POLITECNICO DI TORINO

Master's Degree in Physics of Complex Systems

Master's Degree Thesis

A Time-Varying Network Instance-Based Learning Model of Cyber Threats

Supervisors
Prof. Luca Dall'Asta
Dr. Nicolò Gozzi
Prof. Nicola Perra

Candidate
Gabriele Beltrone

March 2026

Abstract

Cyber threats spread through online social networks by exploiting users' susceptibility, which is shaped by individual characteristics and cognitive mechanisms. Typically, models that simulate the spread of cyber threats assume time-aggregated networks and static susceptibility, which limit the applicability of modeling findings to realistic scenarios. This thesis integrates the temporal dynamics of online social networks with the decision-making processes of individuals by implementing a cognitive model that describes users' phishing susceptibility.

We combine three main components: i) a time-varying network, in particular an Activity Driven Network (ADN) with non-homogeneous susceptibility, expressed in terms of gullibility (i.e. infection propensity) and time to recover, over which cyber threats spread; ii) a Susceptible-Infected-Susceptible (SIS) model; and iii) an Instance-Based Learning Model (IBLM). In an ADN, each node has its own activity, which governs its propensity to create connections. In the IBLM, the decision to interact with potentially phishing content is based on the retrieval of past experiences from memory. We develop a comprehensive model combining these three components. In more detail, we model the messages (e.g. emails) exchanged within the network and define how gullibility is represented within the IBLM. Each node in the network is assigned a memory that represents its past experience with emails. The memory consists of instances containing the features of the message and the associated action, reflecting the node's gullibility through its quality, that is, how accurate the associations between features and actions are. We then model the spread of cyber threats using a SIS model. Contagion is mediated by the IBLM, which generates users' decisions based on the retrieval of past instances most similar to the current message. We also introduce the effect of trust in the sender. The chosen action and the features of the new message are stored in memory and used in future decisions. If a node opens a message containing phishing content, it becomes infected. Unbeknownst to the node, the malware sends phishing emails whenever the infected node contacts another node, propagating through the network.

Our results show three main findings. First, our implementation of the cognitive model reproduces the qualitative behavior reported for humans in the literature. Second, although contagion is mediated by the IBLM rather than by a fixed infection probability, we find that the integrated model still exhibits threshold behavior. We find an approximate mapping to a standard SIS process that recovers the epidemic threshold of the integrated model for large recovery rates. Third, we investigate network effects, showing that vulnerability increases when activity and gullibility are positively correlated. We then propose an extension of the IBLM that adjusts individuals' decision-making processes based on trust in other users, finding that the SIS stationary infected density increases when trust is considered in the model. This result shows that system vulnerability is shaped by the interplay between network structure and cognitive processes, suggesting that susceptibility should be modeled as a dynamic property.

Overall, this thesis proposes a framework that links network science, computational epidemiology, and cognitive science, and provides a basis for future improvements in cyber-threat modeling.

Contents

List of Figures

Chapter 1

Introduction

The spread of cyber threats is one of the most challenging problems facing our digital, interconnected society. Today, many of our activities, services, and communications rely on digital infrastructure, often connected to the internet. As with other human infrastructures and technologies, its vulnerabilities are often exploited by malicious actors for criminal purposes. Cybercrime is responsible for huge financial losses. According to the European Commission, cybercrime costed EUR 5.5 trillion in 2020 [1]; further estimates suggest USD 12.2 trillion by 2031 [2]. To understand the impact of these criminal actions, the International Monetary Fund estimated in a 2018 report, based on cyberattack frequency data from 2011 to 2016, that losses due to cybercrime could amount to approximately 9% of the annual net income of banks [3]. In parallel, spending on cybersecurity has also increased, reaching around USD 208 billion globally in 2024, and it is expected to grow to USD 352 billion by 2030 [4]. Despite the significant amounts spent to tackle cybercrime, the risks associated with digital infrastructures remain high, as indicated by the Internet Security Threat Reports published by Symantec [5].

In this context, our attention turns to social engineering attacks, a class of cyber threats that exploit trust in online relationships, and which are among the most relevant in online social networks [6]. They are characterized by the use of deception to exploit vulnerabilities in human psychology, making them effective against users [7]. In particular, they are able to exploit behavioral and emotional factors typical of human beings, with the aim of provoking specific reactions that allow malicious actors to pursue their criminal goals [7]. Social engineering attacks span a range of different threats. These include, for instance, phishing, where attackers pretend to be a trusted entity; smishing, where phishing is carried out via text messages; file masquerading, where malicious content is presented as legitimate; and sharebaiting, where attackers use social media posts to spread malicious contents [8].

Various research areas have developed over time to contrast this class of cyber threats. The first one focuses on the development of automated threat detection systems, so that cyber threats do not reach users, who are often considered the weak link in the security

chain, even though this assumption has been challenged in more recent years [8, 9]. However, these automated detection systems can only contrast a fraction of total social engineering attacks. Indeed, such threats often exploit legitimate functionalities and leave very few technical and objective characteristics for detection systems to analyze [8].

The second area of research deals with how properties of different types of networks influence the spread of cyber threats. The literature shows how the heterogeneity of online social contacts - the presence of hubs, for instance - in the network makes it more fragile and vulnerable to this type of attack [10, 11]. Topological properties of online interactions therefore play a fundamental role in driving the propagation of such cyber threats [12, 13]. Historically, this line of research has presented two main limitations. First, many studies have ignored the fact that not all users are susceptible to cyber threats in the same way [14]. Second, few studies considering heterogeneous susceptibility often assume it to be a static property, thus disregarding its dynamic nature, which often depends on cognitive processes [15, 16]. One of the first models developed to overcome these limitations was proposed by Brett et al., who addressed the temporal dynamics and heterogeneity of social contacts, as well as heterogeneous user susceptibility [17]. They proposed a model to study the spread of cyber threats on time-varying networks, specifically Activity-Driven Networks (ADNs), as introduced by Perra et al. [18], which realistically capture the heterogeneous and dynamic nature of online social contacts. Furthermore, they account for the non-homogeneous susceptibility of users by introducing Q categories that characterize users' susceptibility in terms of their gullibility and recovery rates; more gullible nodes are more easily infected [17]. Since susceptibility also depends on socio-demographic factors [19], belonging to a given class could influence the link-creation process. Indeed, homophily, the principle that people tend to form ties with others similar to themselves, is a strong social mechanism [17]. The Susceptible-Infected-Susceptible (SIS) model is used to describe the spread of cyber threats in the network. Although the propagation of biological viruses and computer viruses is different, compartmental models are widely used in the literature to study cyber threats [11, 20–24]. The model allows epidemic thresholds to be derived analytically; these are a function of the interplay between social dynamics and contagion dynamics. The research shows how, under certain conditions, this interplay produces a dynamic that increases the vulnerability of the system [17]. Despite developments in this area, susceptibility is still considered a static property of the user.

Finally, the last area of research concerns the characterization of user susceptibility, based on the assumption that it is not possible to ignore the human element from the security chain [25]. The literature shows that several factors contribute to high susceptibility, including obedience to authority [26], submissiveness [27] and certain socio-demographic characteristics [19]. These characteristics are difficult to measure legally and ethically in work contexts [28]. The literature has therefore shown that other easily measurable indicators, such as security training, familiarity with the system and frequency of usage, are good predictors of susceptibility [28–30]. The importance of these indicators has been confirmed by cognitive science methods with the aim of describing decision-making processes [16]. There are models based on Instance-Based Learning Theory (IBLT) [31], called Instance-

Based Learning Models (IBLM), that accurately reproduce user behavior in various cybersecurity tasks, particularly phishing detection [15, 32]. Studies show that susceptibility is influenced by past experiences, cognitive biases (frequency/recency), and limited rationality [15, 16, 32]. According to experimental evidence, key factors in identifying phishing emails are knowledge and experience [30]. However, the typical experimental setup involves testing the individuals alone, ignoring the fact that in reality they are actually connected to other users [33, 34]. Such connections can modulate the probability of being deceived by cyber threats [35, 36]. In fact, in the context of social networks, one factor that can influence susceptibility is trust in other users of the network [36]. Users are much more likely to fall victim to phishing if contacted by someone who seems familiar [37]. Furthermore, the success rate of this type of attack, in cases where the email appears to come from a friend or acquaintance, is not always maximum, but it can succeed in over 70% of cases [37].

This is the context in which our research fits. We define a coherent framework that integrates i) the model proposed by Brett et al. for describing the dynamics and heterogeneity of online social networks, ii) the SIS dynamics evolving on such networks, and iii) the cognitive processes governing users' phishing susceptibility through an IBLM that models decision-making mediated by cognitive mechanisms.

Indeed, despite efforts to improve the realism of current models, important limitations still remain. In models of online social networks, users' susceptibility is typically modeled as a static property, even when heterogeneity across individuals is taken into account. In reality, however, it depends on experience and knowledge, which inherently evolve and adapt over time. Similarly, contagion dynamics are often parameterized using fixed probabilistic parameters, neglecting the fact that contagion outcomes emerge from individual decision-making processes that take place in a dynamic context. Lastly, although cognitive models of phishing susceptibility can reproduce individual decision-making with good accuracy by modeling the cognitive mechanisms involved, they usually treat individuals in isolation, ignoring the fact that exposure to cyber threats and learning take place within a dynamic social network. Our goal is to develop a model that provides a more realistic description of the spread of cyber threats; to this end, these limitations must be overcome.

In our integrated model, we define the messages exchanged over the network (e.g. emails), their generation and comparison, and the translation of the concept of gullibility into the mechanisms of the cognitive model. Each node in the network is then assigned a memory that represents its past experience with online content, in particular with emails. The memory consists of instances containing the features of the message and the associated action, reflecting the node's gullibility through its quality, that is, how accurate the associations between features and actions are. We then model the spread of cyber threats on the temporal network using a SIS model. Infection, however, is now mediated by the IBLM sitting on top of each node. Nodes decide whether to open a new message by solving an optimization problem that takes into account both past experience and the trust in the sender. In other words, each node selects the most suitable action by evaluating the similarity between the new message and those stored in memory, retrieving the actions taken

in the most relevant past situations, and considering the level of trust in the sender. The chosen action, together with the characteristics of the new message, is then stored in memory and used in future decisions. If a node opens a message perceived as safe but actually containing phishing content, it becomes infected. Unbeknownst to the node, the malware then sends compromised emails whenever the infected node contacts another node, until it recovers and returns to the susceptible state. In this way, the cyber threat can propagate through the network. To measure the impact of these additions and changes, we conducted extensive numerical simulations.

What we find is that the proposed integrated model offers a coherent framework for studying cyber threats spread in which susceptibility emerges from decisions based on past experience mediated by cognitive mechanisms. The IBLM qualitatively reproduces human phishing-detection behavior and the resulting spreading dynamics preserve the main qualitative features of SIS dynamics on ADN. In addition, the mapping to the standard SIS model provides a useful tool to interpret epidemic thresholds. Finally, we demonstrate through numerical simulations that network effects, such as trust within communities mediated by the cognitive mechanisms, can influence contagion outcomes. Overall, our framework provides a basis for future developments that could lead to more realistic models of cyber threats spread, for instance by incorporating a learning mechanism.

1.1 Thesis organization

We structured the thesis as follows:

- **Chapter 1**: in this first chapter we introduce the topic of the thesis, providing the context in which our project is developed and presenting the state of the art.
- **Chapter 2**: in this chapter we introduce some basic concepts of graph theory. Next, we introduce a class of time-varying networks, the Activity Driven Network (ADN). In this chapter, we focus on the network generation algorithm and its properties.
- **Chapter 3**: in this chapter we focus on the model that allows us to study the spread of cyber threats. In particular, we present an epidemiological model, widely used in the study of biological virus propagation, which belongs to the class of compartmental models. We begin with a brief introduction to compartmental models and then examine a notable case that corresponds to the model we will actually use. We will then focus on analyzing how cyber threats spread on the ADN and derive the epidemic threshold analytically, first in the absence of heterogeneous susceptibility and then with its inclusion.
- **Chapter 4**: in this chapter we focus on presenting the modeling of individuals' learning and cognitive processes. We begin the chapter with the introduction of Instance-Based Learning Theory (IBLT), whose mechanisms, introduced in the Adaptive Control of Thought-Rational (ACT-R) cognitive model, have shown excellent applicability

to the recognition of phishing emails in the literature. This is followed by an explanation of the Instance-Based Learning Model (IBLM) that has been chosen to implement phishing email detection in the integrated model with the network.

- **Chapter 5**: in this chapter we first define the framework of the integrated model, explaining the solutions adopted to integrate all components in a coherent way supported by the literature. We then discuss the results of the simulations, that can be divided into three parts. The first presents the results on the implementation of the IBLM, describing our implementation of the model and showing that it can reproduce human behavior in a qualitatively accurate way. The second defines the mapping between standard SIS dynamics and the integrated model with the IBLM, and tests its accuracy and limitations. The third introduces a community effect into the cognitive model and analyzes the consequences of this addition.

- **Chapter 6**: in this chapter, we summarize the main findings of the thesis on cyber-threat spreading, discuss the strengths and limitations of the proposed cognitive-network framework, and outline possible future developments.

Chapter 2

Network modeling

In order to model the dynamic interaction processes that occur among different individuals, it is necessary to start from some basic concepts and introduce graphs, or networks, depending on the field of science in which we are operating. In the following, the terms graph and network will be used interchangeably. In the following sections, we will introduce a class of time-varying networks, which is essential for our purposes. In this chapter, we focus exclusively on the network generation mechanism and its properties; the interaction with dynamical processes unfolding on it will be discussed in the next chapter.

2.1 Basic concepts of graph theory

A graph is a mathematical structure used to represent the relationships between objects. These objects are called vertices or nodes, forming the set denoted by V. Each vertex/node is assigned a unique index, so that $V = \{1, 2, 3, \ldots, N\}$, where N is the total number of nodes. The relationships between nodes are represented by connections called edges (or links), which together form the set $E \subseteq V \times V$. Each edge is expressed as a pair of nodes, $e = (i, j)$, where i and j are two distinct nodes.

Graphs can be categorized based on the type of interactions between nodes. Specifically, we can distinguish:

- **Directed** or **undirected**: depending on whether the connections have a specific direction or if they are bidirectional.
- **Weighted**: if each edge (i, j) is assigned a real value w_{ij} representing the strength of the connection.
- **Time-varying**: in this case, the structure (i.e. the set of edges E) and/or weights of the network change over time, which is commonly observed in epidemic spreading scenarios.

A graph can be represented by an adjacency matrix $A \in \mathbb{R}^{N\times N}$, where each entry A_{ij} encodes whether an edge between nodes i and j exists. Formally:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

If the graph is undirected, both (i,j) and (j,i) belong to E; the adjacency matrix is therefore symmetric, $A_{ij} = A_{ji}$.

The **degree** of a node corresponds to the number of connections it has. More formally, the **neighborhood** of a node i is the set of nodes that are directly connected to it. This set is denoted by ∂i and is formally defined as:

$$\partial i = \{j \in V : \ (i,j) \in E\}$$

and the degree of node i is $k_i = |\partial i|$. In directed graphs, edges are ordered pairs (i.e. $(i,j) \neq (j,i)$), and it is possible to distinguish between in-degree, k_i^{in}, and out-degree, k_i^{out}; one notes that:

$$k_i^{\text{in}} + k_i^{\text{out}} = k_i \tag{2.1}$$

It is useful to introduce the concept of **path**. A path from node i to node j is defined as a finite sequence of distinct nodes

$$(v_0, v_1, \ldots, v_\ell)$$

such that $v_0 = i$, $v_\ell = j$, and $(v_{m-1}, v_m) \in E$ for all $m = 1, \ldots, \ell$. In directed graphs, a path must respect the direction of the edges.

A graph is said to be **connected** if there exists a path between any pair of nodes. More generally, a graph can be divided into **connected components**, which are maximal subsets of nodes such that any pair of nodes in the same subset is connected by a path.

In this work, we focus on **directed**, **unweighted** and **time-varying** graphs. We furthermore assume that self-loops are not allowed.

2.2 Activity Driven Network (ADN)

As anticipated in Chapter 1, network dynamics play a crucial role in the spread of cyber threats. It is therefore essential to provide a realistic description of how networks evolve over time, taking into account the fact that connections between individuals are changing rapidly and are characterized by processes with a very short timescale [12, 38]. Most models are connectivity driven, meaning that the structural organization of the network provided

the basis for the algorithm generating the network itself. It is important to note that connectivity driven networks, where network topology is central, are effective in capturing the essential characteristics of systems where connections persist over time [39, 40]. It should be noted that connectivity driven networks produce a time-aggregate representation, which can fail to describe the dynamic evolution of many networks. Connectivity driven networks are not effective in replicating the dynamics of connection and disconnection between agents. This is especially true in our case where the timescales of the virus spread are often comparable to those governing network evolution. Indeed, choosing a time-aggregated representation implicitly assumes that the dynamics occurring on them have much shorter characteristic times than those of the network.

Another important factor is the heterogeneity in individuals' propensity to engage in interactions [17, 41, 42]. In real-world scenarios, individuals do not all interact with the same frequency; some are more active than others. In order to account for this heterogeneity, we adopt the activity-driven model introduced by Perra et al. [18]. This model, which we refer to as the Activity Driven Network (ADN), provides a suitable framework for time-varying networks, precisely because it captures the temporal nature of interactions in highly dynamic systems while accounting for heterogeneous activity patterns among individuals.

2.2.1 A data-driven model

The authors begin by analyzing three real-world datasets of networks with time-resolved interactions: collaborations in the journal "Physical Review Letters " (PRL) published by the American Physical Society [43], messages exchanged over the Twitter microblogging network and the activity of actors in movies and TV series as recorded in the Internet Movie Database (IMDb) [44].

What the authors observe, when analyzing the datasets, is that the activity of individual agents in the network - for example, the number of collaborations in the case of the PRL dataset - depends on the time window considered. The choice of the time window clearly affects both the network topology and the degree distribution. The smaller the window, the more the network consists of disconnected components. As the length of the window increases, the link density of the network rises and heterogeneous connectivity patterns emerge.

The authors' aim is to define a model (a data-driven model) to generate a network capable of reproducing empirical evidence, starting from the analysis of datasets. This model should primarily be able to reproduce a heterogeneous distribution of human activity.

2.2.2 The activity potential

The authors analyze three real-world datasets of networks with time-resolved interactions and, for each agent, evaluate an empirical quantity - called the activity potential - that

characterizes its level of interaction with other agents. The **activity potential** is formally defined as the number of interactions performed, in a given time window, by each node divided by the total number of interactions made by all the nodes in the same time window. Focusing our attention on node i, we have that its activity potential is:

$$x_i = \frac{m_i}{\sum_j m_j} \tag{2.2}$$

with m_i the number of interactions in a time interval Δt performed by node i with any of the other $N-1$ nodes in the network and $\sum_j m_j$ the total number of interactions within the same time interval; in the end we have a set of activity potentials $\{x_i\}$. We understand that:

$$\begin{aligned} &1. \quad 0 \leq x_i \leq 1 \quad \forall i \\ &2. \quad \sum_i x_i = \frac{\sum_i m_i}{\sum_j m_j} = 1 \end{aligned}$$

meaning that the activity potential x_i may be interpreted as the probability of node i to be active, where active means that the node is involved in some kind of interaction. The set of activity potentials $\{x_i\}$ can also be used to estimate the probability distribution function $F(x)$, which gives the probability that a randomly selected node i has activity potential x. It is observed empirically that $F(x) \propto x^{-\gamma}$ (see Figure 2.1). To prevent the divergence of the probability distribution, the possible values that x can assume are constrained such that $x \in [\varepsilon,\ 1]$, with $0 < \varepsilon \ll 1$. Imposing the normalization:

$$\int_\varepsilon^1 F(x)\mathrm{d}x = \int_\varepsilon^1 K x^{-\gamma}\mathrm{d}x = 1 \quad \Longleftrightarrow \quad \int_\varepsilon^1 x^{-\gamma}\mathrm{d}x = \frac{1}{K}$$

Then:

$$\int_\varepsilon^1 x^{-\gamma} \overset{\gamma \neq 1}{=} \left[\frac{x^{1-\gamma}}{1-\gamma}\right]_\varepsilon^1 = \left[\frac{1}{1-\gamma} - \frac{\varepsilon^{1-\gamma}}{1-\gamma}\right] = \frac{1-\varepsilon^{1-\gamma}}{1-\gamma}$$

So, given that $\gamma \neq 1$, the probability distribution $F(x)$ reads:

$$F(x) = K x^{-\gamma} \quad \text{with } K = \frac{1-\gamma}{1-\varepsilon^{1-\gamma}} \tag{2.3}$$

2.2.3 Activity driven network model

The following section provides a detailed description of the model proposed by Perra et al. [18]. The authors consider N nodes (the individuals) and assign to each node $i \in [1, \ldots, N]$ an **activity** defined as:

$$a_i = \eta x_i \quad \text{with } x_i \sim F(x) \tag{2.4}$$

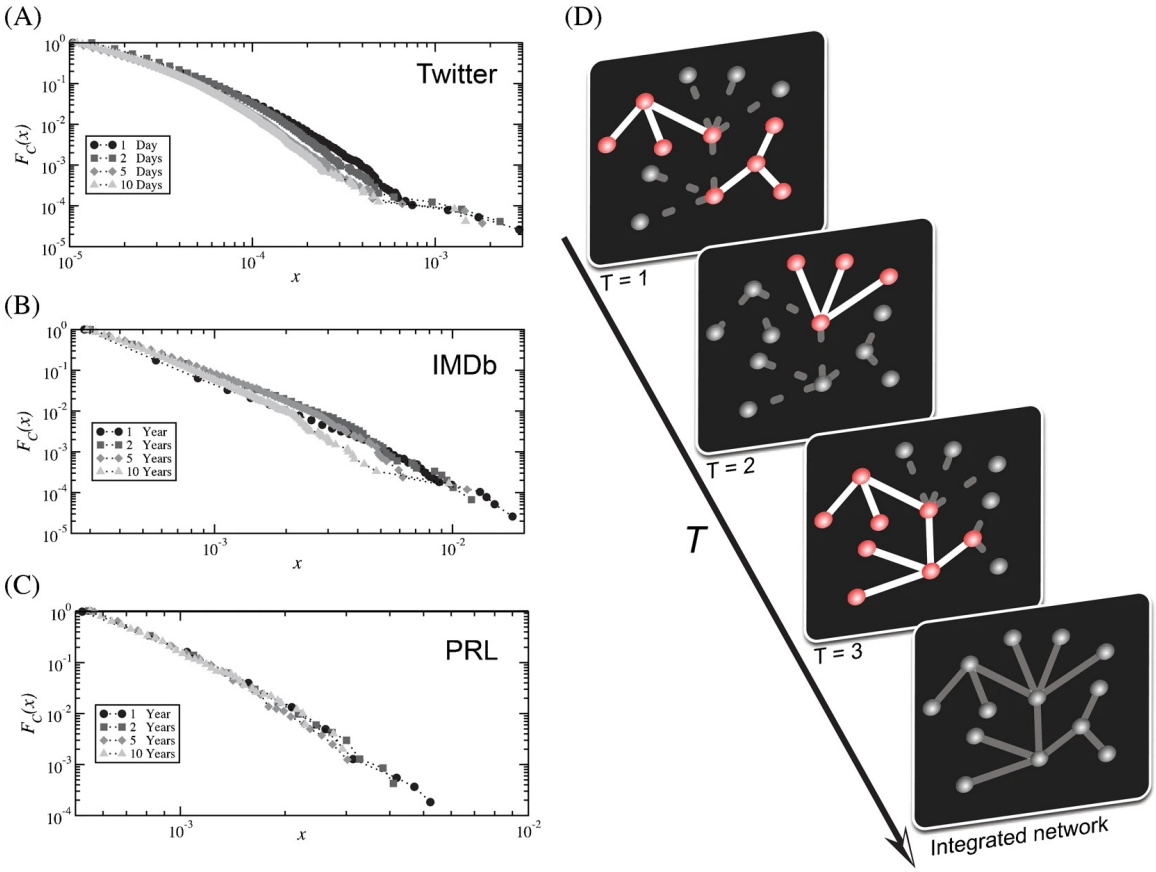


Figure 2.1 Cumulative distribution of activity potential ($F_C(x)$) measured empirically across four time windows and shown schematically. Panel A: Twitter; Panel B: IMDb; Panel C: PRL. Panel D: model schematic. The network is plotted here for 3 different time steps, with red nodes showing active/firing ones. Final visualization shows the network integrated over all time steps. Reproduced from [18].

defined as the probability per unit time to create new contacts or interactions with other individuals. We observe that η is a rescaling factor defined such that the average number of active nodes per unit time is $\eta \langle x \rangle N$. In order to fulfill the definition of probability we must set $\eta = 1$, indeed:

$$1. \quad 0 \leq a_i \leq 1 \quad \Longleftrightarrow \quad 0 \leq x_i \leq \frac{1}{\eta}$$

$$2. \quad \sum_i a_i = \eta \sum_i x_i = \eta$$

Therefore $a_i \sim F(a)$.

