Efficient Collective Action for Tackling Time-Critical Cybersecurity Threats

Gillard Sébastien[1,2]; Percia David Dimitri[1,3];
Mermoud Alain[3]; Maillart Thomas[1,4]

[1] Information Science Institute, Geneva School of Economics and Management, University of Geneva
[2] Department of Defense Economics, Military Academy at ETH Zurich
[3] Cyber-Defence Campus, armasuisse Science and Technology
[4] Citizen Cyber Lab, University of Geneva

Accepted for presentation at the 22nd Workshop on the Economics of Information Security, Tulsa, 2022

Abstract

The latency reduction between the discovery of vulnerabilities, the build-up and dissemination of cyber-attacks has put significant pressure on cybersecurity professionals. For that, security researchers have increasingly resorted to collective action in order to reduce the time needed to characterize and tame outstanding threats. Here, we investigate how joining and contributions dynamics on MISP, an open source threat intelligence sharing platform, influence the time needed to collectively complete threat descriptions. We find that performance, defined as the capacity to characterize quickly a threat event, is influenced by (i) its own complexity (negatively), by (ii) collective action (positively), and by (iii) learning, information integration and modularity (positively). Our results inform on how collective action can be organized at scale and in a modular way to overcome a large number of time-critical tasks, such as cybersecurity threats.

Keywords— cybersecurity, information sharing, collective action, information integration, economies of scales, Malware Information Sharing Platform (MISP)

1 Introduction

From Computer Emergency Readiness Teams (CERT) established in the nineties [1], to information-sharing analysis centers (ISACs) [2], to bug bounty programs [3, 4], collective action has long been used and recognized as key for the gathering, the integration and the sharing of critical cybersecurity information [5, 6]. The reason for resorting to information-sharing as a form of collective action stems from the complexity associated with the continuous and somewhat decentralized (e.g., open source software) adaptation of hardware and software in information systems [7, 8]. Although the Internet has largely developed through an open source spirit [9–11] with significant positive externalities [12, 13], information-sharing has remained difficult when it comes to cybersecurity [6]. The expansion of threats in volume, severity and span has further challenged information infrastructures. Hence, it has forced further cooperation through information-sharing [14]. While their utility has been somewhat confirmed by their wide adoption, there is a dearth of knowledge regarding how these collective action platforms concretely bring performance when addressing cybersecurity threats. For instance, cybersecurity has become increasingly time-critical and demands ever faster reaction time. Determining the chances that a threat will be fully characterized on time for security officers to act upon before attacks actually start has become crucial [15].

Here, we investigate $39,639$ threat events contributed by 485 organizations to a MISP information-sharing platform [14] operated by the Computer Incident Response Center Luxembourg (CIRCL). We specifically study how collective action unravels through information integration and how it brings significant economies of scale in terms of time needed to fully characterize cybersecurity threats (i.e., performance). We resort to a multivariate cross-sectional regression with ordinary least squares method, and we find that (i) the number of organizations engaged in information-sharing, (ii) their acquired experience in the events completion, (iii) the proportion of information integration and (iv) modularity increase performance.

The remainder of this article is organized as follows. Section 2 covers background from the perspectives of social dilemma, productivity and information integration in collective action in general and for cybersecurity. Section 3 introduces MISP and presents the data. In Section 4, we introduce the theoretical framework followed by research hypotheses in Section 5. Section 6 describes the methodological approach. Results are presented in Section 7 and discussed in Section 8 before concluding in Section 9.

2 Background

Knowledge sharing in cybersecurity has been considered as a crucial way to overcome number of vulnerabilities [16] and threats [1]. It is however bound to limiting factors on the one hand, such as social dilemma, as well as enhancing return-on-scale effects on the other hand. Here, we review the literature on (i) social dilemma and productivity of collective action, and on (ii) challenges associated with information integration. We then review the state-of-the-art research in (iii) information sharing for cybersecurity.

2.1 Social Dilemma and Productivity in Collective Action

According to Olson's logic of collective action, small communities are more able to provide collective goods [17]. The central argument is that minor interests will be over-represented and diffuse majority interests trumped, due to a free-rider problem [18]. This free-riding effect is stronger for larger groups [19]. For instance, while Dejean et al. [20] found a positive relation between the size of a community and the amount of collective good provided, they paradoxically also found a decreased propensity by individuals to cooperate

as the size of the community increases. Yet, there is overwhelming evidence that large crowds can be organized in order to establish successful online collective action. Examples include peer-to-peer networks [20, 21], Wikipedia [22], Stack Overflow [23], communities of open source software developers [24, 25]. The Dejean et al. paradox may at least partially resolved by considering that (i) the distribution of effort is highly skewed, with few contributors providing most effort, and (ii) the dynamics of contribution are highly non-linear [26–28]. Taken together, these phenomena are associated with positive return-on-scale of production [26], which may be hindered by coordination costs [29]. Super-linear productivity has been debated at length in organization and management sciences. Investigations of how the number of members, temporal dynamics of events generated can influence positively outputs in way that is greater than the sum of the outputs related to each element of the system (i.e., exhibiting super-linear growth patterns). Research has successfully delivered hints to improve the performance of organization [30–33] by fine-tuning complementary mechanisms within the organization [34], which also foster innovation [35].

2.2 Information Integration and Modularity

One key aspect of generating return-on-scale in knowledge production is information integration. The management of information resources has become central to organizations [36], so that knowledge appears as an utmost strategic resource [37]. For instance, there is growing evidence in science that greater teams create more impacting knowledge [38]. If knowledge is so important, the fundamental capability of an organization has to be considered as the specialized knowledge of each organization member. Its integration shall provide a competitive advantage [37, 39]. With the emergence of virtual exchanges, firms are increasingly seen as distributed knowledge systems [40]. Yet, new interaction methods present various new constraints in term of mutual understanding, contextual knowledge or techniques (e.g., memory, connectivity), which lead to asymmetries in information integration.

In this respect, the tremendous development of online collaboration platforms, as tools for governance strategy and knowledge management, highlights the importance of information-sharing [41]. These platforms promote knowledge transfer by generating modular collaborative units [42]. One may consider that individuals, or groups of individuals, composing a subsystem (i) bring added value in their own specific field (differentiation), in order to (ii) produce a complex good by pooling together this added value (integration). Following Arrow & Debreu [43], differentiation and integration have been a focal point in optimizing the structure of organizations [44, 45]. In fact, differentiation considers segments of a system into subsystems. Each subsystem develops a part of a task, while the integration focuses on the interactions between these subsystems in order to accomplish the entire task [39, 46]. Recently, Engel and Malone used the theory of consciousness as information integration [47] to measure information integration computer systems and on collaborative platforms [46].

