ϵ	Infected Import Rate	Systems that are already in violation; systems that do not fall out of compliance, but are already violating a mandate.
ω	Waning Immunity Rate	After a system recovers, how long for it to be available for violation.

$$
\begin{aligned}
dS(t) &= \epsilon - (\beta * S(t) * I(t)) - \mu * S(t) + \omega * R(t) \\
dE(t) &= (\beta * S(t) * I(t)) - \delta * E(t) \\
dI(t) &= \delta * E(t) - \gamma_d * I(t) - \gamma_r * I(t) \\
dR(t) &= \gamma_r * I(t) - \omega * R(t) \\
dD(t) &= \gamma_d * I(t) + \mu * S(t)
\end{aligned} \tag{1}
$$

Converting the single-strain model to function as a multi-strain model requires minimal additional modeling efforts. Rather than using a single SEIRDS model for all violations, each violation is instead represented by a SEIRDS model. In an environment where there are n violations, there are n SEIRDS models. This is represented in Table III, which adds an indexing term to allow for multiple violations to be represented in the SEIRDS model.

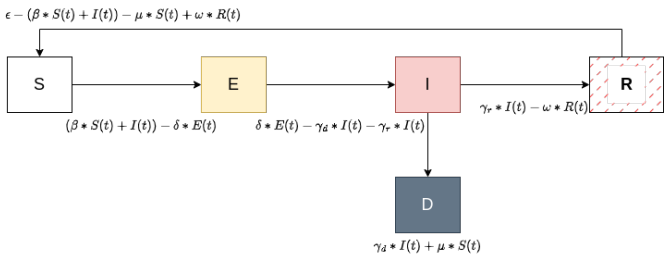

Fig. 1: SEIRDS Epidemiology Model. Display of each compartment in the SEIRDS epidemiology model and the mathematical representation of the relationship between compartments along with the parameters.

TABLE III: Multi-Strain Variable Descriptors for Compliance Graphs. Each variable from the single-strain model can be represented with an indicator of n to denote which SEIRDS violation model the data falls into.

Model Variable	Description
S_n	Susceptible to Violation n
E_n	Exposed to Violation n
I_n	Infected with Violation n
R_n	Recovered from Violation n
D_n	Deceased from Violation n
β_n	Infection Rate for Violation n
δ_n	Incubation Period for Violation n
$\gamma_R n$	Recovery Rate for Violation n
$\gamma_D n$	Death Rate for Violation n
μ_n	Fatality Ratio for Violation n
ϵ_n	Infected Import Rate for Violation n
ω_n	Waning Immunity Rate for Violation n

C. Implementation Details

Many multi-strain epidemiology models are made difficult or are constrained due to strain dependencies, competition, or extinction. In many multi-strain models, two strains A and B are not independent. If an individual is infected with strain A or B, it may prevent the onset of the other strain, increase the likelihood of infection, or decrease the likelihood of infection. These dependencies often have their own dynamics, and can change over time along with the model. These strains also often compete, and strain effects can be dampened or amplified based on other strain populations. With numerous competing strains, the modeling process grows increasingly difficult or infeasible. However, a multi-strain SEIRDS model for compliance graph analysis is able to be simplified. In compliance graph analysis, a simplification can be made that each strain is fully independent - an infection (or lack thereof) of a strain does not impact any other strain. For example, for the automobile maintenance network, a node in the compliance graph can be simultaneously in need of an oil change and a headlight replacement. In addition, a strain curve can be said to have no impact on any other strain curves. Though this

simplification is not fully representative and yields inaccuracy for multi-step or complex, dependent-triggered violations, it serves as a starting foundation for multi-violation analysis in a system with common violations.

This work made use of both Python and R to support the epidemiology modeling. The derivation of parameters and compartmentalization of nodes for the SEIRDS model was implemented using Python and the NetworkX [17] library. In addition to deriving model parameters and compartmentalizing the input compliance graph, the model preparation supports the ability to split the compliance graph into subgraphs based on time steps. Given the prior-knowledge network, the model preparation is able to identify time transitions within the graph and create subgraphs. As part of the main preparation loop, the Python function is able to take input as either the entire compliance graph or a subgraph. This preparation loop iterates through all known violations, and compartmentalizes each node according to each violation. Due to the simplification, it is possible for a node to simultaneously be classified as infected with violation A while being classified as recovered from violation B. For each violation, each node is classified as infected if a violation edge leads to the node, exposed if a node is *n*-steps away from an infected node, recovered if a recovery edge leads to the node, deceased if the node has no children, or susceptible if no other conditions match. After all compartments have been identified, parameters are obtained. The parameter definitions are presented in Equation 2. The compartmentalization and parameter derivation process is shown in Algorithm 1.

$$\begin{aligned}
\beta_n &= \frac{Infected_n * w}{len(CG.nodes)} \\
\delta_n &= 1 \\
\gamma_{r_n} &= \frac{Recovered_n * w}{len(CG.nodes)} \\
\gamma_{d_n} &= \frac{Deceased_n * w}{len(CG.nodes)} \\
\mu_n &= \frac{Deceased_n}{Deceased_n + Recovered_n} \\
\epsilon_n &= \frac{Imported - Infected_n}{len(CG.nodes)} \\
\omega_n &= 1
\end{aligned} \tag{2}$$

This work used the compliance graphs generated from the work presented by the authors of [**?**]. These example graphs are publicly available and serve as foundational test cases for experimental work. Since these example graphs are generated across multiple industries and have unique properties, it allows for a more thorough investigation and analysis of this work. For these example networks, certain model parameters are statically defined. For instance, in Equation 2 both δ and ω are set to 1. As Table III specifies, δ is the incubation period, which defines the incubation period, or how many compliance graph steps (edges) are required before a node can be classified as infected. In a compliance graph format, and specifically for the example networks presented in the work of [**?**], there will always be exactly 1-edge delays between the labeling of an infection and the classification of an infection. Since the edges in a compliance graph are used as the transitionary elements that lead to a change in a system, for node B to be classified as infected, there must be at least one edge between node A and node B for the infected classification to occur. This is also the case for the waning immunity rate, ω. Since there must be at least one edge to transition a node from a recovered state to a susceptible state, the waning immunity rate is also set to a static value of 1.