Let us now proceed to understand the rules governing the network's generative process. The rules are the following:

- At the beginning of each time step t the network G_t consists of N disconnected vertices;
- Each node i becomes active with probability a_i and connects to m other nodes selected at random. Non-active nodes can still receive connections from active nodes;
- Then, at the next time step $t + \Delta t$, all the connections in the network G_t are deleted. This definition implies that all interactions have a constant duration Δt.

The model described is random and Markovian, as agents do not retain memory of past time steps. Consequently, the entire network dynamics and the resulting structural properties are fully determined by the activity potential distribution $F(x)$.

At each time step the network is a simple random graph G_t with low average connectivity (sparse graph); the graph is made up of many small disconnected components. As activity is aggregated over progressively larger time windows T, the accumulation of connections gives rise to a skewed (asymmetric) degree distribution with broad variability. The emergence of heterogeneities and hubs - nodes with a large number of connections - stems from the wide dispersion of activity rates in the system and the presence of highly active agents. Furthermore, the model provides a good reproduction of the empirical observations from the dataset, despite its simplifications.

The model allows for an analytical treatment. The integrated network is defined as the union of all the networks obtained in each previous time step, that is:

$$G_T = \bigcup_{t=0}^{t=T} G_t \tag{2.5}$$

The instantaneous network generated at time t will consist of a set of slightly interconnected nodes corresponding to active agents at that time, plus agents that received connections from them. Each active node will create m links and the average total number of edges per unit time is:

$$\langle E \rangle = Nm\eta \langle x \rangle = Nm \langle a \rangle \tag{2.6}$$

This defines the network's average degree per unit time, or contact rate:

$$\langle k \rangle_{\mathrm{u}} = 2\frac{\langle E \rangle}{N} = 2m \langle a \rangle \tag{2.7}$$

for an undirected graph, and

$$\langle k \rangle_{\mathrm{d}} = \frac{\langle E \rangle}{N} = m \langle a \rangle \tag{2.8}$$

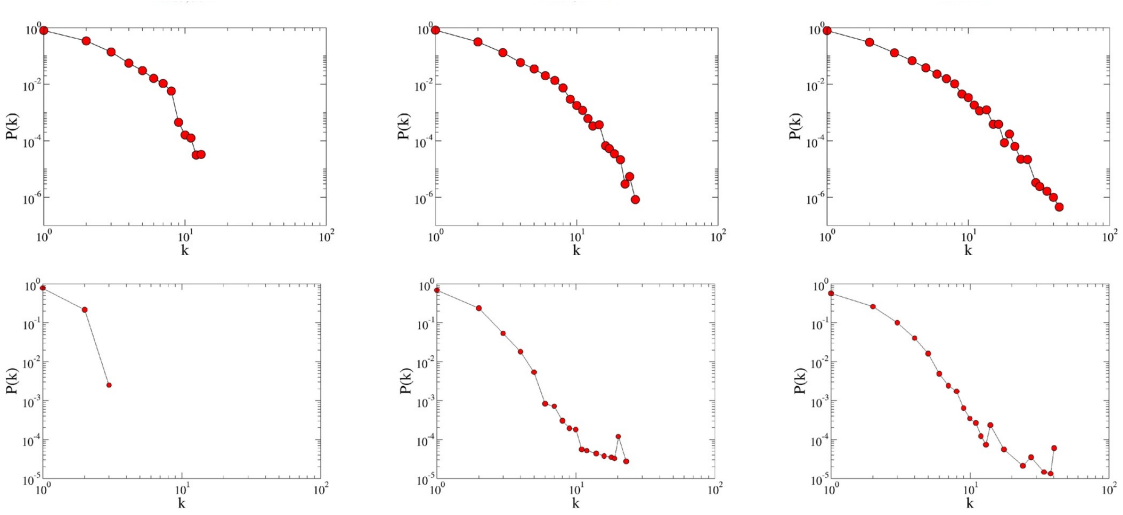


Figure 2.2 The first row shows the degree distribution of the PRL dataset for three different time scales (1, 10 and 30 years). As the degree distributions show, the topological structure of the network is affected by the time scale used to construct it. The second row is the degree distribution obtained with ADN for three different time scales (1, 10 and 20 time steps). Interestingly, despite the fact that the model is random and Markovian by design, we observe behaviour that is qualitatively similar to that observed in the PRL case. Reproduced from [18].

for a directed graph.

Star-like structures form around active nodes in the instantaneous network at each time step. In contrast, the integrated network, which is the union of past time steps, is generally dense (see Figure 2.1). For large T and N, the degree of agent i in the integrated network is shown to be equal to:

$$k_i(T) = N\left(1 - e^{-\frac{Tm\eta x_i}{N}}\right) \tag{2.9}$$

from which the degree distribution $P_T(k)$ can be derived:

$$P_T(k) \sim F\left[\frac{k}{Tm\eta}\right] \tag{2.10}$$

The key result is that the integrated network's degree distribution mirrors the distribution of individual activity. The activity potential distribution function $F(x)$ shapes the network structure over time, linking the emergence of hubs to the heterogeneous activity of its agents. This correspondence is roughly observed in empirical data (see Figure 2.2), though discrepancies arise from real-world factors not captured by the model, such as link memory,

relation persistence, and weighted or multiple connections.

To summarize, the ADN parameters are:

- N, the total number of nodes,
- γ, the exponent of probability distribution $F(x)$,
- m, the number of random connections that an active node creates,
- ε, lower bound introduced in order to avoid divergence of $F(x)$,
- η, the rescaling factor.

2.2.4 ADN with susceptibility classes and homophily

In this section, we will analyze the extension proposed by Brett et al. [17] to the ADN model proposed by Perra et al. [18]. In order to account for the heterogeneity of users' susceptibility, the authors introduce susceptibility classes into the model. Each node is assigned to one of Q classes describing susceptibility to cyber threats, measured in terms of *gullibility* (propensity to fall for deception) and the time needed to recover from a successful attack. The authors then hypothesize that susceptibility classes may influence the link formation process. Since susceptibility is often related to socio-demographic characteristics, the authors hypothesize that belonging to a specific class may influence contact creation mechanisms. This relevant social mechanism is called homophily and is known to influence the structure and organization of social relationships.

As mentioned, to model contact between agents, the authors use a generalization of the ADN framework. They therefore propose an extension of the model that considers homophily, modifying the link creation process:

- with probability p, each node randomly selects a target within the group of nodes belonging to the same category;
- with probability $1 - p$, on the other hand, the target is chosen from nodes belonging to different categories.

The parameter p regulates the network's susceptibility to cyber threats in relation to its level of homophily.

2.2.5 ADN with communities

In the following section, we analyze the addition of modularity, i.e., the presence of communities, in time-varying networks, as proposed by Nadini et al. [45]. To this end, the authors introduce a generative model of a temporal network with adjustable modularity capable of reproducing various characteristics observed in real temporal graphs, starting from the

ADN model proposed by Perra et al. [18]. In this model, each node belongs to a single community. In line with empirical evidence, the size s of the communities is extracted from a heavy-tailed distribution: $P(s) = Cs^{-\omega}$, $s \in [s_{\min}, N]$, where C is a normalization constant. In this way, they do not fix the number of modules a priori, but let it emerge from the model parameters and the distribution of community sizes.

Given this framework, the generative model of the network is defined as follows:

- At each discrete time step t, the network initially consists of N disconnected nodes;
- With probability $a_i \Delta t$, each node i becomes active and can create m connections. Each link generated by i is directed:
  - with probability p_c towards a node in the same community as i;
  - with probability $1 - p_c$ towards a node belonging to a different community.

  In both cases, the target node is selected uniformly at random within the chosen community;
- At the next step $t + \Delta t$, all connections in G_t are removed and the process repeats.

In this scheme, the parameter p_c directly controls the degree of modularity of the contacts over time, allowing us to explore the influence of community structure on the dynamic processes of the network (see Figure 2.3). In general, adding community structure makes the system more fragile when individuals can be reinfected, whereas it slows the spread when they acquire immunity after infection.

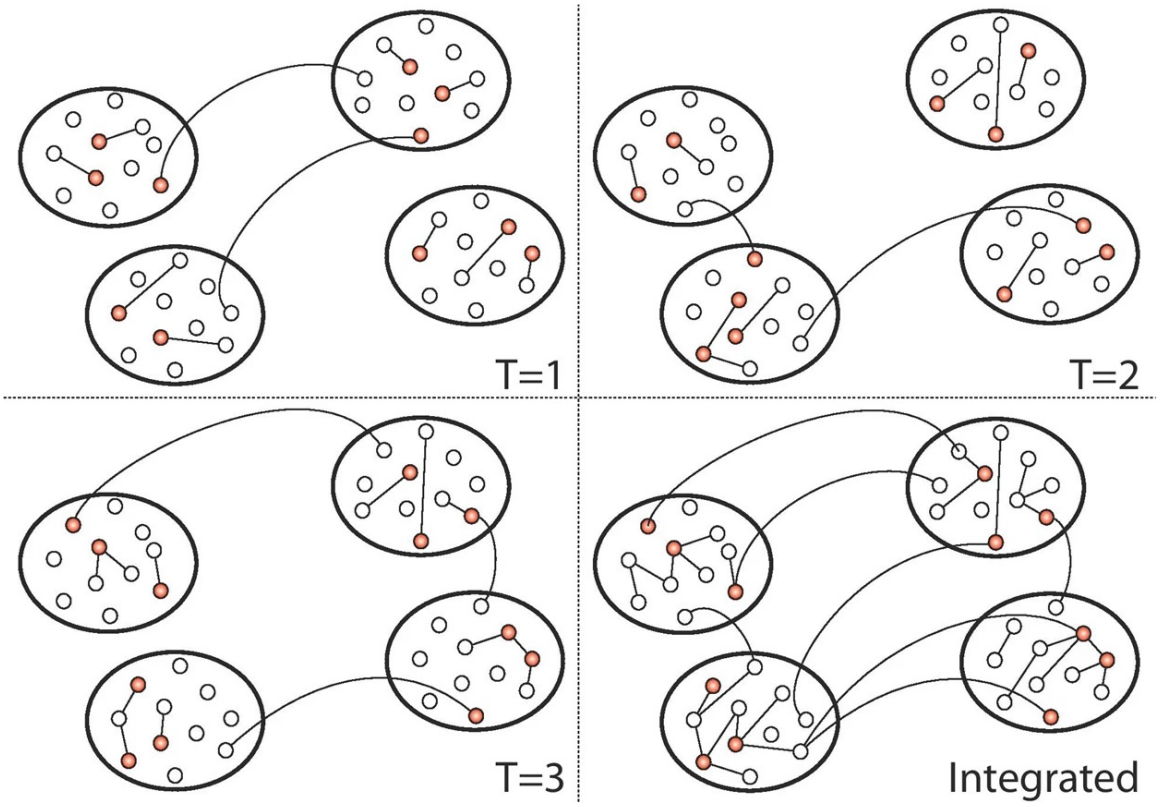


Figure 2.3 This is a schematic representation of the model. Active nodes are shown in red. Straight lines show connections within the same community and links show connections between different communities. The bottom right panel shows the integrated network obtained by combining G_1, G_2 and G_3. Reproduced from [45].

Chapter 3

Cyber threats spreading modeling

In the previous chapter, we described the time-varying network model, outlining its evolution and main characteristics. In this chapter, we focus on the model that allows us to study the spread of cyber threats. In particular, we employ an epidemiological model, widely used in the study of biological virus propagation, which belongs to the class of compartmental models. We begin with a brief introduction to compartmental models and then examine a notable case, that corresponds to the model we will use later. We will then focus on analyzing how cyber threats spread on an ADN and derive the epidemic threshold analytically, first in the absence of susceptibility classes and then with their inclusion.

3.1 Compartmental models for contagion processes

Early compartmental-type models appear in Hamer (in 1906) [46] and in Ross' work on malaria pathometry (formalized in 1916) [47], while Kermack and McKendrick (1927) provided a fundamental formulation of epidemic dynamics [48].

Compartments are used to categorize individuals within a population according to the various stages of infection they may experience during an epidemic. One of the simplest compartmental models is structured as follows:

- **Susceptible (S)**: the compartment contains healthy individuals who can potentially contract the infection;
- **Infected (I)**: the compartment contains infected individuals who are capable of transmitting the disease to susceptible individuals;
- **Recovered (R)**: the compartment contains individuals who have recovered from the disease and are no longer susceptible to infection.

The structure of a compartmental model typically determines its name. The model just described is commonly called the SIR model, derived from the initial letters of its compartments (Susceptible–Infected–Recovered). Additional compartments can also be included,

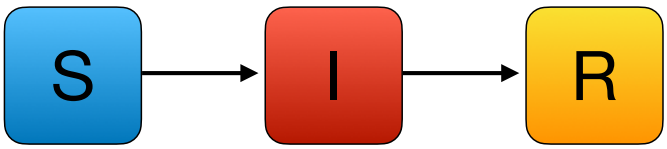


Figure 3.1 Schematic representation of SIR model; the arrows show the direction of the allowed transitions.

for instance for deceased individuals (D), vaccinated individuals (V), or those in a latent/exposed state (E). Furthermore, it is possible to define separate compartments for two or more competing diseases. A wide range of models can be created by combining different compartments and adjusting the transition rates. These models can represent diverse populations and pathogens.

The total population is generally a function of time, $N(t)$, and the sum of the populations in all compartments must satisfy the constraint of being equal to the total population. The total population is constant $N(t) = N$ if birth and deaths are not included in the model; we will consider N as a constant. The number of individuals inside of a compartment is time dependent and is labeled with capital letters $(S(t), I(t), R(t))$.

Individuals move between compartments according to transition rates and interactions. In compartmental models, transitions are generally classified into two basic types:

- **Spontaneous transitions**: an individual spontaneously moves from one compartment to another. An example is the recovering process of an infected $(I \to R)$;
- **Transitions involving interactions**: an individual changes compartment as a result of interaction with another individual. This occurs when susceptible individuals, upon interacting with infected individuals, become infectious themselves, as described in the SIR model $(S + I \to 2I)$.

Transitions of the first type are governed exclusively by the transition rate (i.e. the recovery rate (μ)). The second type of processes depends both on a rate (λ), and on the topology of contacts between individuals of different compartments. The most straightforward method for calculating the contacts between compartments is the homogeneous mixing approximation, in which the probability of interaction is directly proportional to the product of the two individual densities involved in the process. The constant that characterizes the proportionality is the average degree of the network, $\langle k \rangle$, that is as a measure of the average number of interactions an individual experiences with others over a specified time period.

The movement of individuals between compartments can be visualized as a flow. The SIR model is a useful example. The arrows in Figure 3.1 show the possible directions of

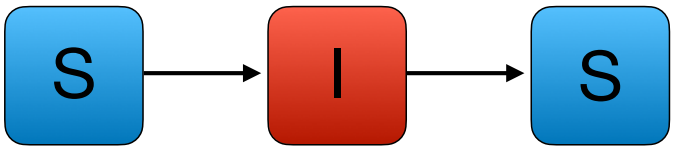


Figure 3.2 Schematic representation of SIS model; the arrows show the direction of the allowed transitions.

transitions; they can be visualized as flows between the various compartments. Focusing on the infected compartment (I), we observe an inflow and an outflow:

- The inflow is characterized by the probability of a susceptible individual being infected by an infected individual, denoted by λ, and average degree $\langle k \rangle$ per unit time; in the literature, the product of λ and $\langle k \rangle$ is typically referred to as β;
- The outflow is characterized by the recovery rate, denoted by μ.

We can therefore derive the discrete equation governing the evolution of the total population in compartment I:

$$I(t + \Delta t) = I(t) + \lambda \langle k \rangle \Delta t S(t) I(t) - \mu \Delta t I(t)$$

In this way, it is possible to write equations for all compartments of the model.

In the next section, our focus will be on the Susceptible-Infected-Susceptible (SIS) model and the derivation of the equations that govern the system's evolution.

3.1.1 SIS model

We introduce now the Susceptible-Infected-Susceptible (SIS) model (see Figure 3.2). Unlike biological viruses, cyber threats do not provide immunity after infection. This makes the SIS model the most suitable for describing the process of "cyber" contagion. In the classical formulation of compartmental models, the population is assumed to undergo homogeneous mixing. Under this assumption, individuals interact uniformly and randomly: each individ-

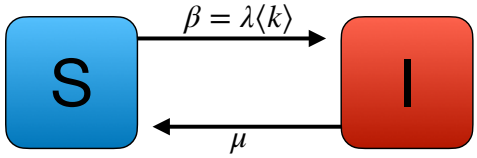


Figure 3.3 Schematic representation of SIS model flow.

ual has the same probability of coming into contact with any other individual.

Within a mean-field approximation and looking at Figure 3.3, the discrete dynamical equations can be written as follows:

$$S^{t+\Delta t} = S^t + \mu \Delta t I^t - \lambda \langle k \rangle \Delta t \frac{S^t}{N} I^t \tag{3.1}$$

$$I^{t+\Delta t} = I^t + \lambda \langle k \rangle \Delta t \frac{S^t}{N} I^t - \mu \Delta t I^t \tag{3.2}$$

Performing the continuous limit ($\Delta t \to 0$), one obtains:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \mu I(t) - \lambda \langle k \rangle \frac{S(t)}{N} I(t) \tag{3.3}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \lambda \langle k \rangle \frac{S(t)}{N} I(t) - \mu I(t) \tag{3.4}$$

One possible approach to solve the system of equations is to use the early-stage approximation, which consists of studying the system at the beginning of the epidemic. This means that the number of infected individuals (I) is negligible (at $t = 0$, the infected population is a minimal fraction of the total population) and that $S \cong N$. Under this approximation, the ODE system becomes:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} \cong \mu I(t) - \lambda \langle k \rangle I(t) = (\mu - \lambda \langle k \rangle) I(t) = K \cdot I(t) \tag{3.5}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} \cong \lambda \langle k \rangle I(t) - \mu I(t) = -(\mu - \lambda \langle k \rangle) I(t) = -K \cdot I(t) \tag{3.6}$$

with $K = \mu - \lambda \langle k \rangle$. In matrix form, we have that:

$$\begin{pmatrix} \dot{S} \\ \dot{I} \end{pmatrix} = \begin{pmatrix} 0 & K \\ 0 & -K \end{pmatrix} \cdot \begin{pmatrix} S \\ I \end{pmatrix} \tag{3.7}$$

Since the solution for $I(t)$ can grow and actually generate an outbreak only if the sign of the exponential is greater than zero, we impose this condition on the largest eigenvalue of the Jacobian. Therefore, we find that the condition for an epidemic not to die out is:

$$\boxed{R_0 \doteq \frac{\lambda \langle k \rangle}{\mu} > 1} \tag{3.8}$$

where we have defined R_0, a fundamental quantity called the reproductive number of an infection. The reproductive number indicates the average number of secondary cases caused by a single primary case within a fully susceptible population. If R_0 is greater than 1, then the epidemic sustains itself and spreads; otherwise, it dies out.

3.2 SIS model on ADN

We have introduced the SIS model into the homogeneous mixing assumption. We will now study it in the case of ADN, presenting the analytical solution proposed by Perra et al. [18].

With an ADN of N nodes, each node is associated with a specific activity $a_i \sim F(a)$, with $F(a) \propto a^{-\gamma}$. Working with the set of activities $\{a_i\}$, it is possible to derive epidemic evolution equations that couple the spread process with network dynamics.

At the mean-field level, the epidemic process is characterized by the number of infected individuals in the activity class (a) at time (t), denoted by I_a^t. The number of infected individuals in class (a) at time $(t + \Delta t)$ is given by:

$$I_a^{t+\Delta t} = I_a^t - \mu \Delta t I_a^t + \lambda m \frac{S_a^t}{N} \int \mathrm{d}a' a' \Delta t I_{a'}^t \tag{3.9}$$

where:

- $\lambda m \frac{S_a^t}{N} \int \mathrm{d}a' a' \Delta t I_{a'}^t$ is the inflow term, proportional to the expected number of infected nodes that are active during Δt, $\int \mathrm{d}a' a' \Delta t I_{a'}^t$, the probability of contacting a susceptible nodes in class a, S_a^t/N, the number m of links created by each active node and the transmission probability λ per contact;
- $\mu \Delta t I_a^t$ is the outflow term, proportional to the number of infected nodes and the recovery rate.

Multiplying by $\int \mathrm{d}a$ the latter equation, we get:

$$\int \mathrm{d}a \; I_a^{t+\Delta t} = \int \mathrm{d}a \; I_a^t - \int \mathrm{d}a \; \mu \Delta t I_a^t + \int \mathrm{d}a \; \lambda m \frac{S_a^t}{N} \int \mathrm{d}a' a' \Delta t I_{a'}^t$$

Then:

$$\begin{aligned} I^{t+\Delta t} &= I^t - \mu \Delta t I^t + \frac{\lambda m \Delta t}{N} \int \mathrm{d}a \; S_a^t \int \mathrm{d}a' a' \; I_{a'}^t = \\ &= I^t - \mu \Delta t I^t + \frac{\lambda m \Delta t}{N} \int \mathrm{d}a \; \left(N_a - I_a^t\right) \int \mathrm{d}a' a' \; I_{a'}^t = \\ &\cong I^t - \mu \Delta t I^t + \frac{\lambda m \Delta t}{N} \int \mathrm{d}a \; N_a \int \mathrm{d}a' a' \; I_{a'}^t \end{aligned}$$

where we have used the early-stage approximation (neglecting second-order terms in the infected population I^t). Then:

$$I^{t+\Delta t} \cong I^t - \mu \Delta t I^t + \lambda m \Delta t \theta^t$$

where we have defined $\theta^t \doteq \int \mathrm{d}a \; a I_a^t$. Performing the continuous limit ($\Delta t \to 0$):

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = -\mu I(t) + \lambda m \theta(t) \tag{3.10}$$

The next step is to find the dynamical equation for $\theta(t)$. Multiplying by $\int \mathrm{d}a\ a$ equation (3.9), we get:

$$\begin{aligned}
\theta^{t+\Delta t} = \int \mathrm{d}a\ a I_a^{t+\Delta t} &= \int \mathrm{d}a\ a I_a^t - \int \mathrm{d}a\ a\mu\Delta t I_a^t + \int \mathrm{d}a\ a\lambda m \frac{S_a^t}{N} \int \mathrm{d}a' a' \Delta t I_{a'}^t = \\
&= \theta^t - \mu\Delta t\theta^t + \frac{\lambda m \Delta t}{N} \int \mathrm{d}a\ a S_a^t \int \mathrm{d}a' a' I_{a'}^t = \\
&= \theta^t - \mu\Delta t\theta^t + \frac{\lambda m \Delta t}{N} \int \mathrm{d}a\ a \left(N_a - I_a^t\right) \int \mathrm{d}a' a' I_{a'}^t = \\
&\cong \theta^t - \mu\Delta t\theta^t + \frac{\lambda m \Delta t}{N} \int \mathrm{d}a\ a N_a \int \mathrm{d}a' a' I_{a'}^t = \\
&= \theta^t - \mu\Delta t\theta^t + \lambda m \Delta t \langle a \rangle \theta^t
\end{aligned}$$

where we have used again the early-stage approximation. Performing the continuous limit ($\Delta t \to 0$):

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\mu\theta(t) + \lambda m \langle a \rangle \theta(t) \tag{3.11}$$

This system of two equations can be rewritten in matrix form:

$$\begin{pmatrix} \dot{I} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} -\mu & \lambda m \\ 0 & -\mu + \lambda m \langle a \rangle \end{pmatrix} \cdot \begin{pmatrix} I \\ \theta \end{pmatrix} \tag{3.12}$$

Computing the characteristic polynomial, we get:

$$\begin{aligned}
&(-\mu - \Lambda)\left(-\mu + \lambda m \langle a \rangle - \Lambda\right) = 0 \\
&\mu^2 - \lambda m \langle a \rangle \mu + \mu\Lambda + \mu\Lambda - \lambda m \langle a \rangle \Lambda + \Lambda^2 = 0 \\
&\Lambda^2 + \left(2\mu - \lambda m \langle a \rangle\right)\Lambda + \left(\mu^2 - \lambda m \langle a \rangle \mu\right) = 0 \\
&\Lambda^2 + b\Lambda + c = 0
\end{aligned}$$

with $b = 2\mu - \lambda m \langle a \rangle$ and $c = \mu^2 - \lambda m \langle a \rangle \mu$. The eigenvalues are the solution of the characteristic polynomial:

$$\begin{aligned}
\Lambda_{1,2} &= \frac{-b \pm \sqrt{b^2 - 4c}}{2} = \\
&= \frac{-2\mu + \lambda m \langle a \rangle \pm \sqrt{\left(2\mu - \lambda m \langle a \rangle\right)^2 - 4\left(\mu^2 - \lambda m \langle a \rangle \mu\right)}}{2} = \\
&= \frac{-2\mu + \lambda m \langle a \rangle \pm \sqrt{4\mu^2 + \lambda^2 m^2 \langle a \rangle^2 - 4\lambda m \langle a \rangle \mu - 4\mu^2 + 4\lambda m \langle a \rangle \mu}}{2} = \\
&= \frac{-2\mu + \lambda m \langle a \rangle \pm \lambda m \langle a \rangle}{2} = \\
&= -\mu + \frac{\lambda m \langle a \rangle}{2} \pm \frac{\lambda m \langle a \rangle}{2}
\end{aligned}$$

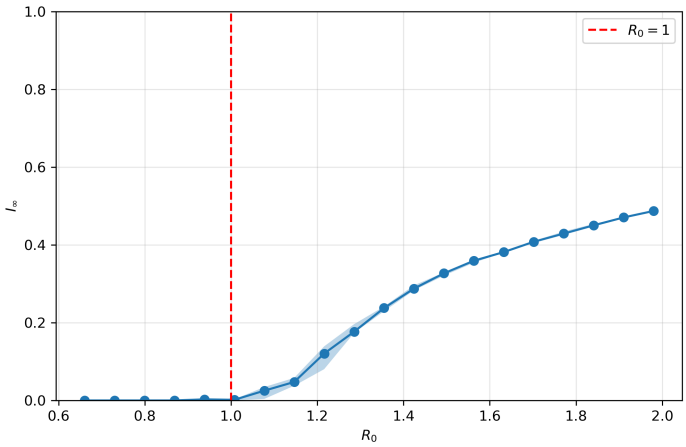


Figure 3.4 In this plot, we show the SIS stationary infected density as a function of R_0. The plot was obtained with $N = 10^5$ and $\mu = 1/40$. We note that the curve starts to increase around $R_0 = 1$.