2.3 Collective Action and Information Integration for Cybersecurity

As early as twenty years ago, the first Computer Emergency Readiness Teams (CERT) and Information Sharing and Analysis Centers (ISACs) have been established as a central resource for sharing information on cybersecurity threats to critical infrastructures [48]. Nowadays, threat intelligence platforms help organizations aggregate, correlate, and analyze threat data from multiple sources in (almost) real-time to support defensive actions [49]. Further, open source solutions have been proposed as a counterweight to cyber-criminals successfully working together [5]. The swift evolution of cyber-threats has forced

organizations and governments to develop new strategies [50] in order to reduce the risks of security breaches [41]. Although information sharing is an interesting way to enhance cybersecurity, it is believed to be thwarted by social dilemma. Without trust, commitment and shared vision between stakeholders, organizations are reluctant to share information due to the fear of disclosure, reputation risk or loss of competitive power [51]. As such, information-sharing can be considered as a marketplace on which transactions occur and knowledge is transferred [52]. However, human beings have a tendency to not optimize organizational goals [53] and in the case of collective action might adopt behaviors that are not conducive to the overall goal of sharing information [6]. As a consequence, cybersecurity professionals share probably less information than desirable, leading to a knowledge asymmetry to the advantage of the attackers [6]. In particular, stakeholders strategically select their contributions to share (i.e., quantity and quality), leading to truncated and imperfect information sharing. Yet, specially crafted forms of cybersecurity information-sharing platforms have developed, such as bug bounty marketplaces. These platforms act as a trusted third-party between security researchers and software editors [3]. Further, in cybersecurity, resource belief, usefulness belief, and reciprocity belief are all positively associated with knowledge absorption, whereas reward belief is not [52]. These empirical results show that functional cybersecurity information-sharing indeed requires to overcome social dilemma and goes beyond simple reward expectations, but foremost requires that information-sharing is efficient in a context that increasingly requires to address time-critical threats.

3 Data

To understand the nuts and bolts of cybersecurity information-sharing, we resort to *MISP Project*,[1] a popular open source platform, which is used e.g., by the North Atlantic Treaty Organization (NATO).[2] MISP stands for *Malware Information Sharing Platform and Threat Sharing*. Although it carries the word malware in its name, MISP is a threat intelligence platform on which people can share, store and collaborate on all sorts of incidents (e.g., COVID-19 MISP community,[3] but primarily cybersecurity threats. These threats (i.e., events) are characterized by indicators of compromise (i.e., attributes), which are contributed by a multitude of organizations.

There are advantages in using MISP as an object of research. First, it is an open source software. This allows to understand in much detail how the platform is designed and works. Second, a number of threat information sharing communities use MISP to share relatively openly their threat intelligence. Here, we use the whole history of a MISP instance maintained by the Computer Incident Response Center Luxembourg (MISP CIRCL), i.e., the Luxembourg CERT.

As of February 8, 2022, the MISP CIRCL instance is a community of $1,908$ organizations (respectively $4,013$ users), which have contributed $39,639$ events, $9,099,685$ attributes and $3,786$ tags since November 10, 2008. Table 1 shows the ten most involved organizations. One can see that the number of events contributed by organizations is highly skewed. Indeed, Figure 1A shows that the complementary cumulative distribution function exhibits a power law $P(X_E > x_E) \sim 1/x_E^{\mu_E}$ with $\mu_e = 0.54(4)$ (c.f., Appendix B for details on the fitting method). One may additionally note that $1,423$, i.e., around 75%, of organizations do not participate in sharing threat information as a collective good with the broad MISP CIRCL community. These organizations may however consume information or share threat information privately within informal sub-groups, which cannot

[1] `https://www.misp-project.org/`

[2] `https://misp.ncirc.nato.int`

[3] `https://covid-19.iglocska.eu`

be observed. Similarly to $P(X_E > x_E)$, the distributions of attributes $P(X_A > x_A)$ and tags $P(X_T > x_T)$ per event, depicted in Figure 2, follow power laws with exponents respectively $\mu_A = 0.64(1)$ (with an upper cut-off around $A_{upper} = 10^5$) and $\mu_T = 2.26(6)$. It is additionally important to consider that only $22,423$ (i.e., around 57%) events have been marked as completed, suggesting that either threat analysis is complicated or that users tend to forget to formally close a large number of events. The cumulative number of tags $N_{T,\text{cum}} = 116,407$ used is bigger than the unique tags amount $N_{T_U} = 3,786$. Thus, there is a massive reuse of already existing tags.

rank	org ID	# users	# events contributed	percentage of total events
1	1092	8	7,682	19.38%
2	1395	2	5,637	14.22%
3	1960	3	3,214	8.11%
4	2	31	2,939	7.41%
5	1857	3	1,411	3.56%
6	201	8	1,247	3.15%
7	1713	1	1,141	2.88%
8	698	2	1,077	2.72%
9	204	56	1,060	2.67%
10	643	12	998	2.52%
		Total	26,406	66.62%

Table 1: 10 of $1,908$ organizations have contributed 66.62% of the $39,639$ events, bringing further evidence of the heavy-tailed nature of the distribution of contributions by organizations in MISP CIRCL.

We further observe that organizations have joined MISP CIRCL following an almost perfect linear relation $N_O(t) \sim \alpha_O \cdot t$ with $\alpha_O = 0.79(1)$ ($R^2 = 0.99$ and $p < 10^{-2}$) with 161 organizations initially joining MISP CIRCL instance on September 14, 2015, the presumed date of official start. Figure 1B, not only shows the almost linear organization joining rate, but also how many events each organization has contributed over time. One see that the contribution effort is highly heterogeneous. It is also worth noting that event contributions started on November 10, 2008, long before the first organizations joined MISP CIRCL instance. This can be explained in the following way: organizations run first their MISP instance locally. At some point, they join the MISP CIRCL community and share at once all their non-private threat intelligence, yet with the nominal event timestamp, which may well be in the past. Also, it is likely that the linear organization joining function may be the result of a highly vetted joining process, controlled by CIRCL.

3.1 Reduction of the Completion Time of Events Δt_C

Following the method described in the Appendix B, we can treat the data and, from them, generate the Figure 3B. As explained in the appendix, by playing with the axis, we remark that when the axes are in linear-logarithmic scale, the data depict two straight lines. From this observation, we can deduce that $\Delta t_C(t)$ follows an exponential decrease in phase. By applying a binning by month and computing the mean value $\overline{\Delta t_C}$ for each bin, we see a first phase that extends from 2011 to 2020 which decrease slower than the second phase from 2020 to today. By applying the linear regression on the data, according to the equation (9), we confirm that Δt_C exhibits an exponential decrease:

$$\Delta t_C(t) = \begin{cases} \sim 10^{\beta^1_\Delta \cdot t}, \text{ for } t \in [2011, 2020[, \\ \sim 10^{\beta^2_\Delta \cdot t}, \text{ for } t \in [2020, 2022], \end{cases} \tag{1}$$