Algorithm 1: SEIRDS Model compartmentalization and Parameter Derivation

Input : Compliance Graph, PKN, Exposure Reachability n
Output: SEIRDS Model Compartments and Parameters

STEP 1: Loop through Compliance Graph and Compartmentalize Nodes Based on Violation.
for *node in Compliance Graph* **do**
  for *violation in possibleViolations* **do**
    if *len(node.in-edges) == 0 and node is infected with violation* **then**
      do set importedInfection[$violation$] += 1;
    if *len(node.out-edges) == 0 and node is infected with violation* **then**
      do set Deceased[$violation$] += 1;
    else if *any node.in-edge has a violation identifier == violation* **then**
      do set Infected[$violation$] += 1;
    else if *node has n-step reachability to a node infected with violation* **then**
      do set Exposed[$violation$] += 1;
    else if *any node.in-edge has a recovery identifier == violation* **then**
      do set Recovered[$violation$] += 1;
    else
      do set Susceptible[$violation$] += 1;
STEP 2: Obtain Model Parameters for each Violation.
for *violation in possibleViolations* **do**
  do set $violation$[parameters] = Equation 2

In order to better capture the recurring maintenance events, the recovery parameters can be adjusted based on a periodic function. This function can be adjusted per component and/or per mitigation strategy, and is intended to be interchanged with ease. For this work, the recovery parameters were altered

to fit a cosine function. Cosine was chosen in order for the recover parameter minimum to be placed at the final known time step (t=0). A sine function with appropriate shifts could also be used, however a cosine function was used for simplicity. A few adjustments are required. To set the period of the cosine function, the default period (2π) is divided by the recurring maintenance event's time step of occurrence. Since the recovery rate must be a value between 0 and 1, a piecewise function is defined for the magnitude, where $y = max(0, cosx) \forall x > 0$. The amplitude of the function was defined by the derived parameter from Algorithm 1. This is shown in Equation 3.

$$\gamma_r = \gamma_{r,derived} * \max(0, cosine(\frac{2 * \pi * t}{RecurringMaintenanceTimeStep})) \quad (3)$$

For accurately capturing the pre-defined events presented in the compliance graphs of [**?**], the infection rate parameters are also designed to be dynamic. Due to the degradation of parts, health, or quality of the asset over time, the infection parameter is expected to steadily increase. The pre-defined events can be used as a baseline for increasing the infection rate to support this functionality. For instance, for the automobile maintenance network, the drive belts have a pre-defined event at time step 3 where a replacement or inspection is required. A function can be defined that increases the infection rate of the drive belt violation as the component reaches its replacement or inspection lifespan. When data is available for the lifespan of a given component, an infection rate parameter increase could model the behavior of $\lim_{t \to expectedPartLifeSpan} f(t) = 1$. For components with pre-defined events, the "expectedPartLifeSpan" can be obtained by identifying the event time of the part failure, repair, or replacement. For components without a pre-defined event, identifying the "expectedPartLifeSpan" would require additional data for each component represented in the compliance graph. As previously stated, this work intends to be self-contained and independent of likelihood estimations. Though tire replacement data may be readily available, identifying the likelihood of forcibly decrypting a database (or the point of its occurrence) requires transitional estimation that is outside of this work. Instead, this work made a simplification of $\lim_{t \to \infty} f(t) = 1$ when the component lifespan data was unavailable.

This work intended for $f(t)$ to be fully replaceable. Better cost models can replace the $f(t)$ functions to better fit to the expected changes in infection parameters. Rather than identifying unique $f(t)$ functions for each component, this work implemented two functions. When the expected component lifespan was known, Equation 4 was used as the $f(t)$ function. When the expected component lifespan was not known, Equation 5 was used as the $f(t)$ function. Equation 4 was intended to be simplistic. Due to the exponential function, the limit approaches 1 as the component reaches its lifespan. There is a shift in the function for axis alignment, but no other modifications are implemented. This function does have a quick rise toward 1, so future adjustments to its gain may be necessary. For Equation 5, more attention was given to its rise. Since the component has an unknown life span, a gradual increase toward 1 at $t = \infty$ may be unrealistic. Likewise, a rapid gain toward 1 in the first few time steps may also be unrealistic. This function was chosen and tuned to allow for a rise toward a infection rate parameter value of 1 near 30 time steps. Since time steps are variable based on the step size selection in the network generation process, this function can be tuned to better fit specific components or networks. Figure 2 displays a plot of this function. After a mitigation or correction scheme is implemented for a component, the infection rate is reset to its original, derived value.