For exactly the same reasons explained in the previous section, we require the largest eigenvalue $-\mu + \lambda m \langle a \rangle$ of the Jacobian to be greater than zero, obtaining:

$$-\mu + \lambda m \langle a \rangle > 0 \quad \Rightarrow \quad \frac{\lambda}{\mu} > \frac{1}{m \langle a \rangle}$$

Therefore, for a directed graph, one finds that the reproductive number is:

$$\boxed{R_0 \doteq \frac{\lambda}{\mu} m \langle a \rangle > 1}$$

To summarize, the parameters of SIS model are:

- λ = probability of a node to be infected by an infected node,
- μ = rate of recovery (of an infected node).

The authors of the article observe that the results of SIS simulations on time-aggregated networks are misleading in terms of both the threshold and the magnitude of the stationary state, compared to simulations on networks generated using the ADN (see Figure 3.5). Time-aggregated networks assume that links are always available to transmit the infection, ignoring the fact that in reality they may or may not be active depending on the particular activity of the agents.

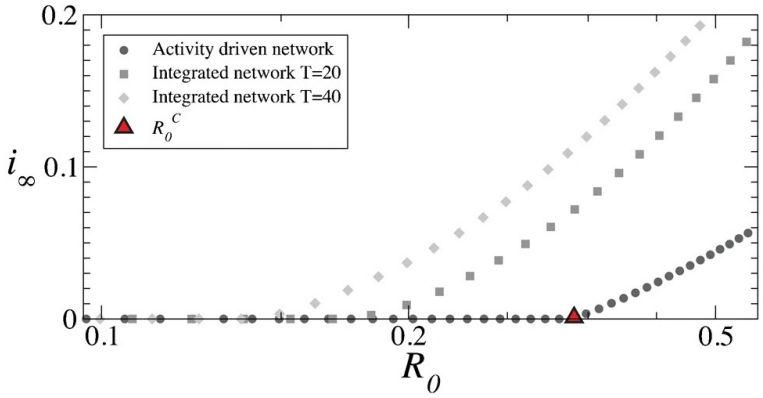


Figure 3.5 The plot shows the value of the fraction of infected individuals in the steady state (i_∞) as the epidemic threshold varies (R_0). Each curve was obtained from numerical simulations of the SIS model on a network generated using the activity driven model and two other networks. The latter two were obtained by integrating the model over 20 and 40 time steps. Reproduced from [18].

3.3 SIS model on ADN with susceptibility classes

When susceptibility classes and the phenomenon of homophily are taken into account, in addition to distinguishing infected nodes based on their activity class (a) at time (t), it is also necessary to consider their membership to a susceptibility class (x), with $x \in \{1, \ldots, Q\}$. Each susceptibility class is characterized by its own *gullibility*, λ_x, and recovery rate, μ_x. We will now present the analytical solution proposed by Brett et al. [17].

The discrete dynamical equation for $I^t_{a,x}$ reads:

$$\begin{aligned} I^{t+\Delta t}_{a,x} = I^t_{a,x} &- \mu_x \Delta t \, I^t_{a,x} \\ &+ \lambda_x m \frac{S^t_{a,x}}{N_x} p \int \mathrm{d}a' \, a' \, \Delta t \, I^t_{a',x} \\ &+ \lambda_x m (1-p) \sum_{y \neq x} \frac{S^t_{a,x}}{N - N_y} \int \mathrm{d}a' \, a' \, \Delta t \, I^t_{a',y}. \end{aligned} \tag{3.13}$$

where the third term on the right-hand side represents the inflow from the same gullibility class and the fourth term represents the inflow from all other classes. Multiplying by $\int \mathrm{d}a$:

$$
\begin{aligned}
I_x^{t+\Delta t} = \int \mathrm{d}a\ I_{a,x}^{t+\Delta t} = \int \mathrm{d}a\ I_{a,x}^t - \mu_x \Delta t \int \mathrm{d}a\ I_{a,x}^t + \\
+ \frac{\lambda_x m p \Delta t}{N_x} \int \mathrm{d}a\ S_{a,x}^t \int \mathrm{d}a' a' I_{a',x}^t + \\
+ \lambda_x m (1-p) \Delta t \int \mathrm{d}a\ S_{a,x}^t \sum_{y \neq x} \int \mathrm{d}a' a' \frac{I_{a',y}^t}{N - N_y} = \\
= I_x^t - \mu_x \Delta t I_x^t + \\
+ \frac{\lambda_x m p \Delta t}{N_x} \int \mathrm{d}a\ \left(N_{a,x} - I_{a,x}^t\right) \int \mathrm{d}a' a' I_{a',x}^t + \\
+ \lambda_x m (1-p) \Delta t \int \mathrm{d}a\ \left(N_{a,x} - I_{a,x}^t\right) \sum_{y \neq x} \int \mathrm{d}a' a' \frac{I_{a',y}^t}{N - N_y} = \\
\cong I_x^t - \mu_x \Delta t I_x^t + \\
+ \frac{\lambda_x m p \Delta t}{N_x} \int \mathrm{d}a\ N_{a,x} \int \mathrm{d}a' a' I_{a',x}^t + \\
+ \lambda_x m (1-p) \Delta t \int \mathrm{d}a\ N_{a,x} \sum_{y \neq x} \int \mathrm{d}a' a' \frac{I_{a',y}^t}{N - N_y} = \\
= I_x^t - \mu_x \Delta t I_x^t + \\
+ \lambda_x m p \Delta t \theta_x^t + \\
+ \lambda_x m (1-p) \Delta t \sum_{y \neq x} \frac{N_x}{N - N_y} \theta_y^t
\end{aligned}
$$

where we have used the early-stage approximation. Performing the continuous limit ($\Delta t \to 0$):

$$
\frac{\mathrm{d}I_x(t)}{\mathrm{d}t} = -\mu_x I_x(t) + \lambda_x m \left[p \theta_x(t) + (1-p) \sum_{y \neq x} \frac{N_x}{N - N_y} \theta_y(t) \right] \tag{3.14}
$$

The next step is to find the dynamical equation for $\theta(t)$. Multiplying by $\int \mathrm{d}a\ a$ equation (3.13), we get:

$$\begin{aligned}
\int \mathrm{d}a\ aI_{a,x}^{t+\Delta t} = \theta_x^{t+\Delta t} = \int \mathrm{d}a\ aI_{a,x}^t &- \mu_x \Delta t \int \mathrm{d}a\ aI_{a,x}^t + \\
&+ \frac{\lambda_x m p \Delta t}{N_x} \int \mathrm{d}a\ aS_{a,x}^t \int \mathrm{d}a' a' I_{a',x}^t + \\
&+ \lambda_x m (1-p) \Delta t \int \mathrm{d}a\ aS_{a,x}^t \sum_{y \neq x} \int \mathrm{d}a' a' \frac{I_{a',y}^t}{N - N_y} = \\
\cong \theta_x^t &- \mu_x \Delta t \int \mathrm{d}a\ aI_{a,x} + \\
&+ \frac{\lambda_x m p \Delta t}{N_x} \int \mathrm{d}a\ aN_{a,x} \int \mathrm{d}a' a' I_{a',x}^t + \\
&+ \lambda_x m (1-p) \Delta t \int \mathrm{d}a\ aN_{a,x} \sum_{y \neq x} \int \mathrm{d}a' a' \frac{I_{a',y}^t}{N - N_y} = \\
= \theta_x^t &- \mu_x \Delta t \theta_x^t + \\
&+ \lambda_x m p \Delta t \left\langle a \right\rangle_x \theta_x^t + \\
&+ \lambda_x m (1-p) \Delta t \left\langle a \right\rangle_x \sum_{y \neq x} \frac{N_x}{N - N_y} \int \mathrm{d}a' a' I_{a',y}^t = \\
= \theta_x^t &- \mu_x \Delta t \theta_x^t + \\
&+ \lambda_x m p \Delta t \left\langle a \right\rangle_x \theta_x^t + \\
&+ \lambda_x m (1-p) \Delta t \left\langle a \right\rangle_x \sum_{y \neq x} \frac{N_x}{N - N_y} \theta_y^t
\end{aligned}$$

where we have used again the early-stage approximation. Performing the continuous limit ($\Delta t \to 0$):

$$\frac{\mathrm{d}\theta_x(t)}{\mathrm{d}t} = -\mu_x \theta_x(t) + \lambda_x m \left\langle a \right\rangle_x \left[p\theta_x(t) + (1-p) \sum_{y \neq x} \frac{N_x}{N - N_y} \theta_y(t) \right] \tag{3.15}$$

To summarize:

$$\begin{cases}
\dot{I}_x = -\mu_x I_x + \lambda_x m \left[p\theta_x + (1-p) \sum_{y \neq x} \frac{N_x}{N - N_y} \theta_y \right] \doteq g_x & \text{for } x = 1, \ldots, Q \\
\dot{\theta_x} = -\mu_x \theta_x + \lambda_x m \left\langle a \right\rangle_x \left[p\theta_x + (1-p) \sum_{y \neq x} \frac{N_x}{N - N_y} \theta_y \right] \doteq h_x & \text{for } x = 1, \ldots, Q
\end{cases}$$

The system of $2Q$ equations can be rewritten in matrix form as:

$$\begin{pmatrix} \dot{I}_1 \\ \vdots \\ \dot{I}_Q \\ \dot{\theta}_1 \\ \vdots \\ \dot{\theta}_Q \end{pmatrix} = J \cdot \begin{pmatrix} I_1 \\ \vdots \\ I_Q \\ \theta_1 \\ \vdots \\ \theta_Q \end{pmatrix} = \begin{pmatrix} \nabla g_1 \\ \vdots \\ \nabla g_Q \\ \nabla h_1 \\ \vdots \\ \nabla h_Q \end{pmatrix} \cdot \begin{pmatrix} I_1 \\ \vdots \\ I_Q \\ \theta_1 \\ \vdots \\ \theta_Q \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial I_1} & \cdots & \frac{\partial g_1}{\partial I_Q} & \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_1}{\partial \theta_Q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_Q}{\partial I_1} & \cdots & \frac{\partial g_Q}{\partial I_Q} & \frac{\partial g_Q}{\partial \theta_1} & \cdots & \frac{\partial g_Q}{\partial \theta_Q} \\ \frac{\partial h_1}{\partial I_1} & \cdots & \frac{\partial h_1}{\partial I_Q} & \frac{\partial h_1}{\partial \theta_1} & \cdots & \frac{\partial h_1}{\partial \theta_Q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_Q}{\partial I_1} & \cdots & \frac{\partial h_Q}{\partial I_Q} & \frac{\partial h_Q}{\partial \theta_1} & \cdots & \frac{\partial h_Q}{\partial \theta_Q} \end{pmatrix} \cdot \begin{pmatrix} I_1 \\ \vdots \\ I_Q \\ \theta_1 \\ \vdots \\ \theta_Q \end{pmatrix}$$

In our case:

$$J = \left(\begin{array}{ccc|ccc} -\mu_1 & \dots & 0 & p\lambda_1 m & \lambda_1 m(1-p)C_{1,2} & \dots & \lambda_1 m(1-p)C_{1,Q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & -\mu_Q & \lambda_Q m(1-p)C_{Q,1} & \lambda_Q m(1-p)C_{Q,2} & \dots & p\lambda_Q m \\ \hline 0 & \dots & 0 & -\mu_1 + \beta_1 p & (1-p)\beta_1 C_{1,2} & \dots & (1-p)\beta_1 C_{1,Q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & (1-p)\beta_Q C_{Q,1} & (1-p)\beta_Q C_{Q,2} & \dots & -\mu_Q + \beta_Q p \end{array}\right)$$

where $\beta_x = m\lambda_x \langle a \rangle_x$ and we have defined $C_{x,y} \doteq \frac{N_x}{N - N_y}$. We observe that:

$$J = \left(\begin{array}{c|c} A & B \\ \hline 0 & C \end{array}\right)$$

Thanks to the specific form of the Jacobian matrix, one can calculate the determinant as:

$$\begin{aligned} \det(J - t\mathbb{I}) &= \det(A - t\mathbb{I}) \cdot \det(C - t\mathbb{I}) = \\ &= \det \begin{pmatrix} -\mu_1 - t & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & -\mu_Q - t \end{pmatrix} \cdot \det \begin{pmatrix} -\mu_1 + \beta_1 p - t & \dots & (1-p)\beta_1 C_{1,Q} \\ \vdots & \ddots & \vdots \\ (1-p)\beta_Q C_{Q,1} & \dots & -\mu_Q + \beta_Q p - t \end{pmatrix} = \\ &= (-\mu_1 - t) \dots (-\mu_Q - t) \cdot \det \begin{pmatrix} -\mu_1 + \beta_1 p - t & \dots & (1-p)\beta_1 C_{1,Q} \\ \vdots & \ddots & \vdots \\ (1-p)\beta_Q C_{Q,1} & \dots & -\mu_Q + \beta_Q p - t \end{pmatrix} \end{aligned}$$

We see that $-\mu_1, \dots, -\mu_Q$ are eigenvalues of matrix A; this means that the largest eigenvalue would be associated with matrix D:

$$\det(D) = \det \begin{pmatrix} -\mu_1 + \beta_1 p - t & \dots & (1-p)\beta_1 C_{1,Q} \\ \vdots & \ddots & \vdots \\ (1-p)\beta_Q C_{Q,1} & \dots & -\mu_Q + \beta_Q p - t \end{pmatrix}$$

The biggest eigenvalue $t_{\max}$ would be of the form:

$$t_{\max} = -\sum_x \mu_x + p\sum_x \beta_x + \Xi > 0$$

with Ξ a function that depends on the combined effects of each class's average activation, infection rate, and recovery rate, as well as on the mixing between classes. The reproductive number R_0 is then:

$$\boxed{R_0 \doteq \frac{p\sum_x \beta_x + \Xi}{\sum_x \mu_x} > 1}$$

To summarize, the parameters are:

- Q = number of gullibility classes,
- $\{\lambda_x\}$ = set of λ's,
- $\{\mu_x\}$ = set of μ's,
- p = prob. to select a node (between m nodes) from the same susceptibility class.

Case $Q = 1$

In this case one finds, thankfully, the solution found in the absence of susceptibility classes, indeed (with $p = 1$):

$$\begin{pmatrix} \dot{I}_1 \\ \dot{\theta}_1 \end{pmatrix} = \begin{pmatrix} -\mu_1 & \lambda_1 m \\ 0 & -\mu_1 + \lambda_1 m \langle a \rangle_1 \end{pmatrix} \cdot \begin{pmatrix} I_1 \\ \theta_1 \end{pmatrix}$$

and so $R_0 = \frac{\lambda_1}{\mu_1} m \langle a \rangle_1$.

Case $Q = 2$

In this case we have that:

$$\begin{pmatrix} \dot{I}_1 \\ \dot{I}_2 \\ \dot{\theta}_1 \\ \dot{\theta}_2 \end{pmatrix} = \begin{pmatrix} -\mu_1 & 0 & p\lambda_1 m & (1-p)\lambda_1 m C_{1,2} \\ 0 & -\mu_2 & (1-p)\lambda_2 m C_{2,1} & p\lambda_2 m \\ 0 & 0 & -\mu_1 + p\beta_1 & (1-p)\beta_1 C_{1,2} \\ 0 & 0 & (1-p)\beta_2 C_{2,1} & -\mu_2 + p\beta_2 \end{pmatrix} \cdot \begin{pmatrix} I_1 \\ I_2 \\ \theta_1 \\ \theta_2 \end{pmatrix}$$

We calculate the determinant as:

$$\begin{aligned} \det(J - t\mathbb{I}) &= \det(A - t\mathbb{I}) \cdot \det(C - t\mathbb{I}) = \\ &= \det \begin{pmatrix} -\mu_1 - t & 0 \\ 0 & -\mu_2 - t \end{pmatrix} \cdot \det \begin{pmatrix} -\mu_1 + p\beta_1 - t & (1-p)\beta_1 C_{1,2} \\ (1-p)\beta_2 C_{2,1} & -\mu_2 + p\beta_2 - t \end{pmatrix} = \\ &= (-\mu_1 - t)(-\mu_2 - t) \cdot \det \begin{pmatrix} -\mu_1 + p\beta_1 - t & (1-p)\beta_1 C_{1,2} \\ (1-p)\beta_2 C_{2,1} & -\mu_2 + p\beta_2 - t \end{pmatrix} \end{aligned}$$

Focusing on the second determinant:

$$\begin{aligned}\det\begin{pmatrix}-\mu_1+p\beta_1-t & (1-p)\beta_1C_{1,2}\\ (1-p)\beta_2C_{2,1} & -\mu_2+p\beta_2-t\end{pmatrix} &= (-\mu_1+p\beta_1-t)(-\mu_2+p\beta_2-t)+\\ &\quad -((1-p)\beta_1C_{1,2})((1-p)\beta_2C_{2,1}) =\\ &= (-\mu_1+p\beta_1)(-\mu_2+p\beta_2)+t^2+\\ &\quad -(-\mu_1+p\beta_1)t-(-\mu_2+p\beta_2)t+\\ &\quad -(1-p)^2\beta_1\beta_2C_{1,2}C_{2,1} =\\ &= \mu_1\mu_2+p^2\beta_1\beta_2-\mu_1p\beta_2-\mu_2p\beta_1+t^2+\\ &\quad -(-\mu_1-\mu_2+p\beta_1+p\beta_2)t+\\ &\quad -\left(1+p^2-2p\right)\beta_1\beta_2C_{1,2}C_{2,1}\end{aligned}$$

Then:

$$\mu_1\mu_2+p^2\beta_1\beta_2-\mu_1p\beta_2-\mu_2p\beta_1+t^2-(-\mu_1-\mu_2+p\beta_1+p\beta_2)t-\left(1+p^2+2p\right)\beta_1\beta_2C_{1,2}C_{2,1} =$$
$$= t^2-(p(\beta_1+\beta_2)-(\mu_1+\mu_2))t+\mu_1\mu_2+p^2\beta_1\beta_2-p(\mu_1\beta_2+\mu_2\beta_1)-(1-p)^2\beta_1\beta_2C_{1,2}C_{2,1}=0$$

Denoting:

$$\begin{cases}B=(p(\beta_1+\beta_2)-(\mu_1+\mu_2))\\ C=\mu_1\mu_2+p^2\beta_1\beta_2-p(\mu_1\beta_2+\mu_2\beta_1)-(1-p)^2\beta_1\beta_2C_{1,2}C_{2,1}\end{cases}$$

we simply obtain:

$$t^2-Bt+C=0$$

The solution is:

$$\begin{aligned}t_{1,2} &= \frac{B\pm\sqrt{B^2-4C}}{2} =\\ &= \frac{p(\beta_1+\beta_2)-(\mu_1+\mu_2)}{2}+\\ &\quad \pm\frac{\sqrt{\left[p(\beta_1+\beta_2)-(\mu_1+\mu_2)\right]^2-4\left[\mu_1\mu_2+p^2\beta_1\beta_2-p(\mu_1\beta_2+\mu_2\beta_1)-(1-p)^2\beta_1\beta_2C_{1,2}C_{2,1}\right]}}{2}\end{aligned}$$

Considering only the argument of the square root:

$$\left[p\left(\beta_1+\beta_2\right)-\left(\mu_1+\mu_2\right)\right]^2-4\left[\mu_1\mu_2+p^2\beta_1\beta_2-p(\mu_1\beta_2+\mu_2\beta_1)-\left(1-p\right)^2\beta_1\beta_2C_{1,2}C_{2,1}\right]=$$

$$\begin{aligned}=&p^2\left(\beta_1+\beta_2\right)^2+\left(\mu_1+\mu_2\right)^2-2p\left(\beta_1+\beta_2\right)\left(\mu_1+\mu_2\right)+\\&-4\mu_1\mu_2-4p^2\beta_1\beta_2+4p(\mu_1\beta_2+\mu_2\beta_1)+4\left(1-p\right)^2\beta_1\beta_2C_{1,2}C_{2,1}=\end{aligned}$$

$$\begin{aligned}=&p^2\beta_1^2+p^2\beta_2^2+2p^2\beta_1\beta_2+\mu_1^2+\mu_2^2+2\mu_1\mu_2-2p\beta_1\mu_1-2p\beta_1\mu_2-2p\beta_2\mu_1-2p\beta_2\mu_2+\\&-4\mu_1\mu_2-4p^2\beta_1\beta_2+4p(\mu_1\beta_2+\mu_2\beta_1)+4\left(1-p\right)^2\beta_1\beta_2C_{1,2}C_{2,1}=\end{aligned}$$

$$\begin{aligned}=&p^2\beta_1^2+p^2\beta_2^2+2p^2\beta_1\beta_2+\mu_1^2+\mu_2^2+2\mu_1\mu_2-2p\beta_1\mu_1-2p\beta_1\mu_2-2p\beta_2\mu_1-2p\beta_2\mu_2+\\&-4\mu_1\mu_2-4p^2\beta_1\beta_2+4p\mu_1\beta_2+4p\mu_2\beta_1+4\left(1-p\right)^2\beta_1\beta_2C_{1,2}C_{2,1}=\end{aligned}$$

$$=p^2(\beta_1^2+\beta_2^2-2\beta_1\beta_2)+(\mu_1^2+\mu_2^2-2\mu_1\mu_2)-2p(\mu_1-\mu_2)(\beta_1-\beta_2)+4(1-p)^2\beta_1\beta_2C_{1,2}C_{2,1}=$$

$$=\left[p(\beta_1-\beta_2)-(\mu_1-\mu_2)\right]^2+4(1-p)^2\beta_1\beta_2\,C_{1,2}C_{2,1}$$

Therefore:

$$t_{1,2}=\frac{p\left(\beta_1+\beta_2\right)-\left(\mu_1+\mu_2\right)}{2}\pm\frac{\sqrt{\left[p(\beta_1-\beta_2)-(\mu_1-\mu_2)\right]^2+4(1-p)^2\beta_1\beta_2\,C_{1,2}C_{2,1}}}{2}$$

The biggest eigenvalue is:

$$\begin{aligned}t_2=&\frac{p\left(\beta_1+\beta_2\right)-\left(\mu_1+\mu_2\right)}{2}+\frac{\sqrt{\left[p(\beta_1-\beta_2)-(\mu_1-\mu_2)\right]^2+4(1-p)^2\beta_1\beta_2\,C_{1,2}C_{2,1}}}{2}=\\=&\frac{-\sum_{x=1,2}\mu_x+p\sum_{x=1,2}\beta_x+\Xi}{2}>0\end{aligned}$$

We have found the specific form of Ξ:

$$\Xi^2=\left[p(\beta_1-\beta_2)-(\mu_1-\mu_2)\right]^2+4(1-p)^2\beta_1\beta_2\,C_{1,2}C_{2,1}$$

When $Q=2$:

$$C_{1,2}=\frac{N_1}{N-N_2}=\frac{N_1}{N-(N-N_1)}=1=C_{2,1}$$

Therefore:

$$\boxed{\Xi^2=\left[p(\beta_1-\beta_2)-(\mu_1-\mu_2)\right]^2+4(1-p)^2\beta_1\beta_2} \tag{3.16}$$

The reproductive number R_0 is:

$$-\sum_{x=1,2}\mu_x + p\sum_{x=1,2}\beta_x + \Xi > 0 \quad \Rightarrow \quad \boxed{R_0 \doteq \frac{p\sum_{x=1,2}\beta_x + \Xi}{\sum_{x=1,2}\mu_x} > 1} \tag{3.17}$$

For $Q = 2$ and $\mu_1 = \mu_2 = \mu$, the authors show that, with $0 < p < 1$:

$$\min_x R_0^{(x)} \leq R_0(p) \leq \max_x R_0^{(x)}$$

with $R_0^{(x)} = \beta_x/\mu$ (i.e. the reproductive number as if the nodes in class x were isolated from all the others, $p = 1$). Furthermore, they show that for $\mu_1 \neq \mu_2$, there is a regime, $p > p^*$, where p^* can be computed analytically, in which mixing can increase fragility, i.e. $R_0(p) > \max_x R_0^{(x)}$. The coupled network is more vulnerable than each isolated category.