where

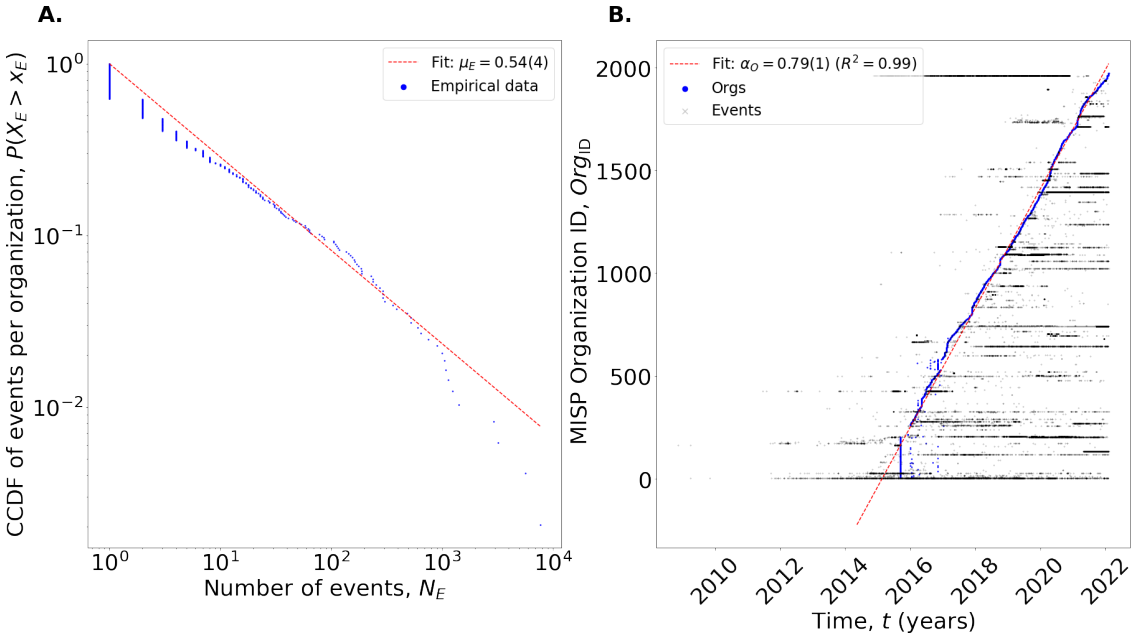

Figure 1: **A.** Complementary cumulative distribution function (CCDF) of events per contributing organization, which is best described by a power law distribution $P(X_E > x_E) \sim 1/{x_E}^{\mu_E}$ with $\mu_E = 0.54(4)$. The fit and the goodness-of-fit, provided by the Kolmogorov-Smirnov statistics test, are obtained with the `Python` library `plfit`. **B.** Curve of the joining organizations (in blue) has followed, after the September 14, 2015, the presumed date of official start, a linear growth with slope $\alpha_O = 0.79(1)$, ($R^2 = 0.99$, p-value $< 10^{-2}$). The events contributed by the organizations have been added (in dark gray), the distribution shows the heterogeneity of organizations efforts.

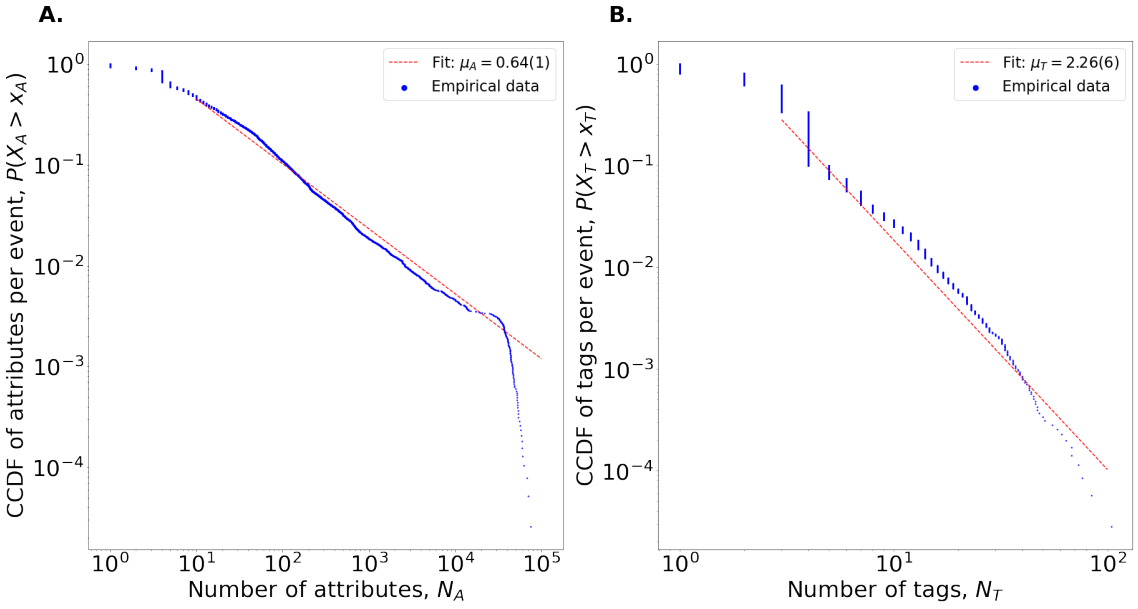

Figure 2: **A.** Complementary cumulative distribution function (CCDF) of attributes encapsulated in an event, which is best described by a power law distribution $P(X_A > x_a) \sim 1/{x_A}^{\mu_A}$ with $\mu_A = 0.64(1)$. **B.** CCDF of tags attached to an event which is best described by a power law distribution $P(X_T > x_T) \sim 1/x_T^{\mu_T}$ with $\mu_T = 2.26(6)$. The fits and the goodness-of-fits, provided by the Kolmogorov-Smirnov statistics test, of panels A and B are obtained with the `Python` library `plfit`.

- $\beta_\Delta^1 = (-6.32 \pm 0.91) \times 10^{-3}$ is the exponential decrease of the first part regression and
- $\beta_\Delta^2 = (-7.12 \pm 0.59) \times 10^{2}$ is the exponential decrease of the second part

regression.

The fit from the linear is of high quality since its Pearson's determination coefficient $R^2 = 0.86$ and its p-value $< 10^{-2}$. Hence, the time Δt_C to complete an event decreases over time, indicating an improvement of performances of the MISP CIRCL instance.