$$f(t) = 1 - \exp((expectedPartLifeSpan - t)) \quad (4)$$

$$f(t) = 1 - \exp(-\frac{t}{1000}^{2.5}) \quad (5)$$

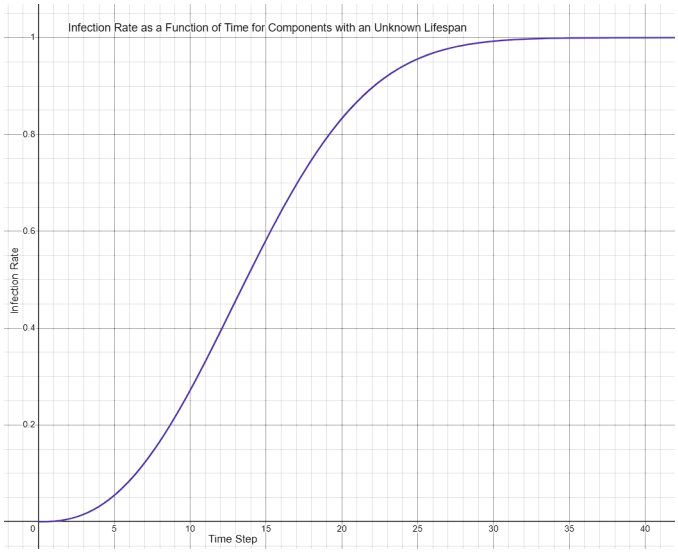

Fig. 2: The infection rate parameter increases in value over time to account for degradation of parts, health, or quality of components. This Figure displays the chosen function for components with unknown lifespans. It is tuned to rise toward its limit value of 1 near 30 time steps. This function is able to be modified or replaced to better fit the degradation factor for other components or environments.

After the Python model preparation was called with the Reticulate library [18] (which acts as an interface between

R and Python), R code was used to evaluate the problem space using a differential equation solver. The implemented solver was ODE45, provided by the pracma library [19]. Equation 1 was defined as a function in R, with inputs of model parameters and model compartments. This function was called through ODE45 over a specified range of time, and would produce an output of epidemic curves. Unless otherwise specified, the epidemic curves would be generated starting with time = "0" (the last known step of the compliance graph), and would evaluate data for the next 5 years. Since this work is approaching the compliance graph through a multi-strain lens, the differential equation solver was applied over each violation. As a result, epidemic curves were produced for each violation that provided insight on the violation trends over time. This process is displayed in Algorithm 2.

Algorithm 2: Producing Epidemic Curves for all Possible Violations in a Compliance Graph

Input : Compliance Graph, PKN, Exposure Reachability n
Output: Epidemic Curves

STEP 0: Initialize the R environment.
do initialize:
  Set working directory;
  Receive inputs;
  Initialize Reticulate with appropriate Python versioning and executable(s);
STEP 1: Compartmentalize $ComplianceGraph$ and Derive Model Parameters.
do call Algorithm 1 via Reticulate;
STEP 2: Format Compartments and Parameters.
do unpack $STEP$ 1 output and load into R variables;
do data type conversions and shaping across time steps and violations;
STEP 3: Produce Epidemic Curves.
for $violation$ *in dim(S)[2]* **do**
  do set vioOut = ODE45 with compartments=SEIRDS.compart[$violation$], parameters=SEIRDS.param[$violation$], time=[0, 5 years];
  do plot vioOut;

D. Results and Analysis

Since a large number of epidemic curves were produced (due to the large number of possible violations across three example networks), only select epidemic curves are displayed and discussed. The selected Figures display the variations in curves based on implemented mitigation strategies. Figure 3 highlights the effectiveness of the routine maintenance for the automobile maintenance network. This Figure displays the decreases in compliance violations for oil changes due to the maintenance that occurs every 6 months or 6000 miles. Likewise, Figure 4 highlights the effectiveness of the routine maintenance that takes place every 1 year and 6 months for the drive belts. Both of these Figures successfully showcase the return to compliance due to the implemented repair strategies. Figure 5 illustrates an output for an ineffective mitigation strategy for the HIPAA network. This Figure showcases that the lack of mitigation and correction implementations leads to a rise in violations, which will be attributed to noncompliance penalties. This Figure successfully showcases the necessity for organizational attention regarding this specific violation.

Another advantage to epidemiology modeling is the ability to identify, predict, and prepare for outcomes if maintenance schedules are adjusted. Figure 6 illustrates an example for the OSHA 1910H network, where the routine maintenance schedule for a component is ended. At the last known time step (t=0), it is evident that the previous maintenance schedule was successful at correcting any known violations: the Susceptible population was at 90.9%, the Exposed, Infected, and Deceased populations were near zero, and the Recovery population was non-zero, indicating that the maintenance schedule was successfully correcting any issues with noncompliance as they arose. Parameters were then altered to determine if maintenance schedules could be removed. For this example, the alteration included setting γ_r to the derived value, rather than based on a function of time relative to the maintenance schedule since it was removed. Figure 6 highlights that the removal of the maintenance schedule would cause an increase in the Infected population over the next 5 years, revealing that the maintenance schedule should likely remain in place.

Some Figures also displayed results with no changes in compartment trends. Figure 7 illustrates an epidemic curve that, though accurate, yields less interesting results. This Figure showcases the trendlines of a violation that is not seen in the given compliance graph, and there is no modeled degradation of components. Since this violation never occurs, the Susceptible population remains at 100%, with all other compartments remaining at 0%. However, this result provides useful insight since it demonstrates that there are no current or expected noncompliance penalties that will be incurred due to this specific violation. These Figures are able to assist in the compliance maintenance process by verifying that budgetary attention can be safely allocated to other, higher priority violations.