Chapter 4

The modeling of cognitive processes

This chapter focuses on presenting the modeling of individuals' learning and cognitive processes. We begin the chapter with an explanation of Instance-Based Learning Theory (IBLT), whose mechanisms, introduced in the Adaptive Control of Thought-Rational (ACT-R) cognitive model, have shown excellent applicability to the recognition of phishing emails in the literature. This is followed by an explanation of the Instance-Based Learning Model (IBLM) that has been chosen to implement phishing email detection in the integrated model with the network.

4.1 The Instance-Based Learning Theory (IBLT)

In this section, we will be looking at the Instance-Based Learning Theory as it was presented in the paper by Gonzalez et al. [31]. The theory is relevant to Dynamic Decision-Making (DDM), which describes the decision-making processes that occur in a dynamic environment involving multiple interdependent decisions made in real time within a changing context.

4.1.1 Introduction to instance-based learning theories

According to the studies of Simon and Langley [49], learning is a process that modifies the cognitive system in a relatively stable manner, improving future performance. Furthermore, they introduce several useful concepts for learning mechanisms: the **Knowledge base** refers to the accumulation of information in a declarative form, such as facts and experiences that can be retrieved and declared, **Recognition** refers to the ability to discriminate among familiar classes of objects and **Strategies** refer to the possible ways of acting and deciding that evolve with experience and feedback. **Evaluation functions** are used to judge the utility of possible alternatives, for example, deciding whether to continue searching for new options or to select the most suitable one.

Simon and Langley observe that decision-makers perform better when they apply heuristics (general rules) in a more flexible way. They interpret this as meaning that decision-makers increasingly rely on their accumulated knowledge over time, using past experience to guide their choices. Based on these findings, they propose that the most realistic learning process in DDM involves creating and recalling specific decision scenarios or examples.

There are several theories that support this intuition; briefly:

- The Chunking Theory by Chase and Simon [50]. It suggests that learning occurs through the accumulation of chunks in long-term memory and that experts in a given field recognize these chunks and save links to them in short-term memory. Experts conduct selective searches, relying on environmental cues to direct their attention. This theory contributes to the Knowledge base, Recognition and Strategies mechanisms.

- The Instance-based recognition model by Hintzman [51, 52]. In the proposed model each experience is stored as a memory trace containing a set of task features. In this framework, retrieval occurs through similarity-based matching between a probe and stored traces.

- The models based on similarity by Medin and Schaffer [53] and Nosofsky [54]. Medin and Schaffer assumes that decisions are based on similarity to previous examples, while Nosofsky presents a mathematical function for quantifying similarity. The concept of similarity is fundamental to many instance-based learning theories, and it is believed to have a crucial role in the mechanisms of Recognition, Strategy and for the Evaluation functions.

- The Instance Theory of Automatization by Logan [55]. The theory presents a model of skill acquisition which relies on the retrieval of examples from memory. Furthermore, it describes the transition from rule-based (algorithm-based) performance to instance-based performance. Indeed the author hypothesizes that people shift from applying rules to retrieving information from memory as instances accumulate. Unlike previous theories and models, Logan's theory does not rely on the concept of similarity in memory retrieval; it retrieves only identical examples. Nosofsky and Palmeri [56] extended Logan's model by proposing that memory retrieval is based on similarity. Logan's theory and its subsequent extensions play a role in both Strategy mechanism and for the Evaluation functions.

- Case-Based Reasoning (CBR) by Gilboa and Schmeidler [57]. It asserts that knowledge of complex tasks is developed by accumulating information about the state of the environment, the chosen solution, and the outcome.

In summary, instance-based learning theories share the following characteristics: the accumulation of instances in memory, the development of situation recognition and alternative

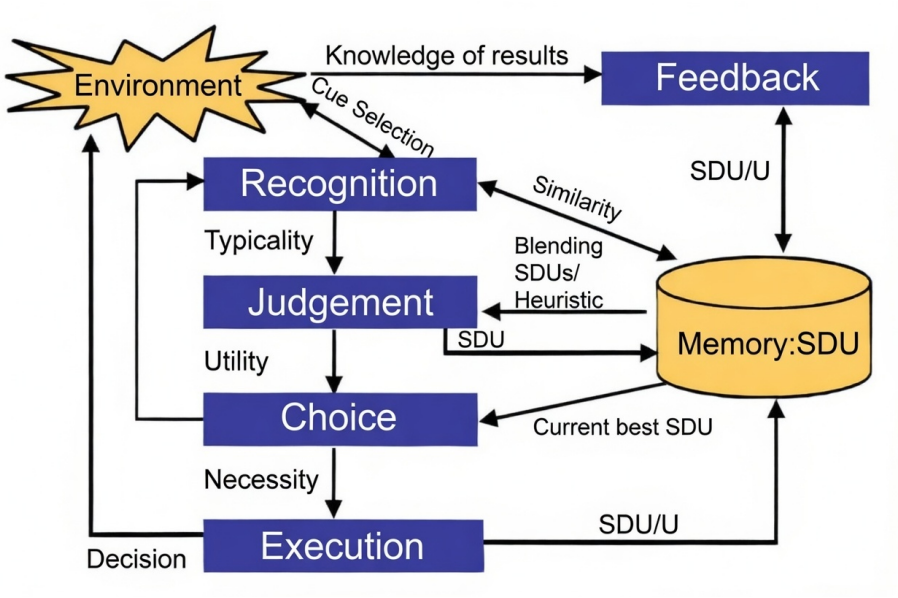


Figure 4.1 Schematic representation of main IBLT steps in DDM. Reproduced from [31].

research, a similarity-based memory retrieval and the transition from rule-based to instance-based performance. The authors argue that the features of instance-based learning theories make them particularly applicable to DDM.

4.1.2 Instance-based mechanisms in dynamic decision making context

This section presents the learning mechanisms within IBLT, as described by Simon and Langley [49], in the context of dynamic decision-making.

The definition of an instance in IBLT is a triplet with slots for situation, decision and utility (SDU). The situation (S) is a set of indicators that characterize the instance and will be useful for understanding how similar it is to a new situation, the decision (D) is the action taken, and the utility (U) is an evaluation of the outcome of that action.

As proposed by Simon and Langley, IBLT operates through five learning mechanisms in the context of a decision-making process:

1. Instance-based knowledge: accumulation of knowledge in the form of instances (as SDU);
2. Recognition-based retrieval: retrieval from memory based on similarity between current situation and past experiences;
3. Adaptive strategies: transition from heuristic strategies to instance-based strategies with time;

4. Necessity-based choice: managing of the trade-off between the research of possible alternatives and the choice of the best solution;

5. Feedback updates: update the utility of past SDUs according to the results of actions.

The main steps of the decision-making process outlined by IBLT are: recognition, judgment, choice and feedback (see Figure 4.1). Let us take a closer look at each of these steps.

Recognition This step involves recognizing the current situation; this skill improves over time, moving from heuristic-based approaches to directly retrieving a solution that depends on the specific context at hand. An inexperienced individual undertakes an inefficient and random exploration of possible options. Over time, they begin to identify familiar situations and retrieve the most appropriate solution. Furthermore, as they accumulate experience, decision makers learn to focus on the most significant elements. Through practice, individuals become increasingly selective about the information they use. IBLT also suggests that this increasing selectivity and attention to the most informative elements derive from the similarity (a measure of the distance between the characteristics of previous instances and those of the current one) between past decisions (SDUs) and the current situation. Recognition depends on attention, which is shaped by prior knowledge and guided by similarity between cases.

A new situation may be typical or atypical:

- Typical. A new situation is considered typical if similar instances exist in memory;
- Atypical. A new situation is atypical when no similar situations can be found in memory.

Judgment After categorizing a new situation as typical or atypical, decision makers evaluate which action might be the most appropriate. The IBLT proposes two judgment procedures, one for atypical situations and one for typical situations. Again, the IBLT proposes that decision makers adapt their strategies from heuristic-based judgments to instance-based judgments. More precisely:

- Atypical situation. In atypical situations, decision makers use heuristics to evaluate the utility of a decision. Judgments can be based, for example, on the instructions provided or even on random heuristics. Time heuristics, for example, recommend making a decision based on the time remaining.
- Typical situation. In typical situations, the IBLT proposes that decision makers determine the utility of an action by combining the utility of similar cases stored in memory. The evaluation of an action is a weighted sum of the similarity between situations and actions, both current and stored. To this end, the IBLT introduces the concept of **activation** associated with each SDU in memory, with the aim of determining its pertinence in relation to the current situation. Activation is a measure of

the relevance of an instance in memory in the current context. Once activation has been evaluated for each instance in memory, the stored utility values of all instances are combined to evaluate the current situation. A new SDU instance is then created whose utility is the result of all previous knowledge.

Choice Following the evaluation process, decision makers determine the action to be taken, presumably the best option. Rational theories of choice assume that all possible alternatives are known, enabling the decision maker to select the optimal option. However, in the context of DDM, this does not work, which is why the IBLT proposes an intermediate strategy. This strategy involves evaluating alternatives one by one, with the decision maker then choosing between continuing the search for further alternatives or implementing the currently preferred alternative after each evaluation. This choice is determined by the decision maker's level of necessity. The level of necessity indicates the need to make a decision and is determined by subjective or objective factors such as the remaining time. The level of necessity directly affects how many alternatives are evaluated before a decision is made. If no time is available, the decision maker will implement the best available alternative. Implementing a decision modifies the environment and the SDU stored in memory to indicate the selected alternative.

Feedback In order to improve performance, decision-makers should be able to measure the consequences of their decisions. For this reason, the process includes a feedback step. Instance-based learning models typically offer little to no information on how to implement a feedback mechanism. The IBLT, however, uses feedback to refine the SDUs of decision makers. When an SDU is created, its utility value is only a prediction based on experience, and its actual utility value remains unknown until feedback is provided. Once this is known, decision makers update the utility value in the SDUs, replacing the initial prediction. This updating process increases the weight that these SDUs will have in the future, as they have been refined based on feedback.

4.2 An ACT-R implementation of IBLT

An Instance-Based Learning Model (IBLM) was constructed in the Adaptive Control of Thought-Rational (ACT-R) cognitive architecture. The ACT-R mechanisms and parameters, as introduced by Anderson and Lebiere [58], are presented next.

ACT-R is an architecture for cognitive modeling, based on principles of cognition, supported by many studies and experiments. The ACT-R architecture is particularly well suited for implementing the IBLT mechanisms, as it represents knowledge in two forms: procedural (If-Then rules) and declarative (chunks). Chunks are used for learning and are formed when a goal is achieved; this knowledge can be retrieved through a range of methods. This framework appears to be the optimal setting for implementing the proposals of

Logan's theory that were presented earlier. Furthermore, in ACT-R, knowledge stored in declarative form is retrieved based on the activation level associated with each chunk.

4.2.1 ACT-R mechanisms and parameters

As we have just mentioned, the ACT-R cognitive architecture includes the concept of chunk activation in memory, which is also present in IBLT itself. Within ACT-R, the activation associated to chunk i is defined as:

$$A_i = B_i + \sum_j W_j S_{ji}$$

with B_i a base-level activation of chunk i, S_{ji} a measure of the relevance of chunk j with respect to chunk i and W_j a weighting factor for the j-th chunk (i.e. the "attention" given to it).

The ACT-R architecture defines the Partial Matching (PM) mechanism, which allows a variable number of chunks to be retrieved that may only partially match the current situation. The Partial Matching Equation reads:

$$M_{ip} = A_i - \text{MP} \sum_{v,d} (1 - \text{Sim}(v, d))$$

where M_{ip} is the match score between chunk i and current situation p, A_i is the activation just defined, $\text{Sim}(v, d)$ is a measure of similarity (similarity function) between the value v, from the current situation, and d, from the chunk, and MP is a constant factor used to calibrate the impact of dissimilarity in the match score. We observe that the subtracted term acts as a penalty in cases of dissimilarity, being equal to zero in cases of perfect matching, $M_{ip} = A_i$. MP can be modulated to find the best compromise between activation and matching.

More generally, the Blending mechanism is introduced. Blending is a generalization of PM that compute a result from multiple chunks of memory instead of selecting a single chunk. While PM retrieves the chunk with the highest match score relative to the current situation, blending involves each chunk participating with a value V_i, with each chunk's contribution weighted by its activation and similarity. The final result, V^*, is therefore a weighted average of the values proposed by the chunks, taking into account how much each chunk is similar to the current situation. The Blending Equation reads:

$$V^* = \underset{V}{\text{argmin}} \left\{ \sum_i P_i \left(1 - \text{Sim}(V, V_i)\right)^2 \right\}$$

with V the set of possible actions. The probability P_i of recovering a chunk i is defined as a function of its match score M_{ip}, compared with the match scores of all other chunks. The

formula is given by the Boltzmann equation:

$$P_i = \frac{e^{\frac{M_{ip}}{t}}}{\sum_j e^{\frac{M_{jp}}{t}}}$$

where t is an activation noise (such as temperature).

Blending is a mechanism in ACT-R that allows an aggregate value to be calculated from multiple chunks (units of knowledge) present in memory. Rather than selecting a single chunk, the idea is to combine information from all relevant chunks based on their activation and similarity to the current situation.

4.3 The Instance-Based Learning Model (IBLM)

In this section, we introduce the Instance-Based Learning Model (IBLM) proposed by Cranford et al. [15]. This framework will be used in the integrated model with the network. The model they developed, in particular, simulates users' responses to receiving phishing emails.

Building on the strong foundations of the IBLT implemented in ACT-R, the authors propose simplifications and changes that are specific to the phenomenon being modeled: the reaction of users to phishing emails. The authors argue that for emails, there is usually a gap between actions and feedback on an email's actual maliciousness. So, for this task, they only represent the context and the action in the elementary knowledge element (chunk), not the outcome (the utility). The context elements (features) of an email are the sender's email address, the subject line, the email body and the link. The possible actions are to either click or not click the link. Therefore, for each new email received (see Figure 4.2), the model considers the email's context and generates an action by retrieving past instances in memory.

In ACT-R the retrieval of an instance depends on the activation strength of the associated chunk in memory. The activation A_i of a chunk i reads:

$$A_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) + \mathrm{MP} \cdot \sum_j \mathrm{Sim}\left(v_j, c_j^{(k)}\right) + \varepsilon_i \tag{4.1}$$

where:

- the first term accounts for the forgetting (recency) phenomena; t_j is the time since the jth occurrence of chunk i and d is the decay rate of each occurrence (ACT-R sets this value equal to 0.5);
- the second term resembles a partial matching process; $\mathrm{Sim}\left(v_j, c_j^{(k)}\right)$ is the similarity between the jth feature value in memory $c_j^{(k)}$ of chunk k and the corresponding value in

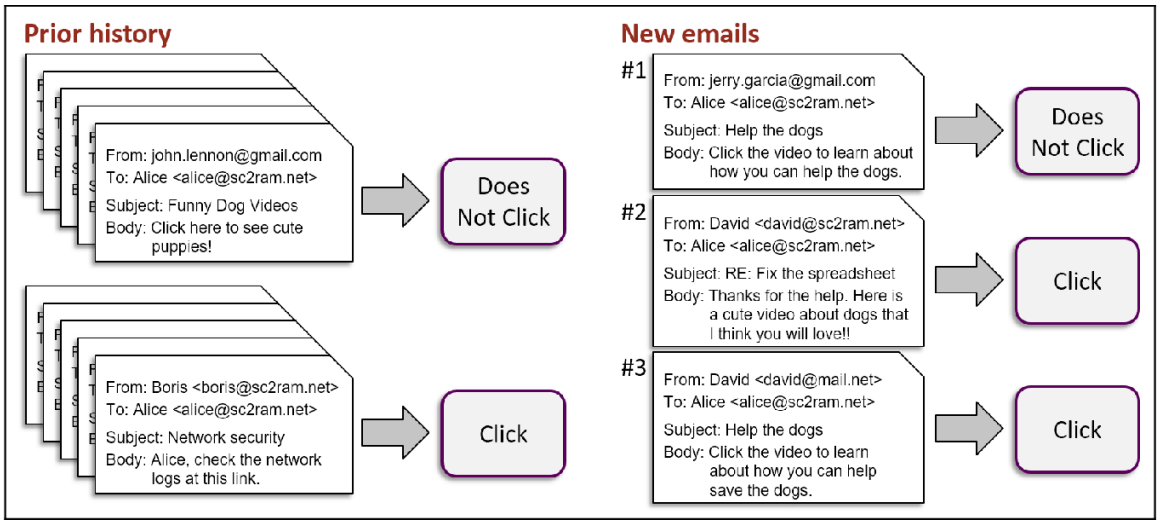


Figure 4.2 An example phishing attack scenario. The third (phishing) email shares some features with both the first (phishing) and the second (safe) email. Reproduced from [15].

the current situation v_j (is scaled by the mismatch penalty (MP), set by default equal to 1). We note that $\text{Sim}(v_j, c_j^{(k)})$ value ranges from -1 to 0, representing maximum dissimilarity and perfect matching, respectively.

- the third term is a noise, sampled from a logistic distribution with a null mean and variance parameter s (ACT-R sets this value equal to 0.25), introduced to add stochasticity in retrieval.

It should be mentioned that the authors used real emails with textual content and employed the University of Maryland Baltimore County's (UMBC) semantic-textual-similarity tool to assess similarity.

The probability of retrieving a specific instance is calculated using the Boltzmann equation, which takes into account the strength of the activation A_i and the temperature t (set to 1):

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}}$$

The IBLM uses ACT-R's blending mechanism. The blending equation reads:

$$V^* = \underset{V}{\text{argmin}} \left\{ \sum_i P_i \cdot (1 - \text{Sim}(V, V_i))^2 \right\}$$

with V the set of possible actions. The solution V^* is the one that minimizes the difference between the possible action V and the actual answer V_i contained in chunk i, weighted

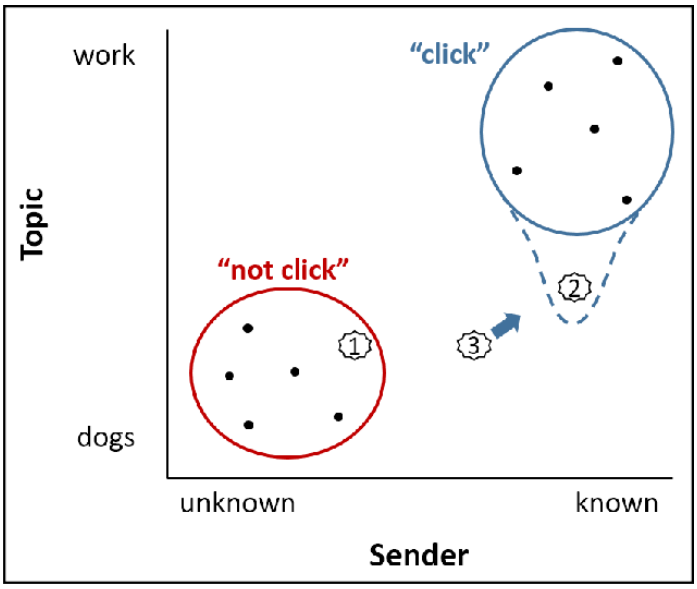


Figure 4.3 Here is represented the memory dynamics. In the plot are shown the effects of the addition of new instances (1, 2, 3) in memory. One observes that the areas associated with the action "click" and "not click" change over time; the third email is a phishing email, but it seems to be more similar to a safe one; the decision maker falls victim to a virus. Reproduced from [15].

by the probability of retrieval P_i (a measure of chunk i relevance in the current situation). After an action is generated, the experience (the context plus the action) is stored in declarative memory as a new instance (or chunk); such instance will influence future decisions (see Figure 4.3).

The authors validated the model with experiments involving human participants. On average, the model can predict human behavior in 58.6% of cases involving benign emails and 63.4% of cases involving phishing emails. The model was more accurate in predicting behavior in cases involving phishing emails than in cases involving benign emails. As with humans, the model responded to a high quantity of phishing emails (39.7% and 39.0%, respectively). However, while humans responded to a higher number of harmless emails (47.9%), the model responded to only 39.5% of harmless emails. This suggests that the model has difficulties distinguishing between the two types of email. This could be partly explained by the fact that participants were less cautious than they would be in a real-life situation, which led, given the structure of the experiment, to a bias in the model's response. The model generally performs well in predicting general human behavior. However, it is less accurate in classifying a benign email as safe.

A few years later, the model was improved by Cranford et al. [32] as part of research

exploring the impact of frequency on phishing detection training. This model was then validated by reproducing the Phishing Training Task (PTT) in Singh et al. [34]. Without going into the details of the experiment at this stage, the changes to the model essentially involve adjusting the parameter values in the activation A_i. These changes are summarized in the following list:

- Mismatch penalty $MP = 1 \rightarrow 2$; increasing this value accentuated the differences between different emails while reinforcing the similarities between similar ones. This increased the overall discriminability of the model;
- The temperature $t = 1 \rightarrow \sqrt{2} \cdot s$; the value of 1 results in a more even distribution of weights among instances during retrieval. Reducing this value creates a less smooth distribution, with peaks around instances with higher activation. This means that more similar instances have a greater weight in the decision-making process;
- Decay rate $d = 0.5 \rightarrow 0$; the decay rate is linked to the fact that the impact of old instances on current decisions should be small (recency effects). Setting it to 0 means that all instances play a more similar role in the retrieval process, which reduces the effects of recency. Studies have shown that, for longer-lasting tasks, this new value is better at representing recency effects in the retrieval process. The ability to recognize phishing emails comes from a great amount of accumulated experience which must play a role. This experience has essentially reached a constant level. An acceptable solution for representing this memory phenomenon is to distribute activation more evenly across old and new instances.

We report that the model was initialized with an equal number of ham and phishing emails, despite the fact that humans actually experience many more ham emails than phishing emails in reality.

Let's now turn our attention to PTT, an experiment conducted by Singh et al. [34], later used by Cranford et al. [32] to validate their IBLM. The aim of the PTT was to study the impact of various learning factors (e.g. frequency effects) on phishing detection decisions. The PTT is divided into three phases:

1. Pre-test: participants receive 10 emails (2 phishing and 8 legitimate) and are asked to classify them.
2. Training: participants are presented 40 emails of which 10,20 or 30 are phishing emails. In this phase, feedback is provided on the actual nature of the emails received.
3. Post-test: participants receive 10 emails (2 phishing and 8 legitimate) and are asked again to classify them, in order to test the effect of frequency.

Cranford et al. run the model 10 times per participants (298 participants), exposing the model to the same stimuli as the humans. To model PTT, Cranford et al. initialized the

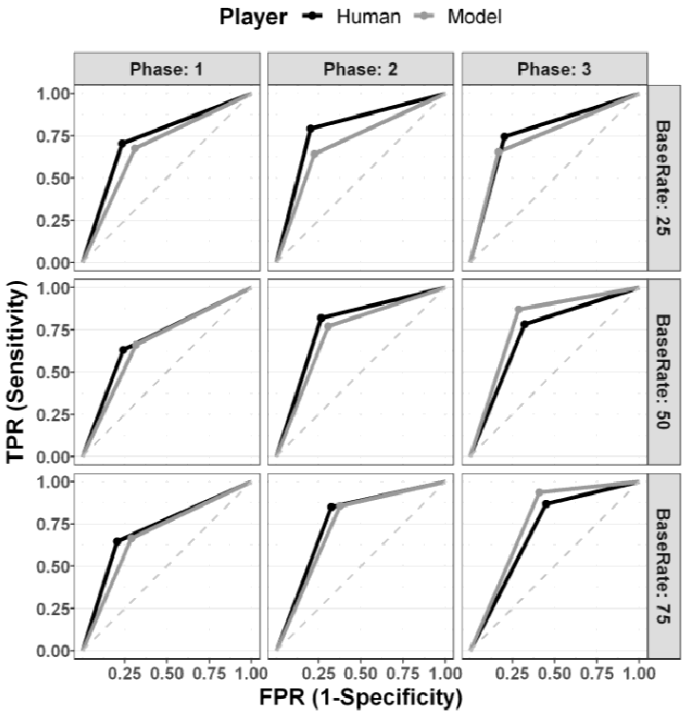


Figure 4.4 TPR vs FPR of phishing decision accuracy across three phases of the PTT for humans (black) and model (gray). The model provides an accurate representation of human behavior. Note, however, that human behavior is far from perfect. Reproduced from [32].

memory of the nodes with 10 emails (half malicious and half ham).

Cranford et al. use the definitions of True-Positive Rate (TPR) and the False-Positive Rate (FPR):

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{4.2}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{4.3}$$

with P the total number of Positive emails (malicious email), N the total number of Negative emails (ham emails), TP the True Positive (malicious email categorized as malicious), FP the False Positive (ham email categorized as malicious), FN False Negative (malicious email categorized as ham) and TN (ham email categorized as ham).

A brief review of the results indicates that the model replicates human behavior accurately across the various phases and conditions (see Figure 4.4).

Chapter 5

Results

5.1 Modeling framework

In previous chapters, we introduced all the necessary components for developing the integrated model. We presented a model for the time-varying network and studied how cyber threats spread on such network. Finally, we presented a model for describing the human cognitive processes exploited by cyber threats. In this chapter, we will define the framework of the integrated model, explaining the measures, simplifications, and solutions adopted to integrate all components in a coherent way supported by the literature.