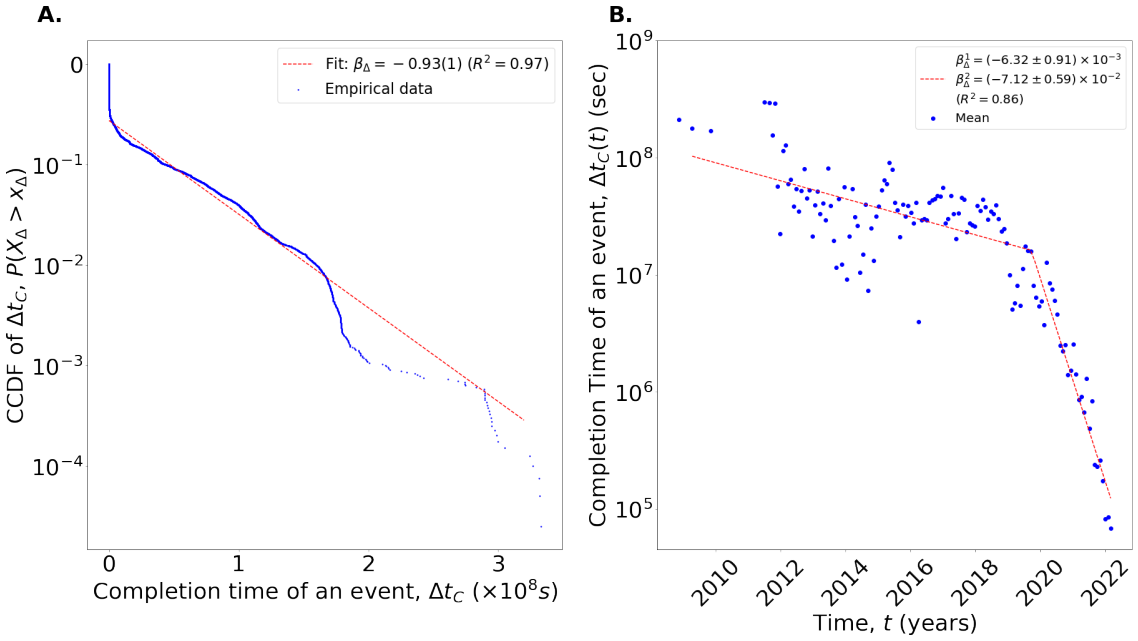

Figure 3: **A.** Complementary cumulative distribution function (CCDF) of the completion time Δt_C, which is best described by a decreasing exponential distribution $P(X_\Delta < x_\Delta) \sim 10^{\beta_\Delta}$ with $\beta_\Delta = -0.93(1)$. **B.** Completion time Δt_C of events over the time. The data (blue dots) represents the mean value of Δt_C binned monthly. The data depict an exponential decrease in two phases, fitted by linear regression (dashed red line), $\Delta t_C(t) \sim (-6.32 \pm 0.91) \times 10^{-2}$ for $t \in [2011, 2020[$ and $\Delta t_C(t) \sim (-7.12 \pm 0.59) \times 10^{-2}$ for $t \in [2020, 2022]$ ($R^2 = 0.86$, p-value $< 10^{-2}$). The fits and their goodness-of-fits, provided by the Pearson's coefficient of determination R^2 and the p-value for the Wald test, of panels A and B are obtained with the `Python` library `scipy.stats.linregress`.

4 Theoretical Framework

Collective action is thought to be a fundamental tool to overcome sprawling and increasing time-critical cybersecurity threats [54–56]. Yet, despite numerous studies of online platforms fostering collective action [57, 58], very little evidence has been uncovered linking the organisation of collective action with group performance as an output. By investigating the MISP threat management platform run by the Computer Incident Response Center Luxembourg (CIRCL), we have a unique chance to better understand how collective action is organized to tackle time-critical cybersecurity threats.

We posit that the performance of collective platforms devoted to the resolution of time-critical tasks at scale, such as MISP, pull from progressively building a knowledge and action environment, made of organizations, which contribute to the resolution of events and, at the same time, bring returns of scale through (i) gaining own experience and (ii) sharing and integrating knowledge, which is associated with increased performance. We further posit that, in order to offset decreasing return-of-scale due to increased groups size and coordination costs [29], the organization of collective action must adapt in a modular way [59], as it has already been witnessed in several open source projects [60, 61].

We test our theory of *collective action for tackling time-critical tasks*, through a set of three hypotheses and six sub-hypotheses to understand how time completion performance

is achieved for events, given (i) the nature of event, (ii) the collective action environment and (iii) the knowledge integration environment at the time of event arrival (c.f., section 5). We proceed with an exploratory approach to test our theory by resorting to a multivariate cross-sectional regression with ordinary least squares method (c.f., sections 6 and 7).

5 Hypotheses

To explain how event completion time has evolved, we consider their *intrinsic nature*, i.e., number of attributes and tags required to characterize events, the *overall collective action environment* and how *knowledge is integrated*. We hypthesize that these three overall factors significantly influence collective action performance, in terms of improved completion time in characterizing threat events.

5.1 Event Complexity Hinders Performance (H1)

First, events are not all equal: while some are fairly simple and require limited input in terms of attributes and of categorization with tags, others are more complex and require more effort. As shown on Figures 2A and 2B, the distribution of respectively attributes and tags is heavy-tailed: while a majority of events have a limited number of attributes (resp. tags), some carry a large numbers of attributes (resp. tags), presumably affecting the time required to complete the characterization of an event. Hypothesis 1 states:

H1: *The number of attributes and tags per event negatively influences performance.*

5.2 Collective Action Improves Performance (H2)

We consider how collective action at scale affects positively or negatively performance. Namely, there are conflicting views on whether having more stakeholders (e.g., contributors, organizations) joining collective action is likely to enhance or hinder performance [17, 25, 27–29]. Yet, to exist and be sustainable, collective action necessarily needs to bring economies of scale of some form, which in turn would attract more contributors. Conversely, having more participants should bring marginally increasing performance. Therefore, we aim to test the following hypothesis:

H2a: *The overall performance increases with the number of organizations participating in collective action.*

Yet, as already shown in [62], the ongoing collective action workload is likely to affect negatively performance, by increasing completion time. Therefore, our second hypothesis states:

H2b: *Given a focal event, the number of simultaneously open events decreases performance.*

5.3 Knowledge Integration Increases Performance (H3)

Having more contributors does not necessarily imply economies of scale [29]. Economies of scale may rather be generated by "the whole is more than the sum of its parts" mechanisms [25], which may stem from productive integration of information [46, 63, 64] as a single entity [25] or through the efficient communication of several modular

sub-systems [65, 66], which in turn may even mitigate free-riding [59]. Here, we recognize that the first form on knowledge integration occurs through experience as *learning* within organizations [67], and one may expect that an organization having accumulated experience in characterizing a large number of threat events is likely to perform better on new events, therefore :

H3a: *More experienced organizations solve events faster.*

On MISP instances, collective action goes beyond coordinating time-critical tasks. As people and organizations contribute, a large corpus of knowledge is built as a library of events, attributes, and tags. In turn, by design of MISP software, this information can be easily reused to quickly characterize new events, proposing matching possibilities according to the preliminary entries.