E. Validation

In order to validate the epidemiology modeling approach, the following characteristics were examined, and test cases were created to compare against expected behavior. The results of these tests are not included in this work, since the test results were a boolean "pass" or "fail". If a failed test was encountered, the validation process failed, and the

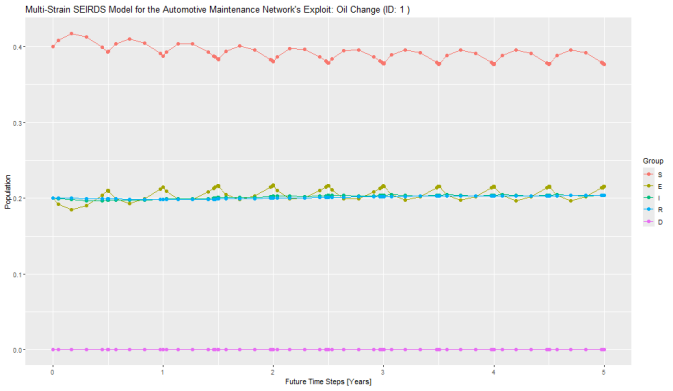

Fig. 3: Epidemic Curve for the Automobile Maintenance Network's Oil Change Violation. This epidemic curve highlights the effectiveness of the routine maintenance for the automobile maintenance network, evident through the regular decreases in the Infected compartments of the SEIRDS epidemiology model. These maintenance events take place every 6 months or 6000 miles, and successfully correct the vehicle to a state of compliance with respect to the oil change violations.

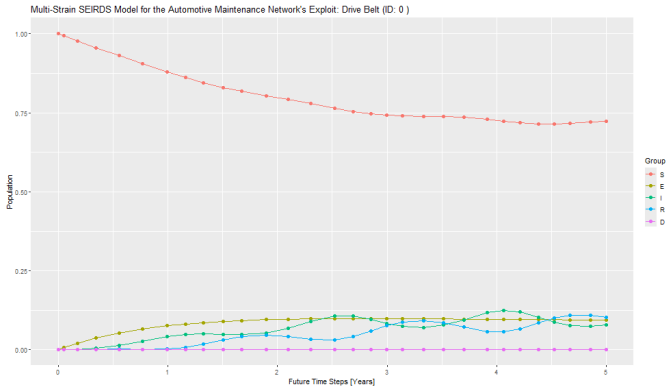

Fig. 4: Epidemic Curve for the Automobile Maintenance Network's Drive Belt Violation. This epidemic curve highlights the effectiveness of the routine maintenance for the automobile maintenance network, evident through the regular decreases in the Infected compartments of the SEIRDS epidemiology model. These maintenance events take place every 1 year and 6 months, and successfully correct the vehicle to a state of compliance with respect to the drive belt violations. However, there is a steady increase of violation (the population of I) due to the increased infection rate parameter caused by the pre-defined events presented in the compliance graphs of [**?**]. This Figure also displays the effectiveness at preventing the spread of a drive belt violation. If the Infected population rose to 1.0, then every node in a compliance graph or subgraph would be expected to posses that violation. Since the Infected population remains below 10%, the mitigation strategy can be considered effective at correcting the drive belt violation before it spreads through the compliance graph or subgraph.

methodology was flawed and in need of correction. For the work presented, each test returned a successful outcome.

- All parameters (β, δ, γ_r, γ_d, μ, ϵ, ω) for all violations are greater than or equal to 0.0.
- The β, γ_r, γ_d, μ, ϵ, ω parameters for all violations are less than or equal to 1.0.
- All initial compartments (S, E, I, R, D) are greater than or equal to 0.0.
- The sum of all compartments (S, E, I, R, D) for all time steps for all violations does not exceed the total number of nodes for the given time-stepped subgraph.
- The sum of all compartments (S, E, I, R, D) in percentage form for all time steps for all violations is equal to 1.0.
- A decrease in the $Susceptible$ compartment causes an increase in either the $Exposed$ or $Infected$ compartment.
- A decrease in the $Exposed$ compartment causes an increase in the $Infected$ compartment.
- A decrease in the $Infected$ compartment causes an increase in either the $Recovered$ or $Deceased$ compartment.
- A decrease in the $Recovered$ compartment causes an increase in the $Susceptible$ compartment.
- A increase in the $Deceased$ compartment causes a decrease in the total number of nodes.
- The total sum of the $Deceased$ compartment across all time steps never decreases.
- The number of nodes in the epidemiology model for the specific example networks in this work never increases.

IV. Risk Assessment

Epidemiology modeling successfully provides insight on the predicted trends and behaviors of compliance violations over time. These trends are useful for identifying the utility of any correction schemes, for future mitigation planning, or for determining which periods of time may require additional attention to prevent noncompliance. However, for presenting quantitative metrics to advisory boards, upper level staff, or management, alternative figures may yield better results, especially when requesting increases in funding. One common method for determining a quantitative metric with respect to expected or predicted challenges is risk assessment. Risk assessment allows for further insight regarding potential risks, damages, or losses, and considers any mitigations, likelihood of occurrences, and can provide information on fault-tolerance in the instance of a risk event [20]. This Section discusses the

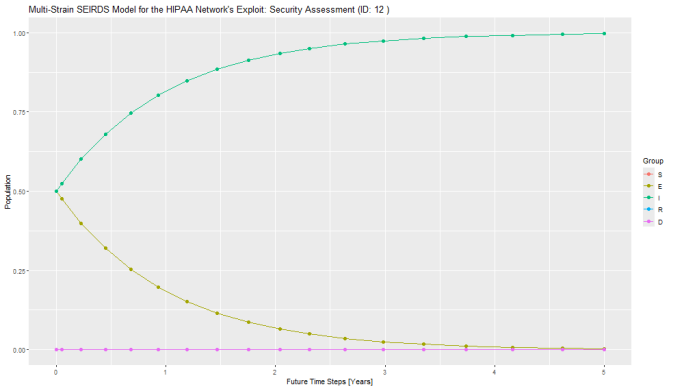

Fig. 5: This epidemic curve highlights the ineffectiveness of the mitigation strategy for the HIPAA network, evident through the continual increase and lack of any decrease in the Infected compartments of the SEIRDS epidemiology model. There are seemingly no mitigation or correction implementations, or they are unable to successfully correct the HIPAA network to a state of compliance with respect to the security assessment violation. At the final known time step (t=0), nearly 50% of the nodes in the last subgraph contained this violation, with the remainder within 2-step exposure of infection.