The system we intend to describe is a dynamic network, whose connections vary over time, in which each node, or individual, is characterized by its own individual experience, which is used to make decisions. The nodes in the network continuously exchange emails, both safe and phishing; infected nodes send both safe and phishing emails. There is a possibility that a malicious email could infect a contacted node, but this would be dependent on its own decision. This enables the cyber threat to spread throughout the network.

5.1.1 Email definition, generation and comparison

In order to perform simulations, it is essential to define how emails are modeled. In the IBLM presented in the previous chapter, real emails are used, with textual content, sender email address, subject line and links. Due to the complexity of the model, it was considered methodologically correct to start with a simplified version of the emails in order to avoid heavy computations, as well as to make the results as interpretable as possible.

In general, an email, even with textual content, can be modeled as vector of s features:

$$\vec{C} = (C_1, \ldots, C_s)$$

The features represent the textual content, the sender's email address, and so on. If we consider node i with M_i instances in memory, the memory chunks (or in can be visualized

as follows:

$$\begin{aligned}
&\text{Chunk } 1 \quad \rightarrow \quad \left[\vec{C}^{(1)} = \left(C_1^{(1)}, \ldots, C_s^{(1)}\right) \quad + \quad \text{choice}^{(1)} = \text{"to click"}\right] \\
&\quad \vdots \\
&\text{Chunk } M_i \quad \rightarrow \quad \left[\vec{C}^{(M_i)} = \left(C_1^{(M)}, \ldots, C_s^{(M_i)}\right) \quad + \quad \text{choice}^{(M_i)} = \text{"to click"}\right]
\end{aligned}$$

The number M_i of instances (or chunks) in memory of node i is not fixed, but it changes over time, $M_i = M_i(t)$. In general $M_i \neq M_j$ for $i \neq j$.

We model the email with one feature ($s = 1$). Therefore:

$$\vec{C} = (C_1)$$

We have defined the single feature such that it takes numerical values $C_1 \in \mathbb{R}$ in a predetermined interval, $C_1 \in [0,\ C_{\max}]$, where the extrema corresponds to "not suspicious at all" and "highly suspicious":

- $C_1 = 0 \quad \longleftrightarrow \quad$ "not suspicious at all",
- $C_1 = C_{\max} \quad \longleftrightarrow \quad$ "highly suspicious".

The feature is therefore an aggregate variable that takes into account all the characteristics of the email, such as its content and subject line. Therefore, if the feature of an email is close to $C_{\max}$, it is perceived as potentially phishing, while if it is close to 0, it is perceived as safe.

Once the emails in our model have been defined, it is necessary to design the email generation process, which is fundamental for studying the spread of cyber threats. Since the feature takes numerical values, the most immediate solution is to sample it from a probability distribution function. In particular, we introduce two probability distribution functions: one for phishing emails, $f_{\text{p}}(C_1)$, and one for safe emails, $f_{\text{s}}(C_1)$. More specifically, the probability distributions will be asymmetric, with a peak at $C_1 = C_{\max}$ and $C_1 = 0$, respectively. In our model, we use truncated Gaussian distributions (see Figure 5.1). Therefore, a phishing email is sampled from the probability distribution $f_{\text{p}}(C_1)$, which, on average, generates values closer to $C_1 = C_{\max}$, although lower values are not excluded. It is precisely these more deceptive phishing emails (i.e. those with lower C_1 values) that are most likely to be mistaken for safe emails and infect their victims. Similarly, a safe email is sampled from the probability distribution $f_{\text{s}}(C_1)$, which, on average, generates values closer to $C_1 = 0$.

We know that a necessary step of the IBLM is the evaluation of the activation A_i of memory chunk i, which is required to compute P_i and, then, to solve the optimization problem. We recall that the activation A_i is:

$$A_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) + \text{MP} \cdot \sum_j \text{Sim}\left(v_j, c_j^{(i)}\right) + \varepsilon_i \tag{5.1}$$

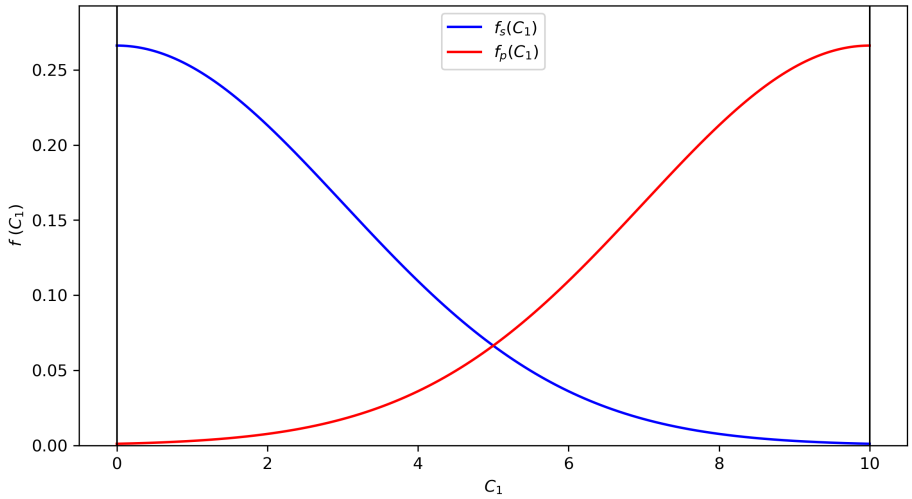


Figure 5.1 In the plot we see the probability distribution functions for the generation of phishing emails in red, $f_{\mathrm{p}}(C_1)$, and safe emails in blue, $f_{\mathrm{s}}(C_1)$. In general the Gaussian distribution takes values in $\mathbb{R}$ and is fully characterized by its average μ (here $\mu = 0, C_{\max}$) and the standard deviation σ (here $\sigma = 3$). Here the truncated Gaussian distributions are normalized over the interval $[0, C_{\max}]$. Here $C_{\max} = 10$.

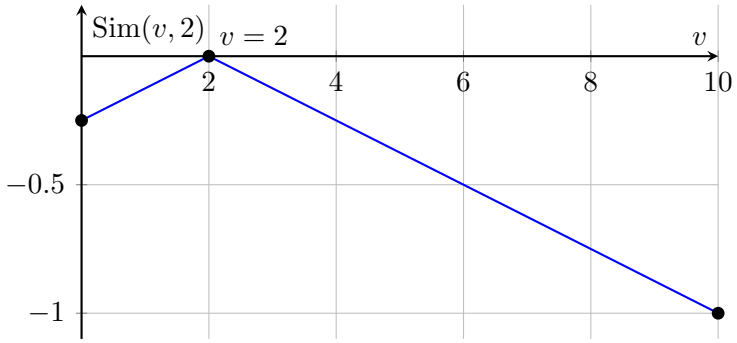
The activation relies on the similarity function $\mathrm{Sim}(v_j, c_j^{(i)})$, which quantifies the distance between the features of the email stored in memory and those of the email just received. In our model, each email is characterized by a single numerical feature $C_1 \in [0, C_{\max}]$; therefore, we can omit the index j. Since $\mathrm{Sim}(v, c) \in [-1, 0]$, we define it as linear function of v:

$$\mathrm{Sim}\,(v, c) = -\frac{|v - c|}{\max\{c,\ C_{\max} - c\}}$$

If we set $c = 2$ and $C_{\max} = 10$ we obtain:

$$\mathrm{Sim}\,(v, 2) = -\frac{|v-2|}{\max\{2,\ 10-2\}} = -\frac{|v-2|}{8} = \begin{cases} 0 & \text{if } v = 2 \\ -\frac{1}{4} & \text{if } v = 0 \\ -1 & \text{if } v = 10 \end{cases}$$

Below we show the plot of $\mathrm{Sim}(v, 2)$:



We note that this definition is consistent with the properties of the similarity function: it equals 0 when the match between v (the feature of the current email) and c (the feature of the email in memory) is perfect, and -1 when they are maximally dissimilar.

5.1.2 Adaptation of gullibility in the cognitive model

A key question is how the concept of gullibility, that in our case determines the phishing susceptibility, can be translated into a framework with the cognitive model. We recall that gullibility is a measure of how easily users can be deceived, i.e. their level of naivety: the higher their gullibility, the greater their propensity to fall into phishing traps. Without the cognitive model, differences in gullibility between individuals translated into different probabilities of infection (λ): the greater the gullibility, the greater the probability of being infected after receiving a phishing email. The introduction of the cognitive model means that this probability no longer exists; contagion is the result of the particular experience and knowledge of the individuals. As mentioned in the Introduction, knowledge and experience are fundamental factors in identifying phishing emails. The knowledge of the nodes (i.e. the instances in their memory) directly determines whether they will click on the link of a phishing email. The solution for the adaptation of gullibility into the integrated framework originates from this fact.

The idea is that knowledge should affect the chances of an individual falling victim to phishing. This depends on the quality of the knowledge, where with quality we mean the correct correspondence between the email feature in memory and the associated action. We expect an expert individual with excellent knowledge - where the majority of stored instances have the correct feature-action correspondence - to have a low probability of falling victim to phishing, as they can recognize safe and malicious emails quite well. In the opposite case we have an inexperienced individual with low-quality knowledge with an incorrect correspondence between features and actions; he decides, on average, randomly. This picture is therefore consistent with what has been shown in the literature.

For this reason, we introduce a fundamental parameter for defining memory quality, and therefore the gullibility of individuals: P_{gul}. The parameter just introduced is a probability, therefore $0 \leq P_{\text{gul}} \leq 1$, which regulates the fraction of instances in initial memory in

which the correspondence between feature and action is random. More precisely, P_{gul} is the probability that the action stored in the initial memory chunks is random, and therefore uncorrelated with the email feature. We have just mentioned that each node has an initial knowledge (memory). Such initial memory is initialized when the simulation starts; we will discuss this in more detail in the next section. We observe that:

- Low gullibility: the probability P_{gul} of having instances in initial memory with uncorrelated feature and action is small
- High gullibility: the probability P_{gul} of having instances in initial memory with uncorrelated feature and action is large

So, to summarize, in the integrated framework, gullibility is no longer modeled as a probability of infection (λ), but as a property of the individual's initial knowledge. We encode it through the memory quality parameter (P_{gul}), which sets the ratio for the random association between the feature and the action stored in the initialized memory chunks. A low value of P_{gul} corresponds to high-quality memory and lower susceptibility to phishing, while a high value of P_{gul} produces a more confused initial knowledge and a higher probability of selecting incorrect actions when receiving new emails.

5.1.3 Integrated model flow

The following section will go through the flow of the integrated model. The integrated model involves the study of SIS dynamics on a time-varying network, specifically an ADN, with the addition of the IBLM. The network dynamics are those described in the dedicated chapter, Chapter 2. The contagion dynamics is described by the SIS model as presented in Chapter 3. The mechanisms of the IBLM are those presented in Chapter 4.

In the integrated model, we have N individuals, each represented by a node in the activity-driven network. The fundamental difference is that the parameter λ is no longer applicable, as it is replaced by the IBLM. Indeed, during the simulation, when a node receives a message (i.e., an email), it must decide whether to click on the link; this decision is made through the IBLM decision-making process, mediated by the individuals' cognitive mechanisms.

Memory initialization We observe that individuals require an initial memory in order to make an action. For this reason, the first step is to initialize the individuals' memory. We do this by fixing the number of instances in the initial memory, M, and the gullibility level, i.e. the parameter P_{gul}. As suggested in the IBLM literature, the memory initialization process for each node consists of providing an initial memory containing an equal number of safe and phishing emails. For each of the M chunks in memory, an email is generated according to the procedure described in the previous section (either safe or phishing). Then, with probability P_{gul} the corresponding action is assigned at random, while with probability $1 - P_{\text{gul}}$ the action is assigned consistently with the email type, i.e. depending on whether its feature was sampled from the safe or phishing probability distribution.

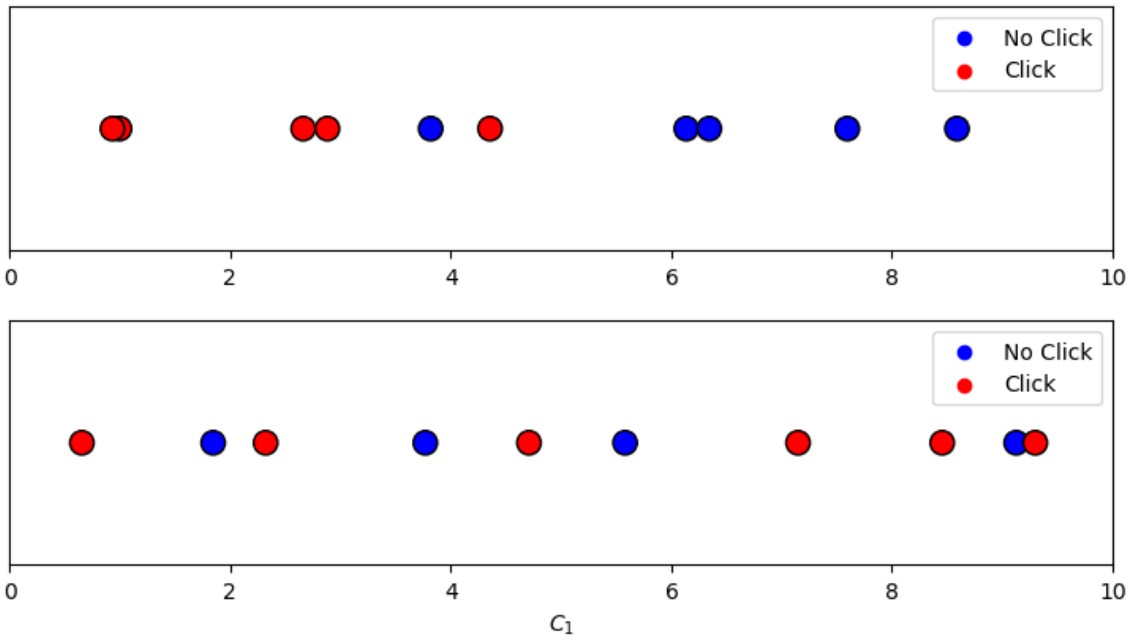


Figure 5.2 We visualize the initial memory of a node with $P_{\text{gul}} = 0$ (first plot) and $P_{\text{gul}} = 1$ (second plot) of size $M = 10$; the position of the dots represent the value of the feature stored in each chunk, from 0 to $C_{\max} = 10$. The dot color is the action associated in the chunk.

At the end of the memory initialization, all nodes have an initial memory that reflects their own gullibility and has size M_i, with $i = 1, \ldots, N$. We visualize in Figure 5.2 the memory of a node with $P_{\text{gul}} = 0$ (first plot) and $P_{\text{gul}} = 1$ (second plot). We recall that $C_1 = 0$ stands for a legitimate email, while $C_1 = C_{\max}$ stands for a malicious email. We observe that:

- when $P_{\text{gul}} = 0$ (first plot), the correspondence between the sampled feature and the associated action is always respected. We observe a dot whose feature lies within the range typically associated with legitimate emails, even though the node registered not to click. This can be interpreted as a phishing email that was successfully detected despite being well disguised (as suggested by its feature value). Of course, we can also observe the opposite case: an email whose feature suggests it is phishing, but the node decided to click (possibly because the sender looked familiar);
- when $P_{\text{gul}} = 1$ (second plot), the associated action is always random; it is as if a person with no prior experience with email decided whether to click or not randomly, regardless of the email feature, precisely because they lack the specific knowledge needed to recognize phishing emails.

SIS dynamics At the beginning of the SIS dynamics, a small fraction of nodes is infected; this very small fraction of the population is necessary to allow the cyber threats to spread. The SIS model has two compartments, infected (I) and susceptible (S). The recovery rate, indicated by μ, governs the transition $I \to S$, while the transition $I + S \to 2I$ is mediated by the IBLM.

Sending and receiving emails At each iteration of the SIS dynamics, nodes in the network activate according to their activity. Active nodes can be:

- Susceptible: in this case, the node sends one message to each of the m contacts selected according to the mechanisms described in Chapter 2. A susceptible node sends only safe emails; therefore, the features of the m emails are sampled from the safe-email probability distribution;
- Infected: in this case, the node sends one safe email and one phishing email to each of the m selected contacts, for a total of $2m$ messages. The email features are sampled from the corresponding probability distributions.

When a node receives an email, it decides which action to take according to the decision-making process presented in Chapter 4. In particular, it evaluates the activation A_i for all the M instances in memory, it computes all the P_ks (with P_k the probability to retrieve the kth instance) and it finally solves the optimization problem (blending mechanism) in order to decide which is the best action according to its knowledge. All emails received by the nodes are stored in memory once all nodes have completed the email sending and receiving process. Each node stores them as instances (feature + action) in its memory.

5.2 Results of the numerical simulations

In this section, we will discuss the results of the numerical simulations.

5.2.1 IBLM implementation

The first part of this section shows that our implementation of the IBLM - despite its assumptions and simplifications - accurately reproduces the results of the more realistic model and, more importantly, captures human behavior as reported in experiments from the literature. We first show that our IBLM implementation can distinguish phishing emails from legitimate ones when $P_{\text{gul}} = 0$.

Recalling the model definition, each individual is initialized with a memory of size M. Following the literature used to validate the model, this memory is initialized with an equal number of phishing and legitimate emails, representing past experience. In our implementation, we introduced the parameter P_{gul}, which controls the quality of this initial knowledge, as described in the previous chapter.

We observe Figure 5.3 where the outcome of the decision-making process of an individual with $P_{\text{gul}} = 0$ (first and second plots) and $P_{\text{gul}} = 1$ (third plot) is visualized. We note that:

- when $P_{\text{gul}} = 0$ (first plot), the individual's choice reflects past experience. In particular, for low values of the received email's feature, the node decides to click, as it did in the past, whereas for high values it decides not to click, suspecting the email may be malicious. We observe a transition region in which red and blue areas frequently alternate; this happens because, in this range, the node is uncertain about the nature of the email. The noise introduced in the activation of each chunk plays here a bigger role than in the rest of the spectrum of C_1. Further evidence that past experience influences future choices can be seen in the second plot, where a red dot appears in the region typically associated with phishing emails. As a result, for certain feature values, emails whose features are similar to that stored in memory are clicked.
- when $P_{\text{gul}} = 1$, also in this case the individual's choice reflects past behavior. The individual's low-quality experience leads him to decide to open emails across most of the C_1 spectrum, potentially clicking on malicious ones as well. The opposite case can also occur.

We now proceed to show that our implementation of the IBLM can reproduce human behavior. To this end, we rely on the Phishing Training Task (PTT) experiment conducted by Singh et al. [34], later used by Cranford et al. [32] to validate their IBLM.

We already presented the structure of the PTT experiment at the end of Chapter 4. We briefly review that it consists of three phases: pre-test, training and post-test. During the training three groups of participants are exposed to different frequencies of malicious emails (25%, 50% and 75%) and receive a feedback about the real nature of the email; the pre-test provides a ground truth while the post-test captures the effects of the training. The results reported by Cranford et al. are shown in Figure 4.4. We tested our implementation of the IBLM, showing its ability to qualitatively reproduce the results of Cranford et al. and, more importantly, human behavior. For this purpose, it is useful to plot TPR vs FPR, as this provides insight into the nodes' ability to detect malicious emails (TPR) and their propensity to avoid excessive false alarms (FPR).

As a first step, we calibrated σ, the standard deviation of the probability distributions ($f_{\text{p}}(C_1)$ and $f_{\text{s}}(C_1)$) from which emails are sampled during generation, such that the results would reproduce human behavior. Indeed, σ controls how much the two distributions overlap, and therefore how easy it is to distinguish between malicious and legitimate emails. Comparing the column of the pre-test phase in Cranford et al. experiment in Figure 4.4 and the curves in Figure 5.4, we note that an appropriate choice for the standard deviation σ would be $\sigma = 3$. From now on the standard deviation σ of $f_{\text{p}}(C_1)$ and $f_{\text{s}}(C_1)$ is set equal to 3.

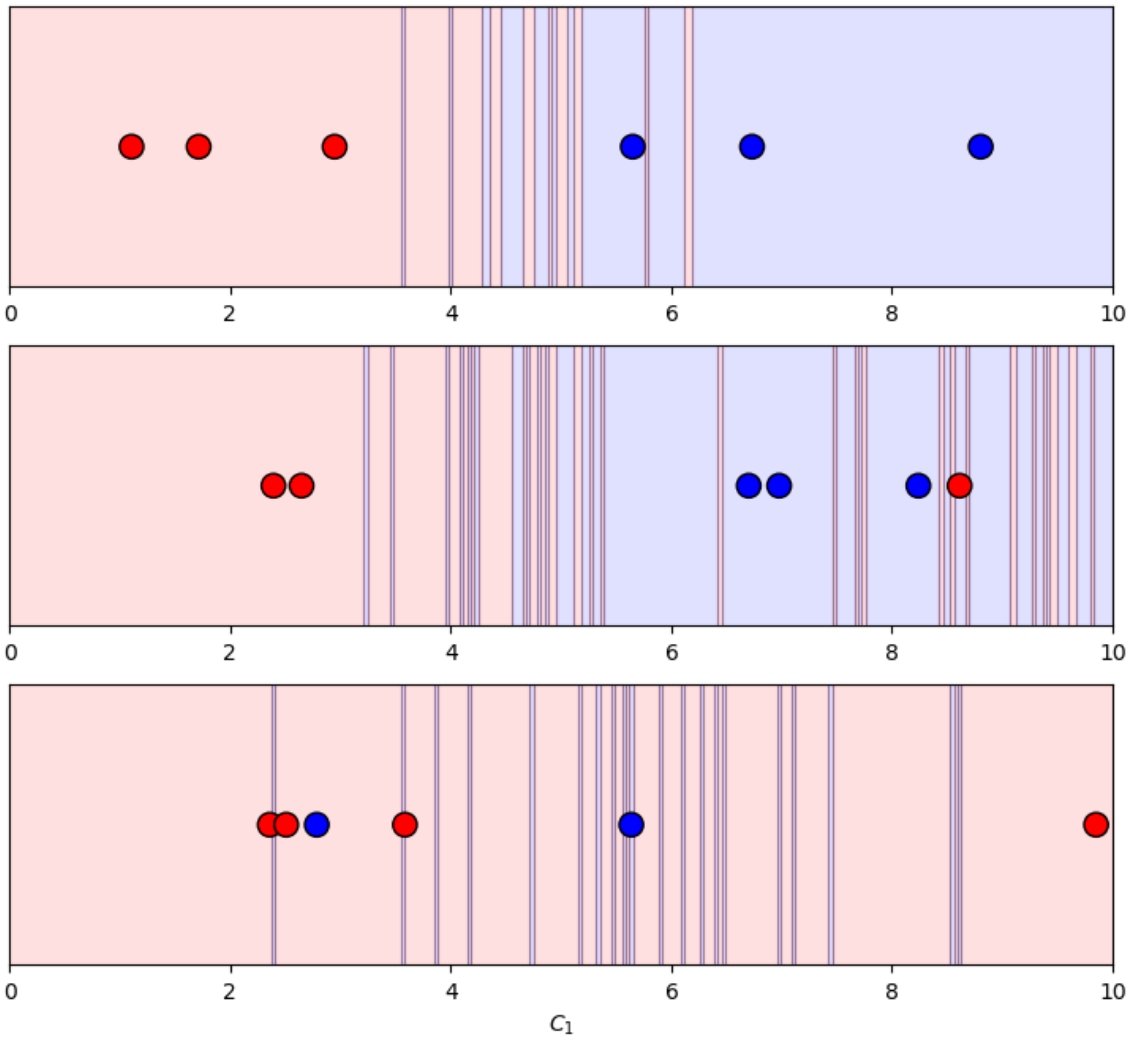


Figure 5.3 The plots show the action selected as a function of the feature value of the received email, for an individual with $P_{\text{gul}} = 0$ (first and second plots) and $P_{\text{gul}} = 1$ (third plot); each node has a memory size of $M = 6$ and $C_{\text{max}} = 10$. In the red regions, an email with those feature values would be clicked, whereas the blue regions correspond to feature values for which the individual would not click. The dot color corresponds to the action of the instance in memory: red ("click") and blue ("no click").

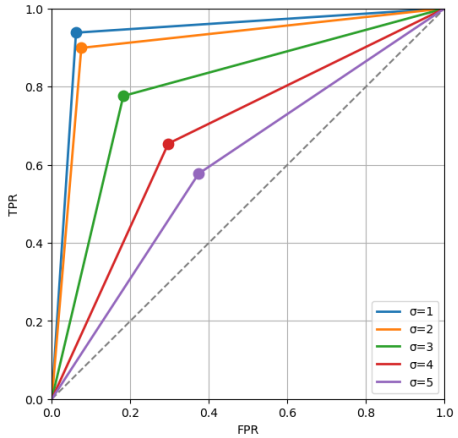


Figure 5.4 This plot shows the points (TPR, FPR) for different values of the standard deviation σ. Each point is obtained by reproducing the pre-test phase of PTT running the model 10 times (with $P_{\text{gul}} = 0.2$) for each of the 298 participants.

We investigated how the ability to detect phishing emails and to avoid excessive false alarms changes, for an expert node ($P_{\text{gul}} < 0.5$) and an inexperienced one ($P_{\text{gul}} > 0.5$), as the number of received emails increases (see Figure 5.5). We observe no substantial difference between more experienced and less experienced individuals: in both cases, the average TPR and FPR remain roughly constant, with a slight improvement in TPR for small numbers of received emails. So both expert and less experienced individuals, in absence of a feedback, maintain their initial skills, when they are exposed, on average, to an equal mix of safe and phishing emails.