Hence, the reuse of knowledge simplifies the emission of attributes and the knowledge is integrated by the creator of the new events. These new events are thus composed of a certain percentage of *inherited* attributes which are likely to impact positively performance:

H3b: *The reuse of tags and attributes from existing events contributes positively to performance in the completion of new events.*

The capacity of an entity to integrate knowledge is tightly related to its modular organization [47, 59, 60]. As MISP clusters of events or attributes, called "Galaxies", were progressively introduced and developed on MISP CIRCL, we have an opportunity to test for modularity. We therefore formulate the following hypothesis:

H3c: *Modularity in collective action positively influences performance.*

By testing these three hypotheses (and six sub-hypotheses), we expect to gain robust insights on how collective action on MISP brings performance in terms of characterizing time-critical cybersecurity threats.

6 Method

We proceed to validate our theory through the testing of three hypotheses, divided in six sub-hypotheses (c.f., Section 5). For this, we specify an econometric model with *completion time* as the main dependent variable representing the key performance indicator in our posited *theory of collective action for tackling time-critical threats* (c.f., Section 4).

We define the following set of events,

$$\Omega_e = \{e | e \leq N_e, e \in \mathbb{N}^*\}, \tag{2}$$

where N_e corresponds to $22,423$ events, which have explicitly been marked as completed (i.e., with field $Analysis = 2$, see section 3). For each event, we define $\Delta t_{C,e}$ the completion time of events as

$$\Delta t_{C,e} = t_{f,e} - t_{c,e}, \tag{3}$$

with $t_{c,e}$ the event creation date and $t_{f,e}$ the last event modification.

To determine the relation between the dependent variable, i.e. the completion time $\Delta t_{C,e}$ for the events, we proceed to a multivariate cross-sectional regression [68]. Specifically, we investigate if completion time $\Delta t_{C,e}$ for the events can be explained by the selected explanatory variables. The corresponding `Python` variable is `CompletionT`. For each event e, the multivariate cross-sectional regression writes:

$$\log(\Delta t_{C,e}) = \zeta + \sum_{k=1}^{N_k} \cdot \sum_{e=1}^{N_e} \kappa_k \cdot \log(Z_{k,e}) + \varepsilon_e, \tag{4}$$

with:

- Δt_{C_e} : time completion for event e,
- ζ : constant,
- N_k : number of explanatory variables,
- κ_k : autoregressor parameter corresponding to $Z_{k,e}$,
- $Z_{k,e}$: k-th explanatory variable for event e,
- ε_e : error term (i.e., $\log(\Delta t_{C,e}) - \log(\widehat{\Delta t_{C,e}})$).

This multivariate cross-sectional regression is performed with the ordinary least squares (OLS) method. The choice of this model is adapted to deal with data without time series, which is the case here. Then, the explicated and explanatory variables are linked with a set of points in time. This set of points in time is given by the creation $t_{c,e}$ of the different e and contains $22,423$ elements, corresponding to the number of completed elements N_e considered. Thanks to this model, it is easy to consider all chosen independent variables. However, due to the heavy-tailed behaviour of the variables and their difference of magnitude (see Section 3), we transform the variables in logarithm in base of 10 [69]. However, the results are indicated as a percentage change of $\Delta t_{C,e}$ when $Z_{k,e}$ varies by a certain percentage [69].

We specify the following explanatory variables in relation with the formulated hypotheses (c.f., Section 5). To test hypothesis **H1** (i.e., *event complexity hinders performance*), we resort to two explanatory variables:

- $N_{A,e}$: the number of attributes per event e. The corresponding `Python` variable is `AttrCount`, which is expected to positively influence `CompletionT` (i.e., reduce performance).
- $N_{T,e}$: the number of tags per event e, The corresponding `Python` variable is `NTags`, which is expected to positively influence `CompletionT` (i.e., reduce performance).

To test hypothesis **H2** (i.e., *collective action improves performance*), we resort to two explanatory variables:

- $N_{O,e}$ stands for the number of organizations listed on MISP CIRCL at the creation $t_{c,e}$ of event e. The corresponding `Python` variable is `CumOrgs`. `CumOrgs` is expected to negatively influence `CompletionT` (i.e., increase performance) and to demonstrate the overall benefits of collective action for tackling time criticial threats (**H2a**).
- $E_{\text{sim},e}$ is the number of simultaneously open events on MISP CIRCL at the creation $t_{c,e}$ of event e. The corresponding `Python` variable is `SimEvents`, which is expected to positively influence `CompletionT` (i.e., reduce performance) and to show that collective action performance is bound to circumstantial operational constraints associated with time as a scarce resource (**H2b**) [62, 70].

To test hypothesis **H3** (i.e., *knowledge integration increases performance*), we resort to three explanatory variables:

– $E_{C,e}$ takes into account the number of already completed events by the organizations at the creation $t_{c,e}$ of a new event e on their behalf. The corresponding `Python` variable is `CumCompE`, which is expected to negatively influence `CompletionT` (i.e., increase performance) (**H3a**).

– $I_{\%A,e}$ is the inherited percentage of attributes per event e. The corresponding `Python` variable is `InhPer`, which is expected to negatively influence `CompletionT` (i.e., increase performance) (**H3b**).

– $N_{G,e}$ counts the number of galaxies created on MISP CIRCL instance at the creation $t_{c,e}$ of the e. The corresponding `Python` variable is `NbGalaxies`, which is expected to negatively influence `CompletionT` (i.e., increase performance) (**H3c**).

– $N_{E_G,e}$ considers the number of events in its corresponding aforementioned galaxy at the creation $t_{c,e}$ of a new event e in this galaxy. The corresponding `Python` variable is `NbEventsinhisG`, which is expected to negatively influence `CompletionT` (i.e., increase performance) (**H3c**).

The pairwise correlations of the dependent variable and the independent ones provide the correlation matrix (see Table 2).

	$\log(\Delta t_C)$	$\log(N_{A,e})$	$\log(I_{\%A,e})$	$\log(N_{T,e})$	$\log(E_{\text{sim},e})$	$\log(N_{O,e})$	$\log(E_{C,e})$	$\log(N_{G,e})$	$\log(N_{E_G,e})$
$\log(\Delta t_C)$	1.00								
$\log(N_{A,e})$	0.11	1.00							
$\log(I_{\%A,e})$	-0.07	-0.27	1.00						
$\log(N_{T,e})$	0.07	0.08	-0.59	1.00					
$\log(E_{\text{sim},e})$	0.74	0.06	0.01	0.04	1.00				
$\log(N_{O,e})$	-0.23	-0.03	0.05	0.01	0.02	1.00			
$\log(E_{C,e})$	-0.60	0.023	-0.02	0.01	-0.53	0.33	1.00		
$\log(N_{G,e})$	-0.16	0.01	-0.07	-0.02	-0.42	0.19	0.23	1.00	
$\log(N_{E_G,e})$	-0.12	0.00	-0.07	0.07	-0.11	0.42	0.43	0.14	1.00

Table 2: Correlation matrix of dependent and explanatory variables.