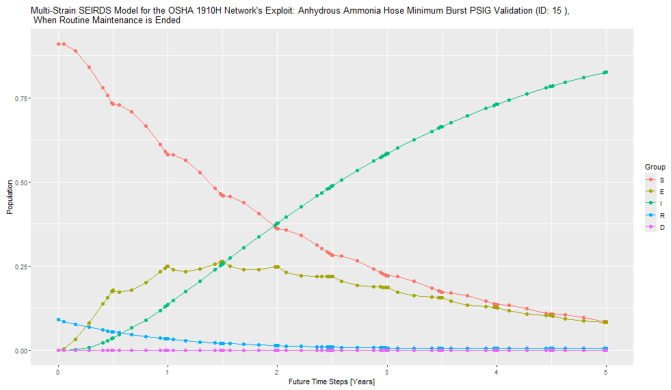

Fig. 6: This Figure illustrates an example for the OSHA 1910H network where the routine maintenance schedule for a component is ended. At the last known time step (t=0), it is evident that the previous maintenance schedule was successful at correcting any known violations: the Susceptible population was at 90.9%, the Exposed, Infected, and Deceased populations were near zero, and the Recovery population was non-zero, indicating that the maintenance schedule was successfully correcting any issues with noncompliance as they arose. Parameter alters were then altered to determine the outcome if maintenance schedules could be removed. For this example, γ_r was set to be the derived value, rather than based on a function of time relative to the maintenance schedule. This Figure highlights that the removal of the maintenance schedule causes an increase in the Infected population over the next 5 years.

risk assessment strategies that are implemented for compliance graph analysis in Section IV-A, with results presented in Section IV-B.

A. Implementation Details

A primary motivator for this work is to provide analysis on compliance graphs that is independent of (but compatible with) any external probability or state transition estimations. Though works exist for determining transitions of nodes and edges through Markov Decision Processes, empirical analysis, or investigations of the National Vulnerability Database (NVD) and Common Vulnerabilities and Exposures (CVE) list, this work aimed to stay as self-contained and self-sufficient as possible. In order to achieve this, this work presents a modification to Annualized Loss Expectancy (ALE) computations. Equation 6 displays the traditional approach for computing ALE.

$$ALE = AnnualRateofOccurrence * SingleLossExpectancy, \quad where: \\ SingleLossExpectancy = AssetValue * ExposureFactor \tag{6}$$

Included in the presentation of ALE in Equation 6, Annual Rate of Occurrence (ARO) is displayed. In traditional computations of ALE, this is obtained through a likelihood estimation. Since this work is intended to stay self-contained,

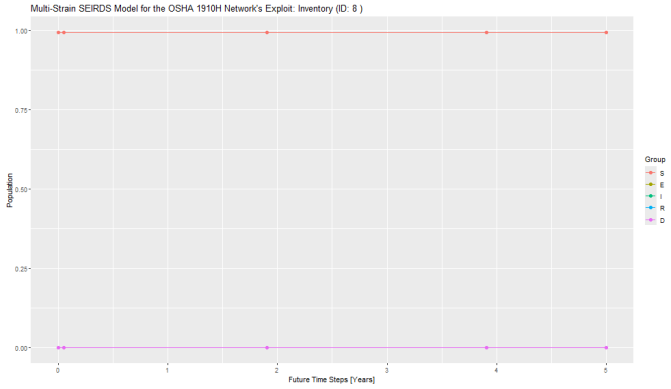

Fig. 7: This epidemic curve displays results with no changes in compartment trends for the OSHA 1910H network due to the lack of any known or expected inventory violations. This epidemic curve verifies that budgetary attention can be safely allocated to other, higher priority violations.

an alternative approach is required. The SEIRDS model presented in Section III-B included a discussion of model parameters. Included with these parameters was a derivation of the β parameter in Algorithm 1 and Equation 2. As Table III displays, the β parameter is used to describe the infection rate of a violation. This parameter, when presented as a yearly rate, can be directly substituted for ARO in the ALE computation. Likewise, the Noncompliance Penalty can be obtained through the prior-knowledge network, which describes the penalties for noncompliance. This is shown in Equation 7 as the Single Loss Expectancy. Since this work specifically does not estimate the probabilities of fines or assessments, the Exposure Factor is set to 100%. This work aims to ensure compliance is fully met, and does not introduce any analysis that allows for organizations to base their mitigation strategies based on the likelihood of being fined. This work is binary in compliance measurement only - either the organization is maintaining a compliant standing and will not be fined, or they are in violation and will receive a penalty.

$$ALE_n = \beta_n * NoncompliancePenalty \tag{7}$$

Using the modified ALE equation, a process can be defined for obtaining necessary components. The β parameter is obtained from the SEIRDS model parameter derivation, which receives input of a compliance graph and the prior-knowledge network. Figure 8 shows the flow diagram of computing risk assessment with ALE given a compliance graph and its input files, operating on the SEIRDS model parameters.

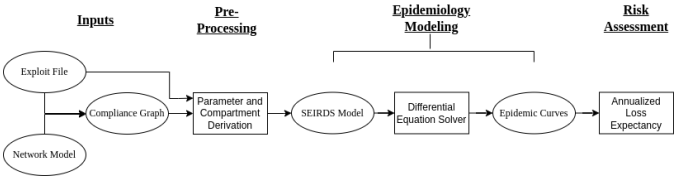

Fig. 8: Risk Assessment Using Epidemiology Modeling. Display of the risk assessment process. Using a compliance graph and its input files, the parameters and compartments can be derived for a SEIRDS model as discussed in Section III. Using the produced multi-strain epidemic curves, Annualized Loss Expectancy can be used to quantify the costs of compliance or regulation mandate violations.