We explored the effect of different values of P_{gul} on the TPR vs FPR plot (see Figure 5.6). What we observe is that, as individuals' gullibility increases (from $P_{\text{gul}} = 0$ to $P_{\text{gul}} = 1$), their ability to detect malicious emails worsens. The limit case $P_{\text{gul}} = 1$ is equivalent to an individual who makes a random choice between clicking and not clicking.

Finally we performed the complete PTT experiment using our implementation of the IBLM (see Figure 5.7). Comparing the results in Figure 4.4, we observe that we obtain the same qualitative behavior:

- At low frequency (25%): human participants showed no substantial change in TPR and an improvement in FPR (fewer false alarms). The model appears to reproduce this behavior, although it shows a slight decrease in TPR.
- At medium frequency (50%): human participants improved TPR, while slightly worsening FPR. The model shows slight changes between the pre-test and post-test, with an improvement in the TPR value.

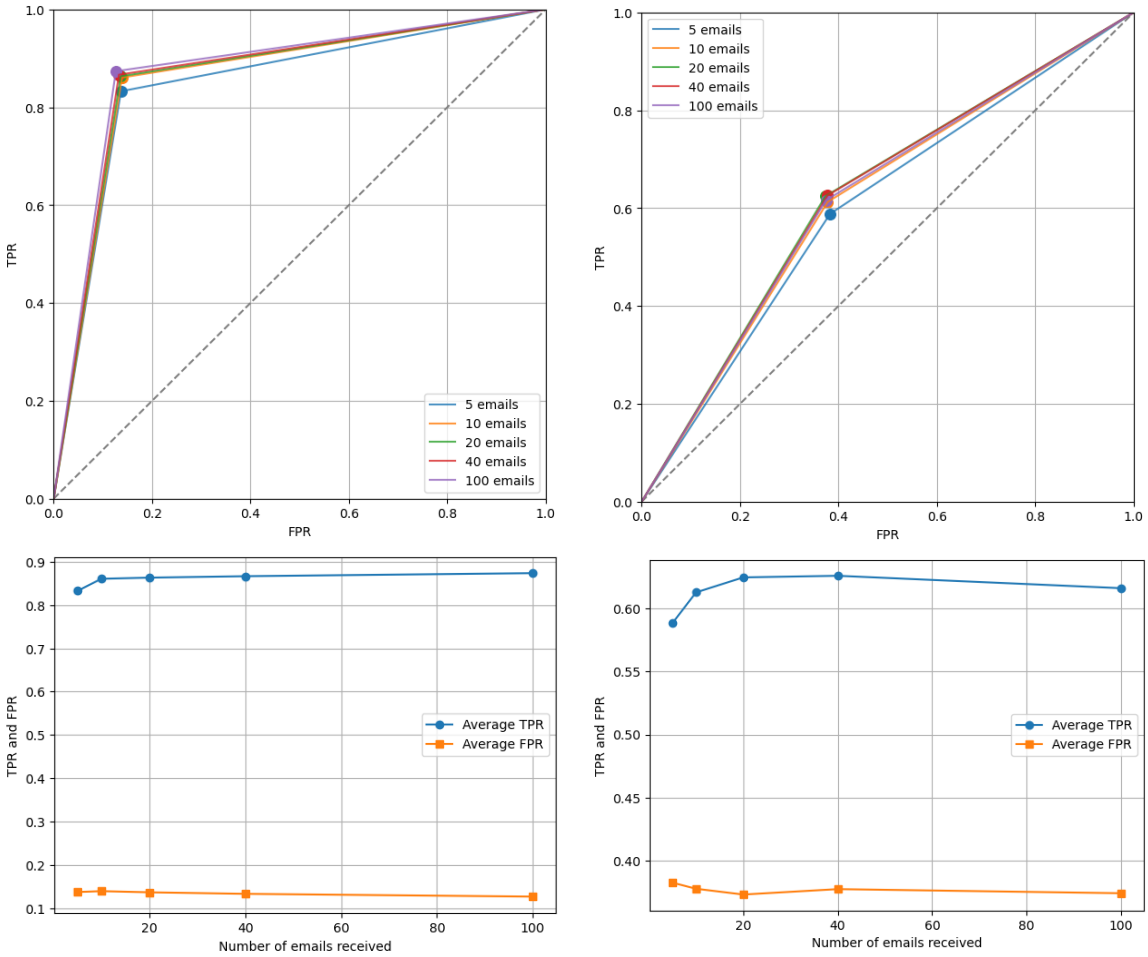


Figure 5.5 In this plot, we show how TPR (i.e. the ability to detect malicious emails) and FPR (i.e. the propensity to generate false alarms) change as a function of the number of received emails. Each point is an average over 3000 simulations; in each simulation, a single node is exposed to an increasing number of emails (on average, half malicious and half legitimate). In this case the initial memory size is $M = 10$.

- At high frequency (75%), human participants improved TPR but worsened FPR (more false alarms). The model seems to capture this trend, although the relative changes in both TPR and FPR are smaller; in general false alarms increase.

Therefore, our implementation of the IBLM can qualitatively recover human behavior when the model is exposed to different phishing-email frequencies. At low frequencies, individuals become less subject to false alarms, but their ability to detect malicious emails decreases. The opposite occurs at higher phishing frequencies: individuals are better at identifying potentially malicious emails, at the cost of a larger number of false alarms, exactly as ob-

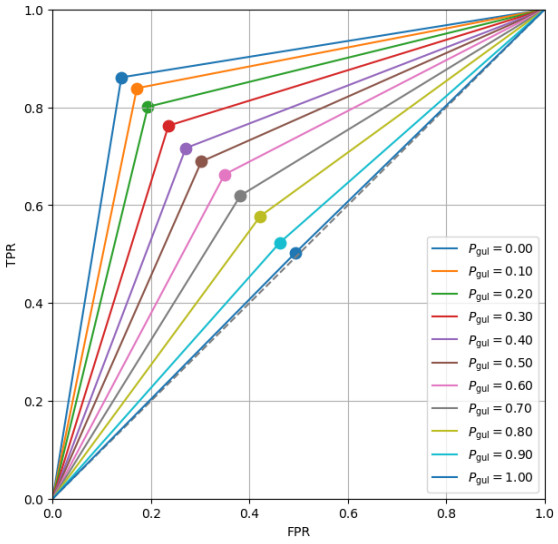


Figure 5.6 In this plot, we show how TPR and FPR vary as a function of individuals' gullibility (i.e. P_{gul}).

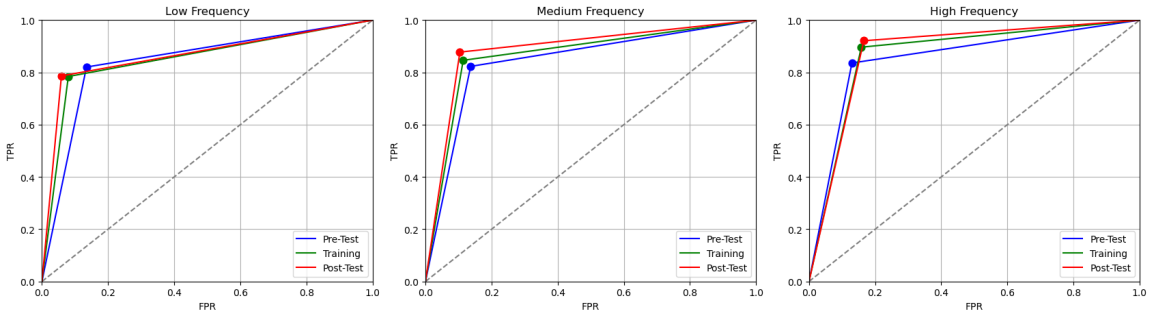


Figure 5.7 We reproduced the PTT experiment with 3000 individuals modeled by IBLM divided into three groups (Low, Medium and High frequency); all nodes have gullibility $P_{\text{gul}} = 0$.

served in human participants.

This email representation, while simplifying the model and reducing its realism, enables simulations with very large populations, N, without requiring, for example, a (necessarily finite) database of real emails with textual content, while still providing evidence of its reliability in reproducing human behavior.

5.2.2 Mapping between standard SIS dynamics and the IBLM

One of the limitations of the IBLM is that it does not allow the epidemic threshold to be derived analytically, due to the optimization problem behind it. For this reason, we looked for a relationship that would connect the model in which contagion is purely probabilistic (parameterized by λ) and our model that incorporates the IBLM. We derived a mapping that, in its current formulation, relies on a linear relationship. This mapping allows us to use the existing results for the probabilistic SIS model to a certain extent.

Before defining the mapping, we investigated the equilibrium value of the infected population (the SIS steady-state infection ratio, I_∞) by studying how it varies with P_{gul}. These preliminary analyses helped us to construct the mapping. We consider the Figure 5.8, which shows, in the first column, the value of I_∞ as a function of P_{gul} for three different values of the recovery rate μ. In all three cases, we can identify a threshold, a critical value of P_{gul}, such that the fraction of infected individuals is null below it and becomes greater than zero above it. We also observe that this critical value of P_{gul} differs for the three recovery rates μ. The critical value increases as μ increases, in other words, as the average time spent in the infected state decreases. This is consistent with what intuition suggests: for the epidemic to be sustained when μ is larger, individuals must be more susceptible in order to compensate for the faster average return to the susceptible state. In our integrated model, this increase in susceptibility is associated with an increase in P_{gul}. In the second column of Figure 5.8 is shown the result of the lifetime computation. The lifetime is defined as the time the virus needs either to die out or to reach a fraction Y of the population. The lifetime behaves like a second-order phase-transition susceptibility and enables a numerical estimation of the SIS threshold; the peak of the lifetime corresponds to the critical value of P_{gul}. We observe good agreement between the critical values obtained from I_∞ and the lifetime.

From these results, we see that plotting I_∞ as P_{gul} varies leads to results with the same qualitative behavior to those of the SIS model with probabilistic contagion, as in Figure 3.4, where I_∞ is plotted as a function of R_0. We therefore find that also in the integrated model with the IBLM the system exhibits a threshold behavior. Since we obtain a similar behavior to the model with probabilistic contagion as P_{gul} varies, then P_{gul} must be related to the threshold.

The main idea is that the infection probability depends on individuals' susceptibility to phishing emails. In our implementation of the IBLM, this susceptibility is controlled by the parameter P_{gul}; it directly affects an individual's initial experience and, consequently, their propensity to fall victim to a cyber threat. This motivates linking the infection probability λ of the probabilistic SIS model to P_{gul}. We also expect λ to increase with P_{gul}, since the probability of contagion should increase as an individual's expertise decreases (i.e. as P_{gul} increases).

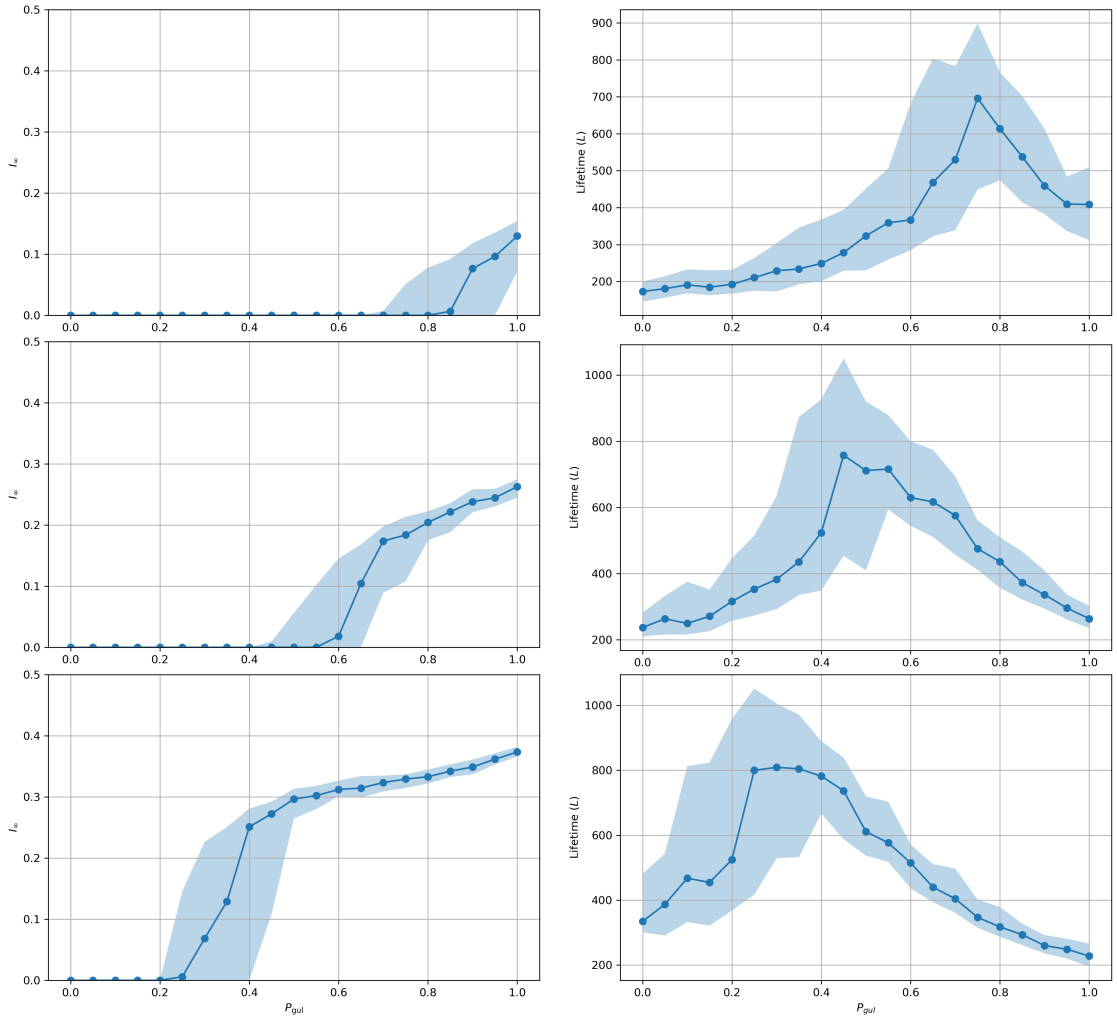


Figure 5.8 In this plot we observe the following: in the first column, the SIS stationary infected density as a function of P_{gul}, and in the second column, the lifetime as a function of P_{gul}. The first row shows the results for $\mu = 1/40$, the second for $\mu = 1/50$, and the third for $\mu = 1/60$. All nodes in the network have the same P_{gul}. In these plots: $N = 10^4$, $\gamma = 1.7$, $\varepsilon = 10^{-3}$, $Q = 1$, $M = 20$, $C_{\max} = 10$. Each point is the average of 100 realizations. The shaded blue area represents the 50% confidence interval.

We would like then to find a relation between λ and P_{gul}, a function like $\lambda = \lambda(P_{\text{gul}})$. To do so, we recall that both $P_{\text{gul}} \in [0, 1]$ and $\lambda \in [0, 1]$. We first asked what values this function should take in two limit cases:

- When $P_{\text{gul}} = 1$, what $\lambda\,(P_{\text{gul}} = 1) \doteq \lambda_1$ is equal to?
  In this case the memory of the individuals is fully randomly initialized; the probability of the individuals to get infected is then on average $\lambda_1 = 0.5$.

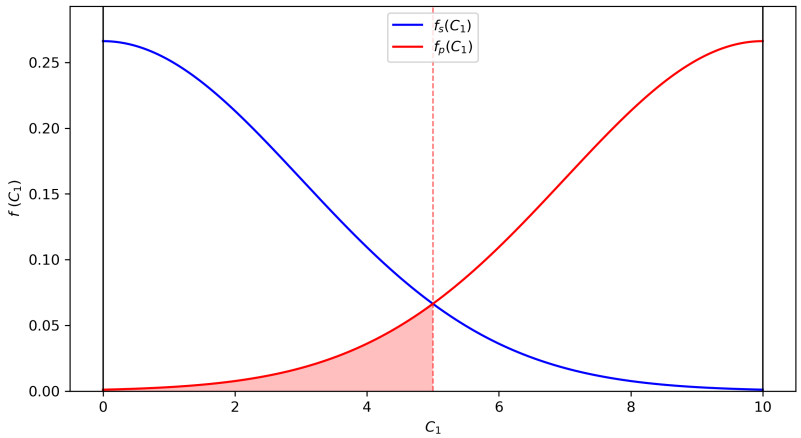


Figure 5.9 In this plot we visualize the probability $\lambda\,(P_{\text{gul}} = 0)$; that is the probability that the feature of the received phishing email is between 0 and $C_{\max}/2$ ($C_{\max} = 10$), as $M \to \infty$.

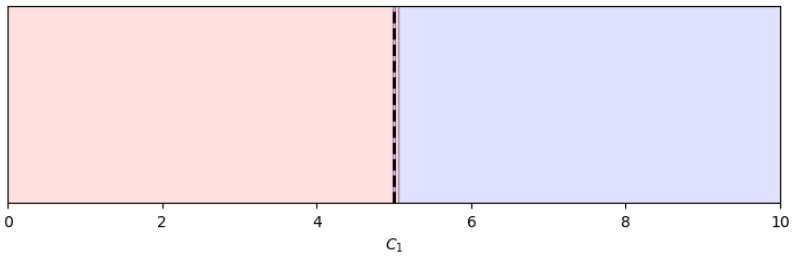


Figure 5.10 We observe in this plot the predicted individual's action for large M (in this case $M = 10000$ instances). The red region corresponds to "click" action, while blue region corresponds to "no click"; the dashed line corresponds to $C_1 = C_{\max}/2$, with $C_{\max} = 10$. We observe that the dashed line well separates the two regions.

- When $P_{\text{gul}} = 0$, what $\lambda\,(P_{\text{gul}} = 0) \doteq \lambda_0$ is equal to?
  In this case the memory of the individual is initialized at its best quality. So the individual is an expert and in the limit $M \to \infty$, with M the number of instances in the memory (considering that the probability distribution functions in the generation process of the emails are symmetric with respect to $C_{\max}/2$), the probability of the expert node to get infected may be considered equal to the probability that a phishing email (sampled from the phishing email distribution) has a feature smaller than $C_{\max}/2$ (see Figure 5.9). In fact, in the limit where the individual's memory is large, thanks to the symmetry of $f_s(C_1)$ and $f_p(C_1)$ with respect to $C_{\max}/2$, the individual's action, on average, depends only on their position relative to $C_{\max}/2$, as can be seen in Figure 5.10. We therefore compute such probability as:

$$\lambda\left(P_{\text{gul}}=0\right)=\mathbb{P}\left(C_1\Big|_{\text{phishing}}<C_{\max}/2\right)$$

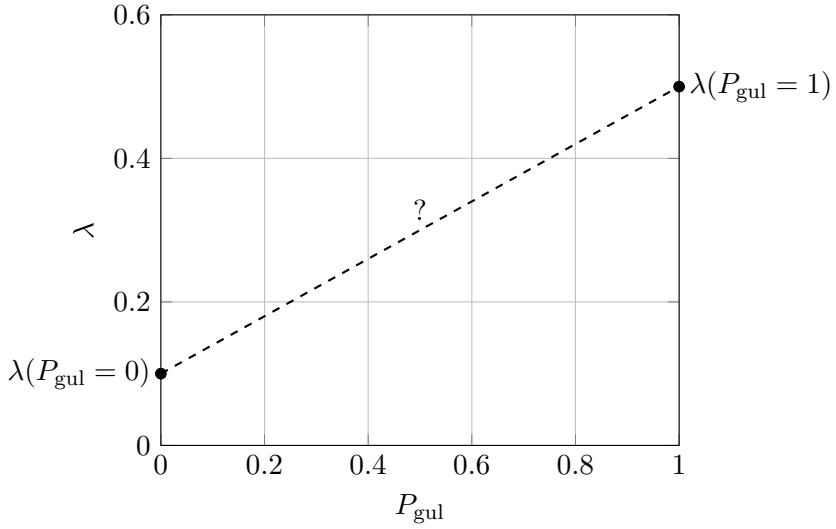
where $C_1\Big|_{\text{phishing}} \in [0, C_{\max}]$ is the feature sampled from $f_{\text{p}}(C_1)$, the probability distribution function for the phishing emails. Then:

$$\lambda\left(P_{\text{gul}}=0\right)=\int_0^{C_{\max}/2} f_p(C_1)\ \mathrm{d}C_1$$

If we set $C_{\max}=10$ and $\sigma=3$, we find that:

$$\lambda\left(P_{\text{gul}}=0\right)=\int_0^{C_{\max}/2} f_p(C_1)\ \mathrm{d}C_1=\lambda_0\cong 0.095$$

We now visualize the two points we have just identified:



Having identified the mapping values at the extremes of the domain, the next question was to determine how λ depends on P_{gul}. For simplicity, we assumed a linear dependence and verified that this was a reasonable assumption. Defining $\Delta_\lambda \doteq \lambda_1-\lambda_0$, we have that the mapping is:

$$\lambda\left(P_{\text{gul}}\right)=\lambda_0+\Delta_\lambda P_{\text{gul}} \tag{5.2}$$

We will now discuss the results of the simulations to determine whether this mapping is reasonable. Looking at Figure 5.11, we can see that the two extreme points of the domain match our assumptions. For $P_{\text{gul}}=1$, the simulated value and the theoretical value essentially coincides. For $P_{\text{gul}}=0$, there is a slight discrepancy between the simulated and theoretical value, although the agreement remains good. This difference is due to the fact

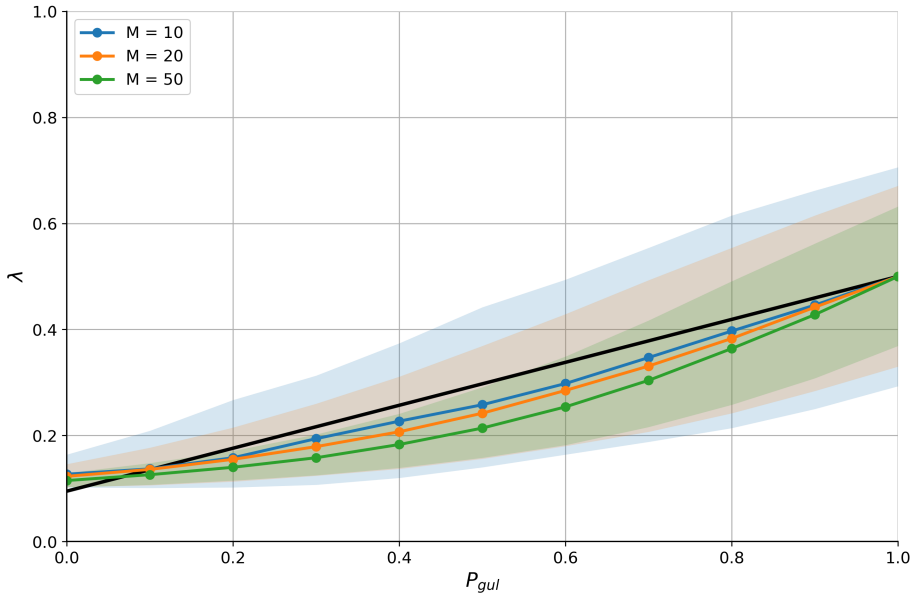


Figure 5.11 In this plot we compare the real dependence of λ on P_{gul} and the mapping. The curves are evaluated for different values of size memory $M = \{10, 20, 50\}$. Each point is the median obtained sending 10^3 phishing emails to $N = 10^5$ isolated individuals, without adding the new emails to the memory; the shaded area corresponds to the 50% confidence interval.

that the theoretical value was derived in the $M \to \infty$ limit, i.e. assuming many instances stored in memory. Indeed, we observe that the value of λ at $P_{\text{gul}} = 0$ approaches the theoretical prediction as M increases.

The largest discrepancy between the simulation outcomes and the mapping occurs for intermediate values of P_{gul}. In this regime, the actual value of λ is lower (and therefore the infection probability is smaller) than the mapping prediction. Individuals modeled with the IBLM therefore appear to be slightly less susceptible than what the linear assumption suggests. Furthermore, susceptibility decreases with increasing memory size M at constant P_{gul}. It should be noted that the plot in Figure 5.11 is evaluated using isolated nodes that receive a certain number of phishing emails, without adding those emails to their memory. This does not reflect what happens in real network dynamics, both because individuals interact with one another (through the network) and because the emails received by the nodes become part of their experience. The actual behavior of the mapping as P_{gul} varies also depends on other parameters, including the recovery rate μ.