With the explanatory variables of our model being defined, we are in position to formulate the econometric model by developing the equation (4):

$$\begin{aligned}\log(\Delta t_{C,e}) = \zeta &+ \kappa_{N_A} \cdot \log(N_{A,e}) + \kappa_{I_{\%A}} \cdot \log(I_{\%A,e}) + \kappa_{N_T} \cdot \log(N_{T,e}) \\ &+ \kappa_{E_{\text{sim}}} \cdot \log(E_{\text{sim},e}) + \kappa_{N_O} \cdot \log(N_{O,e}) + \kappa_{E_C} \cdot \log(E_{C,e}) \\ &+ \kappa_{N_G} \cdot \log(N_{G,e}) + \kappa_{N_{E_G}} \cdot \log(N_{E_G,e}) \\ &+ \varepsilon_e \end{aligned} \tag{5}$$

Model validation is performed as follows. When handling a multivariate regression, one must pay particular attention to multi-collinearity between the Z_k's, which may distort the model. For that, the variance inflation factor (VIF) resulting from the regression of the explanatory variable Z_k on the other explanatory variables which provide R_k^2, must be computed. The VIF_k is then given as $\text{VIF}_k = 1/(1 - {R_k}^2)$ and must be < 10 [68]. The stability of the variance has to be examined, namely by studying heteroskedasticity, which is ruled out if the p-value obtained from a White test is lower than

a threshold $\alpha = 0.05$ [68]. The computation steps are performed with the `Python` libraries `statsmodels.api.OLS` for the regression, `statsmodels.stats.outliers_influence` for the VIF and `statsmodels.stats.diagnostic` for the White test.

7 Results

In order to establish evidence of collective action as an efficient way for tackling time-critical cybersecurity threats, we have resorted to data the MISP instance, which is run by the computer Incident Response Center Luxembourg (CIRCL). We used a multivariate cross-sectional regression analysis of *completion time* (i.e., performance) required to characterize a threat event with both event related and collective action explanatory variables.

Dep. Variable		Completion Time	
Method	OLS	F-Stat	2.251×10^3
No. Observations	22423	Prob (F-Stat)	0.00
R-squared	0.413	Log-likelihood	-5.030×10^4
	coeff		std err
Const	$16.505^{(***)}$		0.135
CountAttr	$0.230^{(***)}$		0.011
InhPer	$-0.089^{(***)}$		0.014
NTags	$0.951^{(***)}$		0.090
CumOrgs	$-0.346^{(***)}$		0.024
CumCompE	$-0.629^{(***)}$		0.006
NbGalaxies	$-0.083^{(***)}$		0.019
NbEventsinhisG	$0.160^{(***)}$		0.005
Skew	-0.011	Durbin-Watson	1.302
Kurtosis	2.833	Cond No.	76.4

Table 3: Results of the ordinary least squares (OLS) regression with the explained variable `CompletionT` and the explanatory variables: `CountAttr, InhPer, NTags, CumOrgs, CumCompE, NbGalaxies` and `NbEventsinhisG`, namely the number of attributes per event, the inherited percentage of attributes per event, the number of tags per event, the cumulative number of organizations at the creation of the event e, the number of already completed events by the organization at the creation of his new event e, the number of galaxies at the creation of the event e and the number of events populating these galaxies at the creation of the event e. For each explanatory variable, the autoregressor coefficient (in the column `coeff`), as well as its standard deviation (in the column `std err`) are provided. The significance of the explanatory variables is given by the p-value and its threshold, i.e. $p - \text{value} < 0.1 : (*), < 0.05 : (**)$ or $< 0.01 : (***)$ and the goodness-of-fit by the `R-squared`. The other added information are not necessary for the evaluation of the model.

The regression results are shown in Table 3. Overall, the regression model is robust and explains 43% of the variance ($R^2 = 0.413$). Testing for hypothesis 1, the model shows that indeed event complexity measured by the number of attributes `CountAttr` and tags `NTags` influences performance negatively, i.e., event characterization completion time is increased. Hypothesis H1 is supported. Regarding how collective action improves performance (H2), the model shows that overall performance (i.e., completion time reduced) is positively associated with the number of organizations participating in MISP: Hypothesis H2a is supported. Hypothesis H2b could not be tested as a result of unexplained strong multi-collinearity between `CumOrgs` and `SimEvents`. Turning to Hypothesis 3 (i.e., knowledge integration increases performance), we find that more experienced organizations perform better in reducing event completion time. Hypothesis H3a is supported. We also find that the proportion of attributes that an event e inherits from previous events, i.e., from the MISP CIRCL knowledge base, also positively influences performance. Hypothesis

H3b is supported. Finally, testing for hypothesis H3c, i.e., modularity, we find mixed results. While the number of MISP Galaxies, measuring the number of modular subsystems, influences positively performance, the number of events recorded in MISP Galxies, measuring to some extent the intensity of modularity, influences performance negatively. Hypothesis H3b is only partially supported.

We have checked for multi-collinearity of the explanatory variables. We computed the variance inflation factor (VIF) for each explanatory variables, which happens to be all smaller than 10. This implies that there is no evidence of multi-collinearity between the selected explanatory variables (c.f., Table 4). We also controlled for heteroskedasticity, i.e., a possible instability of the variance by performing a White statistics tests. We obtained p-value $< 10^{-2}$, which implies that there is no heteroskedasticity in our model. The post-analysis for the VIFs and the White statistics test completely validate the used model and its results.

Explanatory variables	Notation	VIF
Number of attributes per event	$N_{A,e}$	5.15
inherited percentage of attributes per event e	$I_{\%A,e}$	1.67
Number of tags per event e	$N_{T,e}$	1.03
Cumulated number of organizations at the creation of e	$F_{\text{cum},e}$	6.73
Cumulated number of completed events at the creation of e	$E_{C,\text{cum},e}$	3.28
Cumulated number of galaxies at the creation of e	$N_{G,\text{cum},e}$	1.12
Cumulated number of events in galaxies at creation of e	$N_{E_G,\text{cum},e}$	2.02

Table 4: Computation of the variance inflation factor (VIF) for the explanatory variables of the econometric model. The values of the VIF allows to detect the presence of multi-collinearity between the considered variables. As all values VIF < 10, there is no evidence of multi-collinearity between the explanatory variables. These results validate the econometric model.

8 Discussion

Organizations are increasingly encouraged to cooperate and share information to overcome cybersecurity threats. Investigating how collective action unfolds and brings performance on information-sharing platforms is necessary as cybersecurity threats have become increasingly time-critical. In other words, not only collective action shall be used to characterize threat events, it also must be used to characterize threat events before attacks unravel [56]. Here, we have investigated collective action on MISP, a popular open source threat intelligence platform, from the perspective of the time required to fully characterize an event as an objective function to be optimized (i.e., completion time or performance). We found that performance is negatively associated with event complexity (Hypothesis 1) and positively associated with collective action (Hypothesis 2). Indeed, as the number of organizations taking part to information-sharing on the MISP instance studied increased, the time required to complete the characterization of events decreased. This result informs on positive returns on scale, which necessarily exist given the increased adoption of MISP as well as other information-sharing platforms. Nevertheless, the mechanisms at work generating these economies of scale have remained unclear. We considered the perspective of knowledge integration [47] as the collective action process at work to generate the "the whole is more than the sum of its parts" [25]. With hypothesis 3, we tested and verified organizational learning, knowledge integration and modularity as positively associated with performance.