B. Results and Analysis

For each example network presented in [**?**], ALE was computed for all possible violations. To display the results, bar graphs were used to highlight the magnitudes of expected losses incurred from an event of noncompliance. This display allows for a simplistic view of violations that are expected to yield unfavorable penalties, as well as those that may yield little to no penalties. The bar graphs were plotted in terms of both monetary and time penalties. Figures 9 and 10 present the results for the Automobile Maintenance Network for its monetary and time expected losses, respectively. Figures 11 and 12 present the results for the HIPAA Network for its monetary and time expected losses, respectively. Figures 13 and 14 present the results for the OSHA 1910H Network for its monetary and time expected losses, respectively.

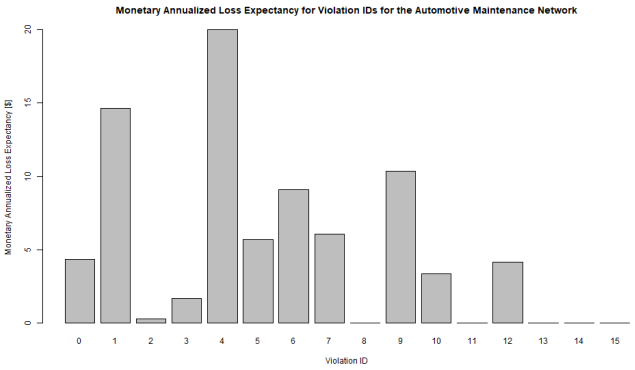

Fig. 9: Annualized Monetary Loss Expectancies for the Automobile Maintenance Network. Bar graph representation of the expected monetary losses across all violations for the automobile maintenance network per year. Each loss is computed through the modified ALE equation presented in Equation 7. Numerous violations have no expected losses due to the lack of presence of violation in the input compliance graph, or due to successful mitigation strategies. This is a favorable result, as it allows for provable visuals that current mitigation strategies are successful in preventing any penalties for a particular mandate, or that there is no current risk of noncompliance penalties for the possible violation.

There are a few notable points of interest in each of these Figures. ALE was computed for all possible violations, and no filtering or post-processing was performed. Due to this, numerous violations list no expected losses. Though these violations were set as possibilities in the analysis methods, since there were no occurrences of noncompliance in the input compliance graph, the β parameter was derived to be 0, and the resulting ALE was likewise 0. This is a favorable result, as it allows for provable visuals that current mitigation strategies are successful in preventing any penalties for a particular mandate.

In all Figures that presented expected time losses, magnitudes are relatively low. This is due to the input of prior-knowledge networks for each example network. For all problem spaces, identifying monetary penalties was made easier due to public availability of data. However, data describing time losses due to temporary shutdowns or

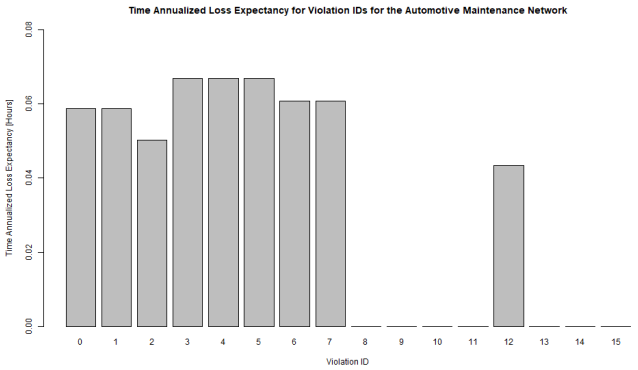

Fig. 10: Annualized Time Loss Expectancies for the Automobile Maintenance Network. Bar graph representation of the expected time losses across all violations for the automobile maintenance network per year. Each loss is computed through the modified ALE equation presented in Equation 7. Numerous violations have no expected losses due to the lack of presence of violation in the input compliance graph, or due to successful mitigation strategies. This is a favorable result, as it allows for provable visuals that current mitigation strategies are successful in preventing any penalties for a particular mandate, or that there is no current risk of noncompliance penalties for the possible violation. However, of the possible violation occurrences, the expected losses are of low magnitude due to the availability of ride-sharing or rental cars that was built into the monetary cost model.

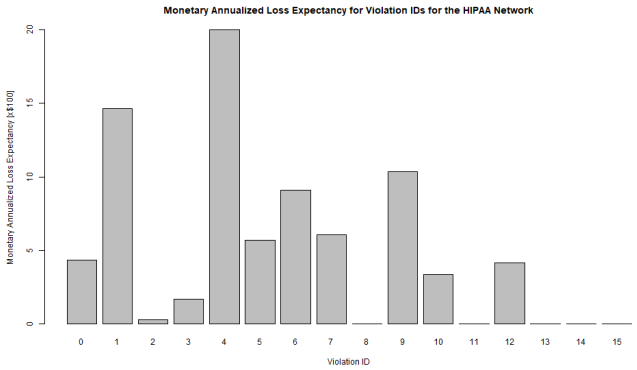

Fig. 11: Annualized Monetary Loss Expectancies for the HIPAA Network. Bar graph representation of the expected monetary losses across all violations for the HIPAA network per year. Each loss is computed through the modified ALE equation presented in Equation 7. Numerous violations have no expected losses due to the lack of presence of violation in the input compliance graph, or due to successful mitigation strategies. This is a favorable result, as it allows for provable visuals that current mitigation strategies are successful in preventing any penalties for a particular mandate, or that there is no current risk of noncompliance penalties for the possible violation.

closures was not readily available. Rather than estimate data, create data, or insert bias, time losses were generally not present in the prior-knowledge networks. For the Automobile Maintenance network, time losses were available due to the expected time of repair, but these losses tended to be overshadowed by monetary losses. Even in instances where repairs were lengthy, costs of public transportation or ride-sharing were built into monetary losses, rather than time losses; individuals were more likely to drop a vehicle off for repair with a return at a later date for pickup. Given more data regarding time losses, these Figures could be expected to see greater variances in magnitude.