Given these considerations, we find a mapping that approximate the simulation behavior, allowing us to reuse analytical results for the SIS model on ADN, in particular the

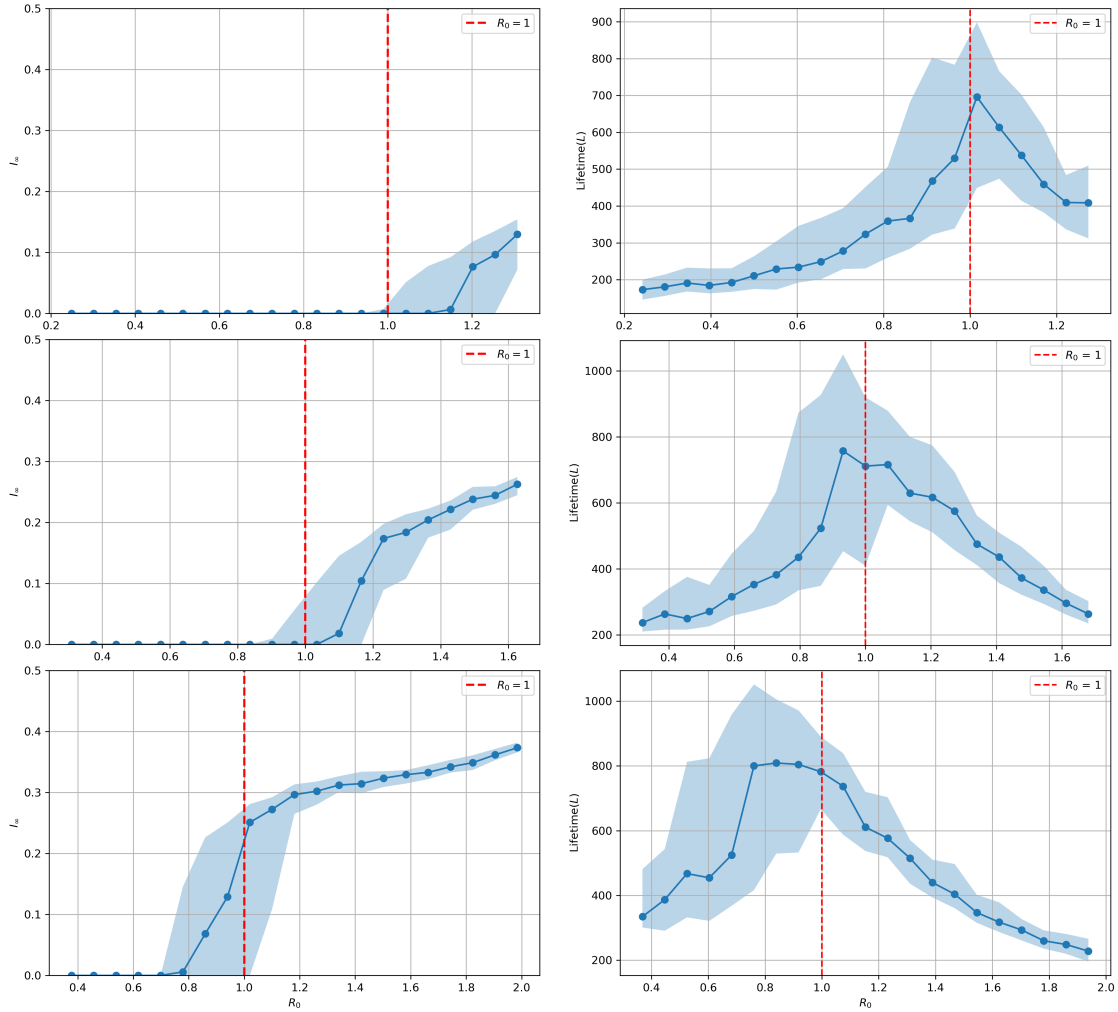


Figure 5.12 In this plot we observe the following: in the first column, the SIS stationary infected density as a function of R_0, and in the second column, the lifetime as a function of R_0. The first row shows the results for $\mu = 1/40$, the second for $\mu = 1/50$, and the third for $\mu = 1/60$. All nodes in the network have the same P_{gul}. In these plots: $N = 10^4$, $\gamma = 1.7$, $\varepsilon = 10^{-3}$, $Q = 1$, $M = 20$, $C_{\max} = 10$. Each point is the average of 100 realizations. The shaded blue area represents the 50% confidence interval.

epidemic thresholds, R_0, as presented in Chapter 3. We recall that:

$$R_0 = \frac{\lambda}{\mu} m \langle a \rangle \qquad \text{and} \qquad R_0 = \frac{p \sum_x \beta_x + \Xi}{\sum_x \mu_x}$$

where the first expression does not consider susceptibility classes, whereas the second does account for it. Now that we have defined a mapping that relate the parameters of the

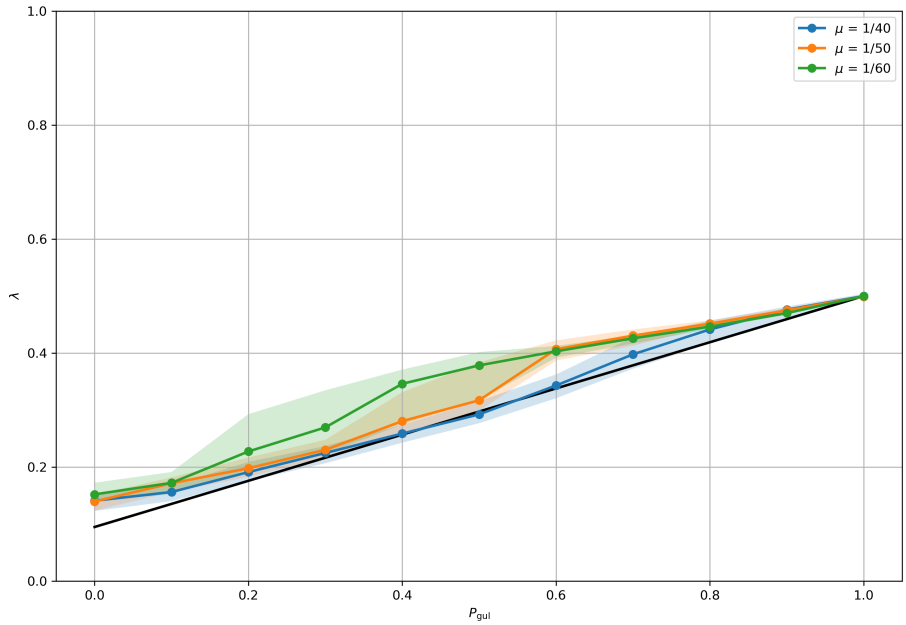


Figure 5.13 In this plot, we show the empirical behavior of $\lambda(P_{\text{gul}})$ obtained from SIS simulations with different values of μ for $M = 20$, the initial memory size. The mapping curve is also shown. In this plot: $N = 10^4$, $\gamma = 1.7$, $\varepsilon = 10^{-3}$, $Q = 1$, $M = 20$, $C_{\max} = 10$. Each point is the median of 50 realizations. The shaded areas represents the 50% confidence interval.

integrated framework to the contagion probability λ, we can determine the thresholds as follows:

$$R_0 = \frac{\lambda\left(P_{\text{gul}}\right)}{\mu}\, m\,\langle a\rangle \tag{5.3}$$

$$R_0 = \frac{p\sum_x \beta_x\left(P_{\text{gul}}\right) + \Xi\left(P_{\text{gul}}\right)}{\sum_x \mu_x} \tag{5.4}$$

with $\beta_x\left(P_{\text{gul}}\right) = m\,\langle a\rangle\,\lambda\left(P_{\text{gul}}\right)$ and $\Xi\left(P_{\text{gul}}\right)$ a function of P_{gul} through $\lambda\left(P_{\text{gul}}\right)$.

We used these thresholds, that exploit the mapping, to validate the thresholds of the integrated model. We observe in Fig. 5.12 that the threshold estimated via R_0, using the mapping, captures the actual epidemic threshold for $\mu = 1/40$ and $\mu = 1/50$. This is not the case for $\mu = 1/60$, where the curve of the infected fraction begins to deviate from zero well before the predicted threshold. To understand the reason of this behavior, we investigated the actual behavior of λ as P_{gul} varies; the results are shown in Figure 5.13. A possible explanation is that, by decreasing μ - and thus increasing the time individuals remain in the infected state - many more malicious emails circulate through the network

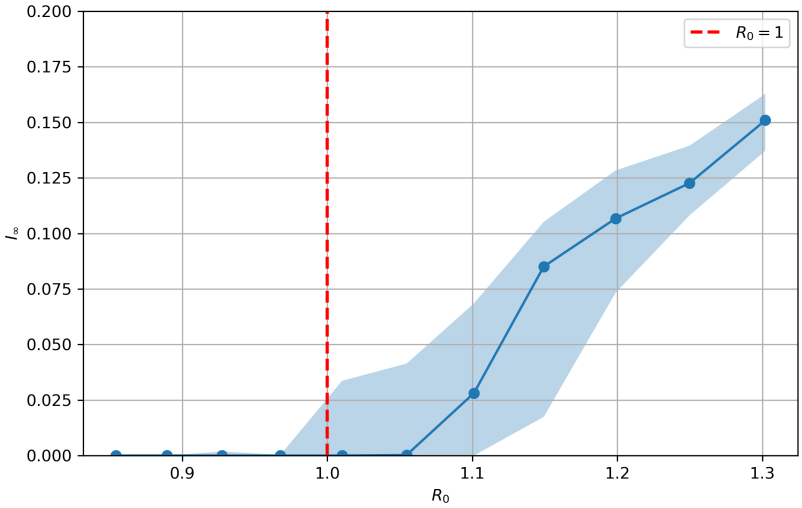


Figure 5.14 In this plot, we show the SIS stationary infected density as a function of R_0 with more than one gullibility class, specifically $Q = 2$, so the class $x \in \{1, 2\}$. Each point is the average over 100 realizations; all nodes in gullibility class $x = 1$ have $P_{\text{gul}} = 1$. In this plot: $N = 10^4$, $\gamma = 1.7$, $\varepsilon = 10^{-3}$, $\mu = 1/40$, $M = 20$, $C_{\max} = 10$, $p = 0.6$. Each point is the average of 100 realizations. The shaded blue area represents the 50% confidence interval.

and end up in the nodes' memory. These emails are incorporated without any feedback mechanism, meaning that a node never learns whether a message was actually safe or not. As a result, for intermediate levels of gullibility, the node's memory becomes less reliable as emails are accumulated; this increases the probability of clicking on malicious emails, making nodes more susceptible. In the end λ is larger. We observe that the differences among the curves (i.e. among the different recovery rates μ) diminish as P_{gul} approaches 0 and 1: for experienced nodes ($P_{\text{gul}} = 0$), the number of phishing emails in the network has little effect, and the same holds for completely inexperienced nodes ($P_{\text{gul}} = 1$). Furthermore, we note that a lower recovery rate μ may be associated with higher phishing susceptibility; a more gullible individual is expected to recover more slowly than a less gullible one. In the absence of a feedback mechanism, having more infected individuals in the network (which depends on μ, i.e. the average time people remain infected) makes high-gullible nodes even easier to infect, increasing λ.

To summarize, we derived a mapping that approximates the actual epidemic threshold R_0. In particular, the mapping used to calculate R_0 is more reliable for larger values of μ, whereas it deviates from the real behavior for lower μ. It is clear that, from these results, the mapping can't capture the effects of different recovery rates μ. Specifically when μ is larger, the fraction of infected nodes in the network increases and the mapping cannot

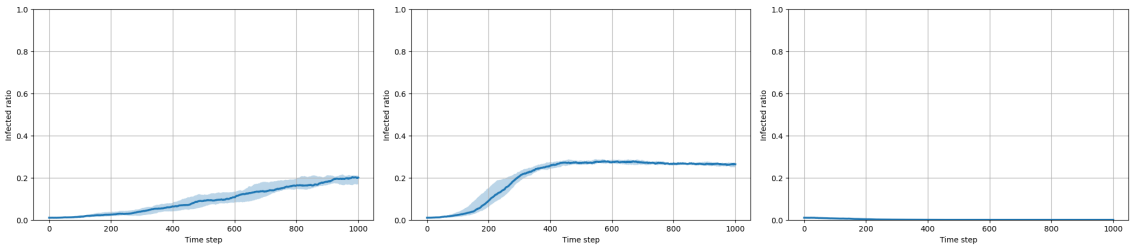


Figure 5.15 In this plot, we show the existence of a correlation between activity and gullibility. The first panel is the baseline, with no correlation. In the second, the most active nodes are those with the highest gullibility ($P_{\text{gul}} = 1$), whereas in the third, the most active nodes are the least gullible ($P_{\text{gul}} = 0$). In this plot: $N = 10^4$, $\gamma = 1.7$, $\varepsilon = 10^{-3}$, $\mu = 1/40$, $Q = 2$, $M = 20$, $C_{\max} = 10$, $p = 0.7$. Each point is the average of 50 realizations. The shaded blue area represents the 50% confidence interval.

capture this effect on nodes gullibility.

Considering the case with the addition of susceptibility classes, we can use the mapping to calculate the threshold, as shown in equation 5.4. Looking at Figure 5.14, we observe that also in this case the mapping is able to roughly capture the critical behavior of the spreading process and its threshold.

We also explored network effects, specifically the correlation between gullibility and activity. We observe in Figure 5.15 how the density of infected nodes in the network behaves in three different scenarios. Specifically, the first case is the baseline, with no correlation between activity and gullibility; the second corresponds to the case in which the most active nodes are also the most gullible; and finally, in the third case the most active nodes are the most experienced and therefore less gullible. We note that in the second case the more gullible nodes allow the contagion dynamics to be sustained, whereas in the third case, since the least gullible nodes are the most active, the cyber threat is not able to spread. We recover, with the integrated model with the IBLM, the result of Brett et al. [17], showing that a correlation between activity and gullibility leads to different responses of the system, and therefore different epidemic thresholds.

5.2.3 Cognitive model with community effect

The goal is to include into the cognitive model an effect driven by the presence of the network and its community structure. As mentioned in the Introduction, emails received from trusted senders (i.e. nodes in the same community) are perceived as safer; consequently, individuals should be more prone to click messages sent by others within their own community. To implement this mechanism within the IBLM framework, we modify the optimization problem to increase the propensity to click the emails when the sender and receiver belong to the same community. To understand how this mechanism was added,

let's begin by making some useful observations.

We recall the IBLM optimization problem:

$$\underset{V\in\{T,F\}}{\operatorname{argmin}}\{S_{M_i}(V)\} = \operatorname{argmin}\{S_{M_i}(T),\ S_{M_i}(F)\} \qquad \text{with } S_{M_i}(V) = \sum_{k=1}^{M_i} P_k \cdot [1 - \mathrm{Sim}(V, V_k)]^2$$

with V the set of the possible actions; in our model $V = \{T, F\}$, with T = "click" and F = "no click". We observe that, since $\mathrm{Sim}(V, V_k) \in [-1, 0]$:

$$1 \leq 1 - \mathrm{Sim}(V, V_k) \leq 2 \quad \Rightarrow \quad 1 \leq [1 - \mathrm{Sim}(V, V_k)]^2 \leq 4$$

Multiplying it by P_k and then summing over all M_i chunks in node's memory:

$$\begin{array}{rcccl} P_k & \leq & P_k \cdot [1 - \mathrm{Sim}(V, V_k)]^2 & \leq & 4P_k \\ \displaystyle\sum_{k=1}^{M_i} P_k & \leq & \displaystyle\sum_{k=1}^{M_i} P_k \cdot [1 - \mathrm{Sim}(V, V_k)]^2 & \leq & \displaystyle 4\sum_{k=1}^{M_i} P_k \\ 1 & \leq & S_{M_i}(V) & \leq & 4 \end{array}$$

So we derived that:

$$1 \leq S(V) \leq 4 \quad \text{with } V \in \{T, F\}$$

where we omitted the M_i to simplify the notation. As already mentioned, our goal is to boost the "click" response, or to penalize the "no click" response, when the sender's community matches the receiver's one. In any case, this addition must remain comparable to $S(V)$. Firstly, we observe the behavior of the median distance between $S(T)$ and $S(F)$ evaluated for the emails sent in one time step. We define the distance between $S(T)$ and $S(F)$ for one received email as:

$$\Delta_S \doteq |S(T) - S(F)|$$

Therefore the median between Δ_S of all the email sent in one time step is:

$$\Delta_t \doteq \operatorname{median}(\Delta_S)\Big|_{\text{emails sent during step } t}$$

We now look at Figure 5.16. We observe that Δ_t increases in time, both for $P_{\text{gul}} = 1$ and $P_{\text{gul}} = 0$. To explain this behavior, we recall that:

$$S(T) = S_{M_i}(T) = \sum_{k=1}^{M_i} P_k \cdot [1 - \mathrm{Sim}(T, V_k)]^2$$

$$S(F) = S_{M_i}(F) = \sum_{k=1}^{M_i} P_k \cdot [1 - \mathrm{Sim}(F, V_k)]^2$$

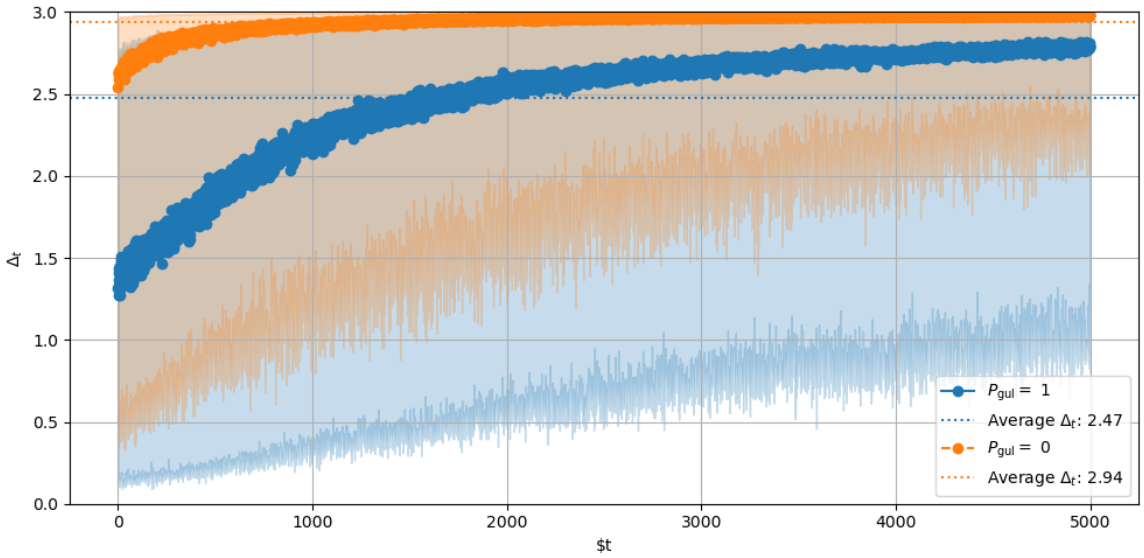


Figure 5.16 In this plot, we show the evolution of Δ_t during the SIS dynamics when all nodes are highly gullible (blue) and when they are weakly gullible (orange). The shaded blue area represents the 90% confidence interval.

where $0 < P_k < 1$ is the probability that measures the relevance of that email in memory with respect to the current context. Considering the contribution of k-th chunk:

$$\text{Contribution to } S(T): \quad P_k \cdot [1 - \text{Sim}(T, V_k)]^2 = \begin{cases} P_k & \text{if } V_k = T \\ 4P_k & \text{if } V_k = F \end{cases}$$

$$\text{Contribution to } S(F): \quad P_k \cdot [1 - \text{Sim}(F, V_k)]^2 = \begin{cases} 4P_k & \text{if } V_k = T \\ P_k & \text{if } V_k = F \end{cases}$$

Therefore, when for example $V_k = T$, the contribution of the k-th chunk is maximum in $S(F)$ and minimum in $S(T)$.

At the beginning, memory is sparse and the stored instances have fairly similar weights in the optimization problem, even though they correspond to different memorized actions. As a consequence, $S(T)$ and $S(F)$ assume similar values. As time progresses, nodes receive many emails that accumulate in memory. Over time, the whole range between 0 and $C_{\max}$ becomes populated. As a result, there is always a small number of instances that closely match the current situation (with similar feature value) and therefore carry significant weight, while an increasingly large number of instances have negligible weight. Since the instances with the highest probability P_k contribute to $S(T)$ and $S(F)$ in opposite ways, the gap between the two increases over time.

For both experienced ($P_{\text{gul}} = 0$) and inexperienced ($P_{\text{gul}} = 1$) nodes, over time, the whole range of possible feature values is explored. This defines the regions of feature space where, according to the node, an email is legitimate or not, regardless of whether this is actually true, as it only reflects the particular knowledge and experience of the node. Without a feedback mechanism, individuals can mistakenly believe that they are learning and think that they are experts.

We furthermore observe that there is a difference at the beginning between Δ_t for $P_{\text{gul}} = 1$ and $P_{\text{gul}} = 0$. In this case, the difference is due to the different memory initialization. For experienced nodes, there is initially a correct correspondence between the email feature value and the associated action; as a result, the terms in the sum $S_{M_i}(V)$ contribute either with the minimum or the maximum value to the presumably "correct" action, T or V. This does not hold for less expert nodes, for which the contributions may be essentially random and can lead to more similar values of $S(T)$ and $S(F)$.

We now introduce the addition in the optimization problem to consider that emails received from nodes in the same community are perceived as safer. At the moment the optimization problem reads:

$$\underset{V\in\{T,F\}}{\operatorname{argmin}} \{S_{M_i}(V)\} \qquad \text{with } S_{M_i}(V) = \sum_{k=1}^{M_i} P_k \cdot [1 - \text{Sim}(V, V_k)]^2$$

The idea is to penalize the action F ("no click") when the community of sender, c_s, is the same as the receiver, c_r. Then the new quantity to be minimized would be:

$$S_{M_i}(V) + \delta_{V,F} \cdot f(c_s, c_r)$$

with $\delta_{V,F}$ the Kronecker delta defined as:

$$\delta_{V,F} = \begin{cases} 1 & \text{if } V = F \\ 0 & \text{if } V \neq F \end{cases}$$

We define:

$$f(c_s, c_r) = \begin{cases} K & \text{if } c_{\text{s}} = c_{\text{r}} \\ 0 & \text{if } c_{\text{s}} \neq c_{\text{r}} \end{cases} = K\delta_{c_{\text{s}},c_{\text{r}}}$$

with $K \in \mathbb{R}_{>0}$ (a positive constant > 0). We define the new quantity to be minimized as:

$$\boxed{\widetilde{S}_{M_i}(V) = S_{M_i}(V) + K\delta_{c_{\text{s}},c_{\text{r}}}\delta_{V,F}}$$

Let us check that it works exactly as intended. For $c_r \neq c_s$ we have that:

$$\begin{cases} \widetilde{S}_{M_i}(T) = S_{M_i}(T) + K\delta_{c_{\text{s}},c_{\text{r}}}\delta_{T,F} = S_{M_i}(T) \\ \widetilde{S}_{M_i}(F) = S_{M_i}(F) + K\delta_{c_{\text{s}},c_{\text{r}}}\delta_{F,F} = S_{M_i}(F) \end{cases} \quad \Rightarrow \quad \boxed{\begin{cases} \widetilde{S}_{M_i}(T) = S_{M_i}(T) \\ \widetilde{S}_{M_i}(F) = S_{M_i}(F) \end{cases} \text{(no change)}}$$

For $c_r = c_s$ we have that:

$$\begin{cases} \widetilde{S}_{M_i}(T) = S_{M_i}(T) + K\delta_{c_s,c_r}\delta_{T,F} = S_{M_i}(T) \\ \widetilde{S}_{M_i}(F) = S_{M_i}(F) + K\delta_{c_s,c_r}\delta_{F,F} = S_{M_i}(F) + K \end{cases} \quad \Rightarrow \quad \boxed{\begin{cases} \widetilde{S}_{M_i}(T) = S_{M_i}(T) \\ \widetilde{S}_{M_i}(F) = S_{M_i}(F) + K \end{cases}}$$

We note that the constant K only appears as a positive addition in the case in which the action is "no click" and $c_s = c_r$, penalizing such action.

Next, we had to determine the value of the constant K. To do so, we consider the case $c_s = c_r$; the individual in the network receives an email. Considering the "old" quantity $S(V)$, if:

1) $S(T) < S(F)$ the individual opens the email
2) $S(F) < S(T)$ the individual does NOT open the email

Let us investigate how $\widetilde{S}(V)$ behaves in these two cases.

1) In this case, we have that:

$$S(T) < S(F) \quad \Leftrightarrow \quad S(T) < S(F) + K \quad \Rightarrow \quad \widetilde{S}(T) < \widetilde{S}(F)$$

since K is a positive constant. Therefore

$$S(T) < S(F) \quad \Rightarrow \quad \widetilde{S}(T) < \widetilde{S}(F)$$

So, if the individual decides to click on the email with the old quantity $S_{M_i}(V)$, he continues to open it even with the new $\widetilde{S}_{M_i}(V)$.