While event completion time is associated with explanatory variables pertaining to event complexity, collective action, and knowledge integration, we could not establish

causality. Although this is a significant limitation to our model, we have organized our multivariate cross-sectional regression in a way that minimizes the risks of uncovering spurious dependencies between the explained variable on the one hand and the explanatory variables on the other hand. And the fact that all our explanatory variables are significant (at the exception of `SimEvents`, the number of simultaneously open events on MISP CIRCL at the creation, which had to be excluded from the model), shows that our proposed theory on *collective action for tackling time-critical tasks* is comprehensive and altogether robust. Yet, the regression analysis approach remains exploratory. Indeed, it does not provide reliable information on which precise collective action mechanisms generate positive returns on scale. Building and testing fine-grained causal models of critical cascades in collective action, inspired from e.g. [25, 27, 28], may surely help better understand the activity, learning, knowledge integration and modularization paths of contributing organizations, as well as how they handle time as a particularly scarce resource [70]. Indeed, when tackling large amounts of time-critical tasks, such as cybersecurity threats or incidents, contingencies necessarily appear [62], which may affect coordination between contributors, and as a result performance, either in a transient way or by triggering long-term instability through cascades of disorganization. At the meso-scale, our model does not account for affinities between events, organizations and the combined commonalities of events and organizations. Indeed, as for number of collective action online platforms, modular *Galaxies* on MISP show that some sub-communities of organizations have specific goals when tackling cybersecurity threats. These specific interests deserve further scrutiny. For instance, are the organizations contributing to a given MISP galaxy active in the same industry? If not, why do they share interest in similar threats? Considering MISP (or other information-sharing platforms) from the perspective of threats, we may investigate kinship between threats, as they most often share attributes. Questioning and perhaps predicting how attributes are "transmitted" from one event to others is likely to be key to anticipate threats and guide organizations in their search of (respectively contributions to) threat information. It may even help decide what information should be shared and with whom.

Finally, our results show that completion time as an objective function in collective action concerned with time-critical tasks can be optimized. This opens further perspectives for computational social science research. One may envision to use machine learning in order to recommend personalized precision strategies that optimize the organization of collective action and knowledge integration. This may help make best use of time as an increasingly critically scarce resource, especially in face of a looming tsunami of cybersecurity threats.

9 Conclusion

Information-sharing in cyber-security has become an increasingly common collective action practice. Yet, its benefits have so far remained unclear. We have investigated MISP, a commonly used open source threat sharing platform, and we found how building a critical mass of contributing organizations and of knowledge to be integrated from past threats brings significant economies of scale. Through collective action, security researchers overcome the challenge of characterizing cybersecurity threats, which appear to be increasingly time-critical. We find that performance, defined as the time needed to fully characterize a threat event, is (i) negatively influenced its own complexity, (ii) positively influenced by collective action, and (iii) positively by learning, knowledge integration and modularity. Our results also inform more generally on how collective action can be organized online at scale and in a modular way to overcome a large number of time-critical tasks.

References

1. Sridhar, K., Householder, A., Spring, J. & Woods, D. W. *Cybersecurity Information Sharing: Analysing an Email Corpus of Coordinated Vulnerability Disclosure* en. in (2021), 39.
2. Gal-Or, E. & Ghose, A. The Economic Incentives for Sharing Security Information. en. *Information Systems Research* **16,** 186–208. ISSN: 1047-7047, 1526-5536 (June 2005).
3. Maillart, T., Zhao, M., Grossklags, J. & Chuang, J. Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *Journal of Cybersecurity* **3,** 81–90 (June 2017).
4. Sridhar, K. & Ng, M. Hacking for good: Leveraging HackerOne data to develop an economic model of Bug Bounties. *Journal of Cybersecurity* **7** (Jan. 2021).
5. Böhme, R. *Back to the Roots: Information Sharing Economics and What We Can Learn for Security* in *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security* (ACM, Vienna Austria, Oct. 2016), 1–2.
6. Laube, S. & Böhme, R. Strategic Aspects of Cyber Risk Information Sharing. *ACM Computing Surveys* **50,** 77:1–77:36 (Nov. 2017).
7. Brady, R. M., Anderson, R. J. & Ball, R. C. *Murphys law, the fitness of evolving species, and the limits of software reliability* en. Tech. rep. UCAM-CL-TR-471 (University of Cambridge, Computer Laboratory, 1999). `https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-471.html` (2022).
8. Stojkovski, B., Lenzini, G., Koenig, V. & Rivas, S. *Whats in a Cyber Threat Intelligence sharing platform?: A mixed-methods user experience investigation of MISP* en. in *Annual Computer Security Applications Conference* (ACM, Virtual Event USA, Dec. 2021), 385–398. ISBN: 978-1-4503-8579-4. `https://dl.acm.org/doi/10.1145/3485832.3488030` (2022).
9. Levy, S. hackers: heroes of the computer revolution. en. **35,** 4 (2010).
10. Benkler, Y. *The Penguin and the Leviathan: How Cooperation Triumphs over Self-Interest* en. ISBN: 978-0-307-59019-0 (Crown, Aug. 2011).
11. Benkler, Y. *The wealth of networks: how social production transforms markets and freedom* en. OCLC: ocm61881089. ISBN: 978-0-300-11056-2 (Yale University Press, New Haven [Conn.], 2006).
12. Katz, M. L. & Shapiro, C. Network Externalities, Competition, and Compatibility. en, 18 (2021).
13. Shapiro, C. & Varian, H. R. *Information rules: a strategic guide to the network economy* en. ISBN: 978-0-87584-863-1 (Harvard Business School Press, Boston, Mass, 1999).
14. Wagner, C., Dulaunoy, A., Wagener, G. & Iklody, A. *MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform* in *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security* (Association for Computing Machinery, New York, NY, USA, Oct. 2016), 49–56.
15. Zibak, A. & Simpson, A. *Cyber Threat Information Sharing: Perceived Benefits and Barriers* en. in *Proceedings of the 14th International Conference on Availability, Reliability and Security* (ACM, Canterbury CA United Kingdom, Aug. 2019), 1–9.
16. Mell, P., Scarfone, K. & Romanosky, S. Common Vulnerability Scoring System. *IEEE Security Privacy* **4.** Conference Name: IEEE Security Privacy, 85–89. ISSN: 1558-4046 (Nov. 2006).

17. Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups, With a New Preface and Appendix* Revised edition. English. ISBN: 978-0-674-53751-4 (Harvard University Press, Cambridge, Mass., Jan. 1971).

18. Anesi, V. Moral hazard and free riding in collective action. *Social Choice and Welfare* **32,** 197 (June 2008).

19. Esteban, J. & Ray, D. Collective Action and the Group Size Paradox. *The American Political Science Review* **95.** Publisher: [American Political Science Association, Cambridge University Press], 663–672 (2001).