C. Validation

In order to validate ALE, the following characteristics were examined, and test cases were created to compare against expected behavior. The results of these tests are not included in this work, since the test results were a boolean "pass" or "fail". If a failed test was encountered, the validation process failed, and the methodology was flawed and in need of correction. For the work presented, each test returned a successful outcome.

- The β parameter for all violations is greater than or equal to 0.0.
- The β parameter for all violations for a single time step is less than or equal to 1.0.
- If the β parameter for any violation when substituted for ARO is greater than 1, then at least two edges are present in the compliance graph when the compliance graph is split into a subgraph across one year.
- All ALE values are greater than or equal to 0.0.
- If the β parameter for a violation is greater than or equal to 0, and is less than or equal to 1, then the resulting ALE does not exceed the costs present for that violation in the prior-knowledge network.
- If the β parameter for a violation is greater than 1, then the resulting ALE does exceed the costs present for that violation in the prior-knowledge network.
- The resulting bar graphs for displaying ALE have an equal number of bars to the number of total possible violations for the example network.

V. Future Works

VI. Conclusions

This work presented and implemented a SEIRDS epidemiology model used for analyzing the trends (and

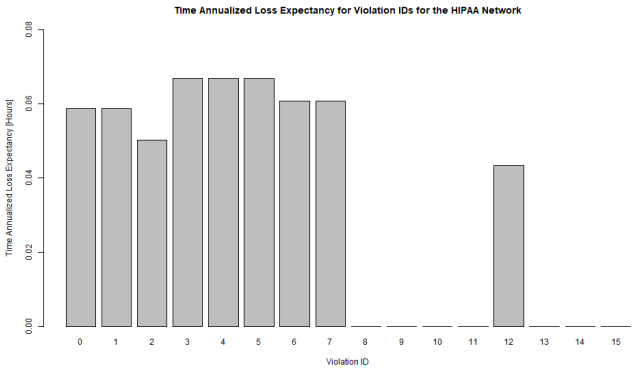

Fig. 12: Annualized Time Loss Expectancies for the HIPAA Network. Bar graph representation of the expected time losses across all violations for the HIPAA network per year. Each loss is computed through the modified ALE equation presented in Equation 7. Numerous violations have no expected losses due to the lack of presence of violation in the input compliance graph, or due to successful mitigation strategies. This is a favorable result, as it allows for provable visuals that current mitigation strategies are successful in preventing any penalties for a particular mandate, or that there is no current risk of noncompliance penalties for the possible violation. However, of the possible violation occurrences, the expected losses are of low magnitude due to the lack of closure or shutdown penalty data.

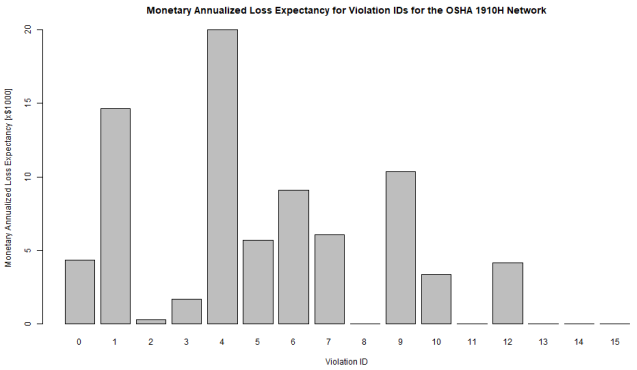

Fig. 13: Annualized Monetary Loss Expectancies for the OSHA 1910H Network. Bar graph representation of the expected monetary losses across all violations for the OSHA 1910H network per year. Each loss is computed through the modified ALE equation presented in Equation 7. Numerous violations have no expected losses due to the lack of presence of violation in the input compliance graph, or due to successful mitigation strategies. This is a favorable result, as it allows for provable visuals that current mitigation strategies are successful in preventing any penalties for a particular mandate, or that there is no current risk of noncompliance penalties for the possible violation.

expected trends) of a compliance graph. This model is able to compartmentalize a compliance graph into five unique and related sets of populations. This model also functions with a variety of tunable, customizable parameters that assist in the analysis of the spread of violations in an environment. These parameters are intended to be used alongside any user-provided data, such as the maintenance or replacement schedules or the expected lifespans of components. Each example compliance graph presented in [**?**] was analyzed using this model. The results were validated as part of the validation process shown in Section III-E. The results presented in Section III-D display notable visualizations about the analysis process. The methodology of this work is able to uncover valuable information regarding the compliance standing of an environment, such as the effectiveness of any mitigation, correction, or prevention strategies, and the exposure to and spread of violations. Additionally, this work allows for the testing and experimentation of mitigation strategies to provide quantifiable results as to the successes or failures of any alterations, removals, or additions toward maintaining compliance. This approach is shown to be successful in all three unique example compliance graphs, and boasts the ability to be modified or fine-tuned given additional data inputs.

In addition, this work presented an approach for risk assessment through a modified Annualized Loss Expectancy (ALE) computation. This technique allows for analysis on a complex environment without the need for estimation regarding rate of occurrence that may otherwise be difficult or impossible to obtain. The results as shown in Section IV-B were presented in the form of bar graphs that display the expected losses for each resource across all examined violations. The results were validated as part of the validation process shown in Section IV-C, and demonstrate that this approach is functional even when no noncompliance instances are observed.