2) In this case, we would like K to be large enough to reverse the ordering $S(F) < S(T)$, obtaining $\widetilde{S}(T) < \widetilde{S}(F)$. The condition to obtain it is:

$$\widetilde{S}(T) - \widetilde{S}(F) = S(T) - S(F) - K < 0 \quad \Rightarrow \quad \boxed{K > S(T) - S(F)}$$

The first observation is that, given:

$$S(T) - S(F) > 0 \quad \text{and} \quad -3 \leq S(T) - S(F) \leq 3 \quad \Rightarrow \quad 0 < S(T) - S(F) \leq 3$$

then:

$$\widetilde{S}(T) - \widetilde{S}(F) = \underbrace{S(T) - S(F)}_{\leq 3} - K \leq 3 - K$$

The individual opens the email if $\widetilde{S}(T) - \widetilde{S}(F) < 0$, that means:

$$3 - K < 0 \quad \Rightarrow \quad K > 3 \doteq K_{\text{upper}}$$

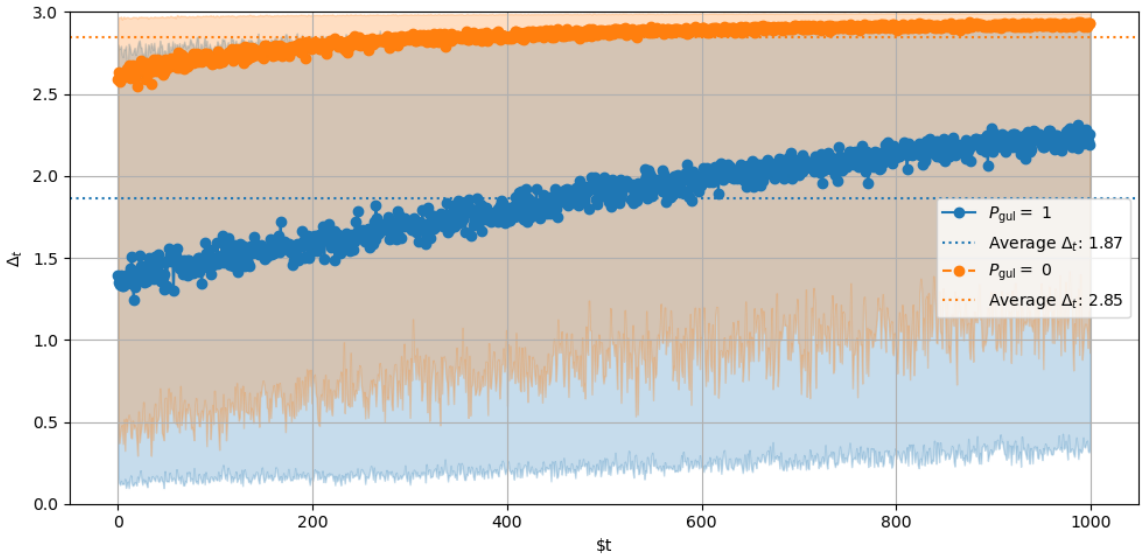


Figure 5.17 In this plot, we show the evolution of Δ_t during the SIS dynamics when all nodes are highly gullible (blue) and when they are weakly gullible (orange). We run until 1000 time steps. The shaded blue area represents the 90% confidence interval.

For $K > K_{\text{upper}}$, it is guaranteed that if an individual receives an email from a member of its own community, the email will be clicked.

The second observation is that, in the extreme case where the difference $S(T) - S(F)$ assumes its maximum value ($S(T) - S(F) = 3$), the parameter K should not be so large as to reverse the preference between the two actions. In this situation, the individual has accumulated very strong evidence in favor of F action ("no click"), which can be interpreted as a high level of experience: the fact that an email comes from a trusted sender is not sufficient to click if the email's features appear typical of phishing. As mentioned in the Introduction, it has in fact been shown that even when an email is received from a sender perceived as safe, the cyber attack does not always succeed. Then:

$$K < K_{\text{upper}}$$

In this way, we obtained an upper bound for K; so $0 < K < K_{\text{upper}}$.

It remains to determine the precise value of this constant. We recall that the condition on K is:

$$K > S(T) - S(F)$$

If this condition is respected for every received email by the individuals in the network, then every node will open the email from a trusted sender. We want only a portion of individuals to rely exclusively on trust in the sender, and this fraction depends on the choice

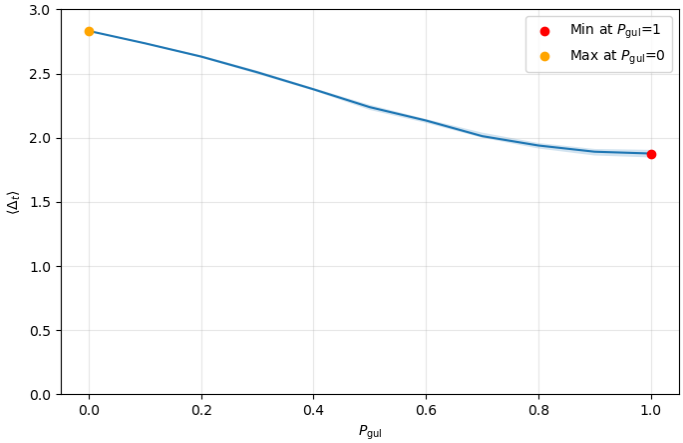


Figure 5.18 In this plot, we show how $\langle \Delta_t \rangle$ varies with P_{gul}, highlighting the minimum and maximum values, found at $P_{\text{gul}} = 1$ and $P_{\text{gul}} = 0$, respectively.

of K. One possibility is to set it equal to the average of Δ_t over the entire simulation, so that, on average, about half of the individuals would click on the email simply because it comes from the same community (see Figure 5.17).

We notice that the average of Δ_t depends on P_{gul} (see Figure 5.18). It is reasonable to expect that choosing K between its minimum ($\langle \Delta_t \rangle$ in $P_{\text{gul}} = 1$) and maximum ($\langle \Delta_t \rangle$ in $P_{\text{gul}} = 0$) values will produce heterogeneous behavior: more gullible nodes will tend to rely on trust in the sender (belonging to the same community) rather than on their past experience, whereas smarter nodes will continue to base their decisions primarily on accumulated evidence, and thus on experience.

We investigated the effects of this addition by evaluating the SIS stationary density of infected nodes, I_∞, for different gullibility values and as a function of K, while keeping all other parameters fixed. We begin by looking at Figure 5.19, where we observe that, in general, the value of I_∞ is higher when the community effect is taken into account. This means that including this effect in the optimization problem allows the cyber threat to spread more easily.

Looking at Figure 5.20, we observe how I_∞ varies for different values of K. It is interesting to observe that for $K = 0$ (i.e. the case without the community effect) the contagion dynamics would not be able to sustain itself, whereas for $K = 1, 2, 3$ we obtain values of I_∞ different from zero, whose value increases with K. Since K modulates the strength of the community effect - and thus of trust - this suggests that trust in a group of individuals (the community), perceived as safer, can make the system more vulnerable, leading to a higher infected population.

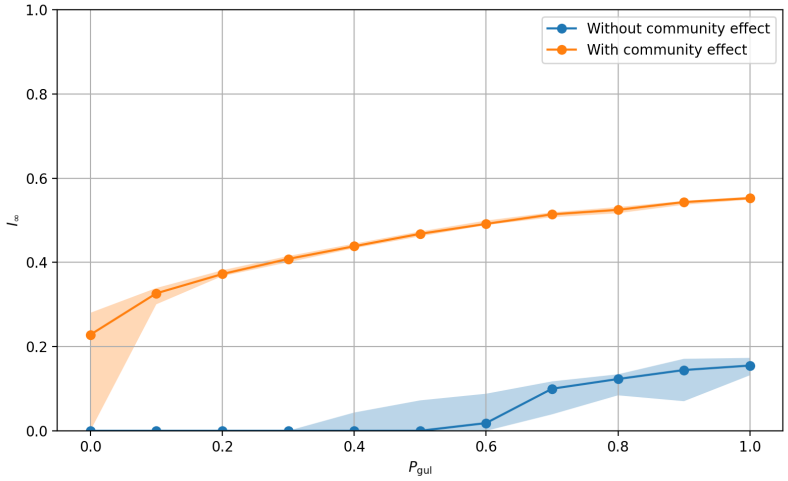


Figure 5.19 In this plot we show how the SIS stationary infected density I_∞ varies with P_{gul}, with the community effect (orange) and without it (blue). In this plot: $N = 10^4$, $\gamma = 1.7$, $\varepsilon = 10^{-3}$, $\mu = 1/50$, $Q = 1$, $M = 20$, $C_{\max} = 10$, $p_c = 0.75$, $C = 500$, $K = 2$. Each point is the average of 20 realizations.

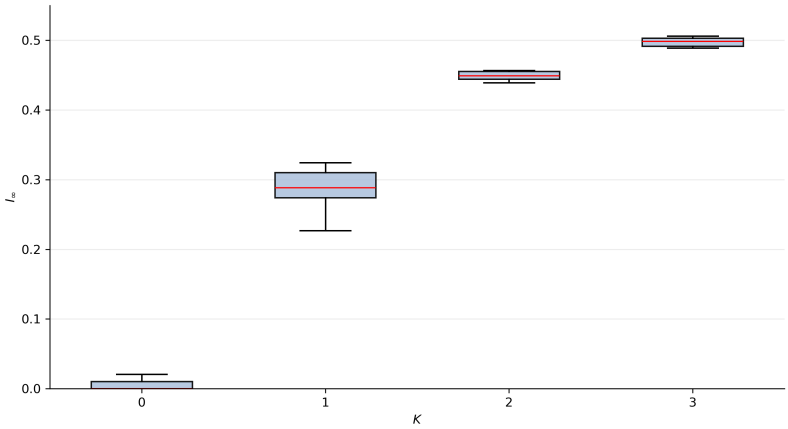


Figure 5.20 In this plot we show the value of the SIS stationary infected density I_∞ for the following values of $K = 0, 1, 2, 3$. We note that the case $K = 0$ corresponds to the case without the community effect.

We note that K is a parameter and should be estimated from real data, if available.

To summarize, in this section we introduced what we have called the community effect. Adding this effect to our integrated framework is supported by evidence in the literature

that users' trust in others within online relationships can be exploited by cyber threats, in particular by social engineering attacks, as a vulnerability, since users are much more likely to fall victim to phishing when contacted by someone who seems familiar. We therefore modified the cognitive model to account for this effect by introducing a parameter K that controls the strength of trust in individuals' decision-making processes. We observe that, for $K > 0$, the fraction of infected individuals in the network is larger than in the $K = 0$ case (i.e. the case without the community effect). This shows that trust has a strong impact on the spread of cyber threats and can actually be exploited to facilitate their propagation. Moreover, we observe that the larger the value of K, and so the strength of trust, the larger the fraction of infected individuals in the network.

Chapter 6

Conclusion

In this thesis, we develop a coherent framework that integrates the dynamics of online social networks, through which cyber threats spread, with individuals' decision-making processes mediated by cognitive mechanisms, as described in cognitive science.

In order to do so, we combine three main components: i) a time-varying network, in particular an Activity Driven Network (ADN) with non-homogeneous susceptibility, expressed in terms of gullibility and time to recover, on which cyber threats spread; ii) SIS dynamics; and iii) an Instance-Based Learning Model (IBLM). The ADN is able to reproduce the heterogeneity of social contacts in a time-varying network, showing how this heterogeneity affects contagion dynamics. The cognitive model is based on Instance-Based Learning Theory (IBLT), which defines a set of cognitive mechanisms that mediate decision-making processes. It is based on evidence that experience, knowledge, and the cognitive biases derived from them play a fundamental role in decision-making in dynamic environments. We therefore define how messages are generated and exchanged within the network, and translate the concept of gullibility into the quality of the memory assigned to each node in the network. Each node is in fact modeled as an autonomous IBLM. The memory of each node consists of a finite set of past instances, each containing the characteristics of a received message and the action taken in response to it. This experience is then used to make decisions about newly received messages, together with the perceived trust in the sender. If a node clicks on a message containing phishing content, it becomes infected, and the malware sends malicious messages to the nodes contacted by that node. In this way, the cyber threat propagates through the network.

The main contribution of this thesis is the development of an integrated framework that bridges models of online social networks, spreading dynamics, and cognitive processes. More specifically, the results can be divided into three main areas. The first concerns the implementation of the IBLM. We show that our implementation is able to distinguish between legitimate and phishing emails. We also validate the model using the results of the Phishing Training Task (PTT) experiment, showing that it is able to qualitatively reproduce not only

the results of more complex IBLMs, but, more importantly, human behavior. The second concethe mapping results. By studying the fraction of infected nodes in the stationary state, we find that threshold behavior is also present with the IBLM. We derive a mapping that reproduces the epidemic threshold quite well for large values of the recovery rate. Finally, the third deals with the results on the community effect. We introduce into the cognitive model an effect associated with network topology; nodes belonging to the same community are perceived as more reliable. Through numerical simulations, we find that trust within communities is a factor that contributes to increasing the number of infected nodes in the network. We also find that the more individuals rely on trust, the more vulnerable the network becomes.

More generally, the proposed framework suggests that susceptibility should be modeled as a dynamic property, defining a model for studying the spread of cyber threats in which susceptibility arises from decisions based on past experience and mediated by cognitive mechanisms, possibly including network effects in users' susceptibility characterization. The reliability of the IBLM implementation, the connection established through the mapping, and the flexibility of the model make this approach a solid basis for future developments. In particular, the framework can be extended to include additional network effects, such as the community effect, providing a more realistic description. By adding a cognitive model, this framework makes it possible to include new dynamics in the description of cyber threats spread, with a potential impact on network vulnerability. The versatility of the model therefore opens the way to more realistic representations of cyber-threat spread.

At the same time, our work has also revealed a number of limitations and critical issues. First, the use of artificial emails rather than real ones may reduce the realism of the spreading dynamics, resulting in an oversimplified model. Second, although the mapping provides a link to the probabilistic model, it still presents some limitations: it performs better for large values of the recovery rate, suggesting that a better understanding of the role of the recovery rate as gullibility varies is still needed. Third, and perhaps most importantly, the absence of a feedback mechanism could influence the dynamics of contagion. Indeed, according to IBLT, learning is a fundamental component of cognitive processes. Finally, computational cost may represent a critical issue. Since each node is modeled as an autonomous IBLM, the model is computationally demanding, which must be compared with the actual improvement in the representation of the spreading phenomenon.

Given these limitations, several future research directions emerge. The most promising is certainly the introduction of a feedback mechanism, which could justify the use of a more computationally demanding model. In addition, one could consider adding further network effects that directly influence user susceptibility, and therefore network vulnerability, improving the applicability of modeling findings to realistic scenarios. Finally, the realism of the model could be improved by using real emails or by refining the way emails are modeled.

Appendix A

Inverse transform sampling

Consider a random variable X with a continuous probability density function $P(x)$, defined for $x \in \mathbb{R}$. We also assume that we know its cumulative distribution function:

$$P_<(x) = P(X < x) = \int_{-\infty}^{x} P(y)\, dy$$

Our goal is to generate random samples according to $P(x)$ using a uniform random generator. Let $u \sim \text{Unif}(0, 1)$ be a random variable uniformly distributed on the interval $[0, 1]$, that is:

$$\text{Unif}(u) = \begin{cases} 1 & \text{if } 0 \leq u \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

We introduce the inverse of the cumulative function, denoted by $P_<^{-1}(x)$, such that:

$$P_<^{-1}(P_<(x)) = x$$

We now define a new variable:

$$x' = P_<^{-1}(u)$$

where u is a uniform random variable. We want to determine the distribution of x'. To do so, we compute its cumulative distribution function:

$$P(x' < x) = P\big(P_<^{-1}(u) \leq x\big)$$

By applying the inverse function property of $P_<$, we can rewrite this as:

$$P\big(P_<^{-1}(u) \leq x\big) = P\big(u \leq P_<(x)\big)$$

Since u is uniformly distributed over $[0, 1]$, we have:

$$P(x' < x) = P(u \leq P_<(x)) = \int_0^{P_<(x)} 1\, du' = P_<(x)$$

Thus, we find that the cumulative distribution of x' is exactly $P_<(x)$. Consequently,

$$x' \sim P(x)$$

We have shown that if the inverse cumulative distribution function $P_<^{-1}(u)$ is known, then by drawing u from a uniform distribution on $[0, 1]$, the transformed variable

$$x' = P_<^{-1}(u)$$

is distributed according to the target probability distribution function $P(x)$.

Bibliography

[1] European Commission. *A cybersecure digital transformation in a complex threat environment. Brochure*. Last update: 2023-01-31. European Commission. June 5, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/cybersecure-digital-transformation-complex-threat-environment-brochure (visited on 02/25/2026).

[2] David Braue. *Cybercrime To Cost The World $12.2 Trillion Annually By 2031*. Cybersecurity Ventures. May 28, 2025. URL: https://cybersecurityventures.com/official-cybercrime-report-2025/ (visited on 02/25/2026).

[3] Antoine Bouveret. *Cyber Risk for the Financial Sector: A Framework for Quantitative Assessment*. IMF Working Paper. International Monetary Fund, 2018. URL: https://www.imf.org/en/Publications/WP/Issues/2018/06/22/Cyber-Risk-for-the-Financial-Sector-A-Framework-for-Quantitative-Assessment-45924 (visited on 02/25/2026).

[4] MarketsandMarkets Research Private Ltd. *Cybersecurity Market by Solution (IAM, Firewall & VPN, Log Management & IEM, Antivirus/Antimalware), Service (Professional (Risk & Threat Assessment, Training & Education), Managed), Security Type (Cloud, Application) - Global Forecast to 2030*. 2025. URL: https://www.marketsandmarkets.com/Market-Reports/cyber-security-market-505.html (visited on 02/25/2026).

[5] Symantec. *Internet Security Threat Report*. Symantec, 2018. URL: https://www.symantec.com/security-center/threat-report (visited on 02/25/2026).

[6] Imrul Kayes and Adriana Iamnitchi. "Privacy and security in online social networks: A survey". In: *Online Social Networks and Media* 3 (2017), pp. 1–21.

[7] Ryan Heartfield and George Loukas. "Protection against semantic social engineering attacks". In: *Versatile cybersecurity*. Springer, 2018, pp. 99–140.

[8] Ryan Heartfield and George Loukas. "Detecting semantic social engineering attacks with the weakest link: Implementation and empirical evaluation of a human-as-a-security-sensor framework". In: *Computers & Security* 76 (2018), pp. 101–127.

[9] Brij B Gupta et al. "Fighting against phishing attacks: state of the art and future challenges". In: *Neural Computing and Applications* 28.12 (2017), pp. 3629–3654.

[10] Justin Balthrop et al. “Technological networks and the spread of computer viruses”. In: *Science* 304.5670 (2004), pp. 527–529.

[11] Romualdo Pastor-Satorras and Alessandro Vespignani. “Epidemic spreading in scale-free networks”. In: *Physical review letters* 86.14 (2001), p. 3200.

[12] Petter Holme and Jari Saramäki. “Temporal networks”. In: *Physics reports* 519.3 (2012), pp. 97–125.

[13] Petter Holme. “Modern temporal network theory: a colloquium”. In: *The European Physical Journal B* 88.9 (2015), p. 234.

[14] Suyu Liu et al. “Controlling contagion processes in activity driven networks”. In: *Physical review letters* 112.11 (2014), p. 118702.

[15] Edward A Cranford et al. “Modeling cognitive dynamics in end-user response to phishing emails”. In: *Proceedings of the 17th ICCM* (2019).

[16] Edward A Cranford et al. “Towards a cognitive theory of cyber deception”. In: *Cognitive Science* 45.7 (2021), e13013.

[17] Terry Brett et al. “Spreading of computer viruses on time-varying networks”. In: *Physical Review E* 99.5 (2019), p. 050303.

[18] Nicola Perra et al. “Activity driven modeling of time varying networks”. In: *Scientific reports* 2.1 (2012), p. 469.

[19] Steve Sheng et al. “Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 373–382.

[20] Alun L Lloyd and Robert M May. “How viruses spread among computers and people”. In: *Science* 292.5520 (2001), pp. 1316–1317.

[21] Alesia Chernikova et al. “Cyber network resilience against self-propagating malware attacks”. In: *European Symposium on Research in Computer Security*. Springer. 2022, pp. 531–550.

[22] Alesia Chernikova et al. “Modeling self-propagating malware with epidemiological models”. In: *Applied Network Science* 8.1 (2023), p. 52.

[23] John C Wierman and David J Marchette. “Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction”. In: *Computational statistics & data analysis* 45.1 (2004), pp. 3–23.

[24] Wanping Liu and Shouming Zhong. “Web malware spread modelling and optimal control strategies”. In: *Scientific reports* 7.1 (2017), p. 42308.

[25] Ryan Heartfield, George Loukas, and Diane Gan. “You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks”. In: *IEEE access* 4 (2016), pp. 6910–6928.

[26] Michael Workman. “Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security”. In: *Journal of the American society for information science and technology* 59.4 (2008), pp. 662–674.

[27] Ibrahim Mohammed A Alseadoon. “The impact of users’ characteristics on their ability to detect phishing emails”. PhD thesis. Queensland University of Technology, 2014.

[28] Ryan Heartfield and George Loukas. “Predicting the performance of users as human sensors of security threats in social media”. In: *International Journal on Cyber Situational Awareness (IJCSA)* 1.1 (2016).

[29] Tian Lin et al. “Susceptibility to spear-phishing emails: Effects of internet user demographics and email content”. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 26.5 (2019), pp. 1–28.

[30] Rundong Yang et al. “Predicting user susceptibility to phishing based on multidimensional features”. In: *Computational Intelligence and Neuroscience* 2022.1 (2022), p. 7058972.

[31] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. “Instance-based learning in dynamic decision making”. In: *Cognitive Science* 27.4 (2003), pp. 591–635.

[32] Edward A Cranford et al. “Modeling phishing susceptibility as decisions from experience”. In: *Proceedings of the 19th Annual Meeting of the ICCM*. Applied Cognitive Science Lab State College, PA. 2021, pp. 44–49.

[33] Shelby R Curtis et al. “Phishing attempts among the dark triad: Patterns of attack and vulnerability”. In: *Computers in Human Behavior* 87 (2018), pp. 174–182.

[34] Kuldeep Singh et al. “Training to detect phishing emails: Effects of the frequency of experienced phishing emails”. In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 63. 1. SAGE Publications Sage CA: Los Angeles, CA. 2019, pp. 453–457.

[35] George Saridakis et al. “Individual information security, user behaviour and cyber victimisation: An empirical study of social networking users”. In: *Technological Forecasting and Social Change* 102 (2016), pp. 320–330.

[36] Samar Muslah Albladi and George RS Weir. “Predicting individuals’ vulnerability to social engineering in social networks”. In: *Cybersecurity* 3.1 (2020), p. 7.

[37] Tom N Jagatic et al. “Social phishing”. In: *Communications of the ACM* 50.10 (2007), pp. 94–100.

[38] Gourab Ghoshal and Petter Holme. “Attractiveness and activity in internet communities”. In: *Physica A: Statistical Mechanics and its Applications* 364 (2006), pp. 603–609.

[39] Réka Albert, Hawoong Jeong, and Albert-László Barabási. “Diameter of the world-wide web”. In: *nature* 401.6749 (1999), pp. 130–131.

[40] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2004.

[41] Albert-Laszlo Barabasi. "The origin of bursts and heavy tails in human dynamics". In: *Nature* 435.7039 (2005), pp. 207–211.

[42] Hang-Hyun Jo et al. "Circadian pattern and burstiness in human communication activity". In: *New J Phys* 14.1 (2012), p. 013055.

[43] APS. *Data sets for research*. 2010.

[44] IMDb. *Internet movie database*. 2010. URL: `http://www.imdb.com/interfaces` (visited on 11/10/2011).

[45] Matthieu Nadini et al. "Epidemic spreading in modular time-varying networks". In: *Scientific reports* 8.1 (2018), p. 2352.

[46] William Heaton Hamer. *The milroy lectures on epidemic diseases in england: The evidence of variability and of persistency of type; delivered before the royal college of physicians of london, march 1st, 6th, and 8th, 1906*. Bedford Press, 1906.

[47] Ronald Ross. "An application of the theory of probabilities to the study of a priori pathometry.—Part I". In: *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 92.638 (1916), pp. 204–230.

[48] William Ogilvy Kermack and Anderson G McKendrick. "A contribution to the mathematical theory of epidemics". In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pp. 700–721.

[49] HA Simon and P Langley. "The central role of learning in cognition". In: *Models of thought* 2 (1981), pp. 102–184.

[50] William G Chase and Herbert A Simon. "The mind's eye in chess". In: *Visual information processing*. Elsevier, 1973, pp. 215–281.

[51] Douglas L Hintzman. "MINERVA 2: A simulation model of human memory". In: *Behavior Research Methods, Instruments, & Computers* 16.2 (1984), pp. 96–101.

[52] Douglas L Hintzman. "" Schema abstraction" in a multiple-trace memory model." In: *Psychological review* 93.4 (1986), p. 411.

[53] Douglas L Medin and Marguerite M Schaffer. "Context theory of classification learning." In: *Psychological review* 85.3 (1978), p. 207.

[54] Robert M Nosofsky. "Choice, similarity, and the context theory of classification." In: *Journal of Experimental Psychology: Learning, memory, and cognition* 10.1 (1984), p. 104.

[55] Gordon D Logan. "Toward an instance theory of automatization." In: *Psychological review* 95.4 (1988), p. 492.

[56] Robert M Nosofsky and Thomas J Palmeri. "An exemplar-based random walk model of speeded classification." In: *Psychological review* 104.2 (1997), p. 266.

[57] Itzhak Gilboa and David Schmeidler. "Case-based decision theory". In: *The quarterly Journal of economics* 110.3 (1995), pp. 605–639.

[58] John R Anderson and Christian J Lebiere. *The atomic components of thought*. Psychology Press, 2014.