REFERENCES

[1] S. Ahn and M. Kwon, "Reproduction factor based latent epidemic model inference: A data-driven approach using covid-19 datasets," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1259–1270, 2023.

[2] T. J. Roy, M. A. Mahmood, A. Mohanta, and D. Roy, "An analytical approach to predict the covid-19 death rate in bangladesh utilizing multiple regression and seir model," in *2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*, pp. 42–45, 2021.

[3] D. Chumachenko, K. Bazilevych, I. Meniailov, S. Yakovlev, and T. Chumachenko, "Simulation of covid-19 dynamics using ridge regression," in *2021 IEEE 4th International Conference on Advanced Information and Communication Technologies (AICT)*, pp. 163–166, 2021.

[4] S. Zhang and H. Yang, "Spatial modeling and analysis of human traffic and infectious virus spread in community networks," in *2021 43rd*

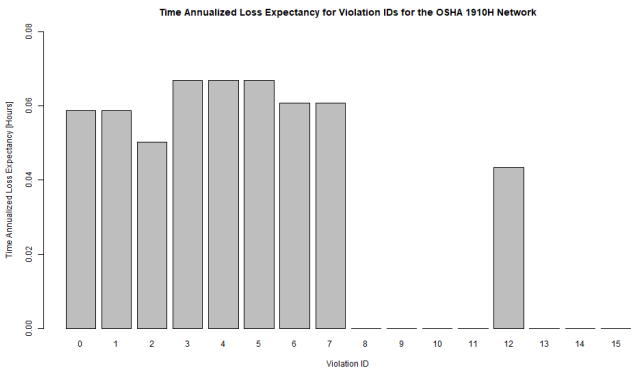

Fig. 14: Annualized Time Loss Expectancies for the OSHA 1910H Network. Bar graph representation of the expected time losses across all violations for the OSHA 1910H network per year. Each loss is computed through the modified ALE equation presented in Equation 7. Numerous violations have no expected losses due to the lack of presence of violation in the input compliance graph, or due to successful mitigation strategies. This is a favorable result, as it allows for provable visuals that current mitigation strategies are successful in preventing any penalties for a particular mandate, or that there is no current risk of noncompliance penalties for the possible violation. However, of the possible violation occurrences, the expected losses are of low magnitude due to the lack of closure or shutdown penalty data.

Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2286–2289, 2021.

[5] H. Kim and R. Anderson, "An experimental evaluation of robustness of networks," *IEEE Systems Journal*, vol. 7, no. 2, pp. 179–188, 2013.

[6] R. Mitchell, "Epidemic-resistant configurations for intrusion detection systems," in *2017 IEEE Conference on Communications and Network Security (CNS)*, pp. 487–494, 2017.

[7] N. Dakhno, O. Leshchenko, Y. Kravchenko, A. Dudnik, O. Trush, and V. Khankishiev, "Dynamic model of the spread of viruses in a computer network using differential equations," in *2021 IEEE 3rd International Conference on Advanced Trends in Information Theory (ATIT)*, pp. 111–115, 2021.

[8] B. Shan, "The spread of malware on the wifi network: Epidemiology model and behaviour evaluation," in *2009 First International Conference on Information Science and Engineering*, pp. 1916–1918, 2009.

[9] Y. Tang and R. A. Williams, "Investigating relationship conflict within the social network of large is projects using a sir model," in *2022 IEEE International Symposium on Technology and Society (ISTAS)*, vol. 1, pp. 1–5, 2022.

[10] M. A. Parwez, M. Abulaish, and J. Jahiruddin, "A social media time-series data analytics approach for digital epidemiology," in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 852–859, 2020.

[11] D. Mathebula, "Novel data-based model for future epidemiology," in *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1–6, 2022.

[12] Y. Wang, K. N. Plataniotis, J. Z. Wang, M. Hou, M. Zhou, N. Howard, J. Peng, R. Huang, S. Patel, and D. Zhang, "The cognitive and mathematical foundations of analytic epidemiology," in *2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, pp. 6–14, 2020.

[13] D. Fedorov, Y. Tabarak, A. Dadlani, M. S. Kumar, and V. Kizheppatt, "Dynamics of multi-strain malware epidemics over duty-cycled wireless sensor networks," in *2021 International Balkan Conference on Communications and Networking (BalkanCom)*, pp. 1–5, 2021.

[14] Y. Lou and R. B. Salako, "Control strategies for a multi-strain epidemic model," *Bulletin of Mathematical Biology*, vol. 84, p. 10, Nov 2021.

[15] E. F. Arruda, S. S. Das, C. M. Dias, and D. H. Pastore, "Modelling and optimal control of multi strain epidemics, with application to covid-19," *Plos One*, vol. 16, pp. 1–18, 09 2021.

[16] M. B. Alaya, W. B. Aribi, and S. B. Miled, "Mathematical analysis of a delayed seirds epidemics models: Deterministic and stochastic approach," 2022. arXiv:2208.07690.

[17] A. Hagberg, P. J. Swart, and D. A. Schult, "Exploring network structure, dynamics, and function using networkx," Available: https://www.osti.gov/biblio/960616.

[18] K. Ushey, J. Allaire, and Y. Tang, *Reticulate: Interface to 'Python'*, 2023. R package version 1.28. Available: https://CRAN.R-project.org/package=reticulate.

[19] H. W. Borchers, *Pracma: Practical Numerical Math Functions*, 2022. R package version 2.4.2.

[20] M. Rausand, "Introduction," in *Risk Assessment: Theory, Methods, and Applications*, ch. 1, pp. 1–28, John Wiley & Sons Inc., 2013.