20. Dejean, S., Pénard, T. & Suire, R. *Olson's Paradox Revisited: An Empirical Analysis of incentives to contribute in P2P File-Sharing Communities* SSRN Scholarly Paper ID 1299190 (Social Science Research Network, Rochester, NY, July 2010).

21. Asvanund, A., Clay, K., Krishnan, R. & Smith, M. D. An Empirical Analysis of Network Externalities in Peer-to-Peer Music-Sharing Networks. *Information Systems Research* **15.** Publisher: INFORMS, 155–174 (June 2004).

22. Klein, M., Maillart, T. & Chuang, J. *The Virtuous Circle of Wikipedia: Recursive Measures of Collaboration Structures* in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Association for Computing Machinery, New York, NY, USA, Feb. 2015), 1106–1115.

23. Wang, S., Lo, D. & Jiang, L. *An empirical study on developer interactions in StackOverflow* in *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (Association for Computing Machinery, New York, NY, USA, Mar. 2013), 1019–1024. ISBN: 978-1-4503-1656-9. `https://doi.org/10.1145/2480362.2480557` (2022).

24. Hippel, E. v. & Krogh, G. v. Open Source Software and the Private-Collective Innovation Model: Issues for Organization Science. *Organization Science* **14.** Publisher: INFORMS, 209–223. ISSN: 1047-7039. `https://pubsonline.informs.org/doi/abs/10.1287/orsc.14.2.209.14992` (2022) (Apr. 2003).

25. Sornette, D., Maillart, T. & Ghezzi, G. How Much Is the Whole Really More than the Sum of Its Parts? 1 1 = 2.5: Superlinear Productivity in Collective Group Actions. *PLoS ONE* **9** (ed Perc, M.) e103023 (Aug. 2014).

26. Sornette, D., Maillart, T. & Ghezzi, G. How much is the whole really more than the sum of its parts? $1 \boxplus 1 = 2.5$: Superlinear productivity in collective group actions. *Plos one* **9,** e103023 (2014).

27. Maillart, T. & Sornette, D. Aristotle vs. Ringelmann: On superlinear production in open source software. en. *Physica A: Statistical Mechanics and its Applications* **523,** 964–972 (June 2019).

28. Muri, G., Abeliuk, A., Lerman, K. & Ferrara, E. Collaboration Drives Individual Productivity. en. *Proceedings of the ACM on Human-Computer Interaction* **3,** 1–24. ISSN: 2573-0142 (Nov. 2019).

29. Scholtes, I., Mavrodiev, P. & Schweitzer, F. *From aristotle to ringelmann: A large-scale analysis of team productivity and coordination in open source software projects* in (Gesellschaft für Informatik e.V., 2016).

30. Tziner, A. & Eden, D. Effects of crew composition on crew performance: Does the whole equal the sum of its parts? *Journal of Applied Psychology* **70.** Place: US Publisher: American Psychological Association, 85–93. ISSN: 1939-1854(Electronic),0021-9010(Print) (1985).

31. Sundstrom, E., De Meuse, K. P. & Futrell, D. Work teams: Applications and effectiveness. *American Psychologist* **45.** Place: US Publisher: American Psychological Association, 120–133 (1990).

32. Cohen, S. G. & Bailey, D. E. What Makes Teams Work: Group Effectiveness Research from the Shop Floor to the Executive Suite. en. *Journal of Management* **23.** Publisher: SAGE Publications Inc, 239–290. ISSN: 0149-2063 (June 1997).

33. Neuman, G. A. & Wright, J. Team effectiveness: Beyond skills and cognitive ability. en. *Journal of Applied Psychology.* ISSN: 0021-9010 (Jan. 1999).

34. Ennen, E. & Richter, A. The Whole Is More Than the Sum of Its Parts Or Is It? A Review of the Empirical Literature on Complementarities in Organizations. en. *Journal of Management.* ISSN: 0149-2063 (Jan. 2010).

35. Sacramento, C. A., Chang, M.-W. S. & West, M. A. in *Innovation through collaboration* (Emerald Group Publishing Limited, 2006).

36. Nonaka, I. A Dynamic Theory of Organizational Knowledge Creation. en. *ORGANIZATION SCIENCE* **5,** 25 (1994).

37. Grant, R. M. Prospering in Dynamically-Competitive Environments: Organizational Capability as Knowledge Integration. *Organization Science* **7,** 375–387 (1996).

38. Wuchty, S., Jones, B. F. & Uzzi, B. The Increasing Dominance of Teams in Production of Knowledge. en. *Science* **316,** 1036–1039. ISSN: 0036-8075, 1095-9203 (May 2007).

39. Lawrence, P. R. & Lorsch, J. W. Differentiation and Integration in Complex Organizations. en. *Administrative Science Quarterly* **12,** 1. ISSN: 00018392 (June 1967).

40. Majchrzak, A., Griffith, T. L., Reetz, D. K. & Alexy, O. Catalyst Organizations as a New Organization Design for Innovation: The Case of Hyperloop Transportation Technologies. en. *Academy of Management Discoveries* **4,** 472–496. ISSN: 2168-1007. `http://journals.aom.org/doi/10.5465/amd.2017.0041` (2022) (Dec. 2018).

41. Safa, N. S. & Von Solms, R. An information security knowledge sharing model in organizations. en. *Computers in Human Behavior* **57,** 442–451. ISSN: 07475632. `https://linkinghub.elsevier.com/retrieve/pii/S0747563215303083` (2022) (Apr. 2016).

42. Mockus, A., Fielding, R. T. & Herbsleb, J. A case study of open source software development: the Apache server. *Software Engineering, 2000. Proceedings of the 2000 International Conference on,* 263–272 (2000).

43. Debreu, K. J. A. a. G. Existence of an Equilibrium for a Competitive Economy. *The Econometric Society* **22,** pp. 265–290 (1954).

44. Ravasi, D. & Verona, G. Organising the process of knowledge integration: the bene"ts of structural ambiguity. en, 26 (2001).

45. Huang, J. C. & Newell, S. Knowledge integration processes and dynamics within the context of cross-functional projects. en. *International Journal of Project Management,* 10 (2003).

46. Engel, D. & Malone, T. W. Integrated information as a metric for group interaction. *PLOS ONE* **13** (ed Dovrolis, C.) e0205335 (Oct. 2018).

47. Tononi, G. Consciousness and Complexity. en. *Science* **282,** 1846–1851 (Dec. 1998).

48. Zheng, D. E. Cyber Threat Information Sharing: Recommendations for Congress and the Administration. en, 18 (2015).

49. He, M., Devine, L. & Zhuang, J. Perspectives on Cybersecurity Information Sharing among Multiple Stakeholders Using a Decision-Theoretic Approach: Cybersecurity Information Sharing. en. *Risk Analysis* **38,** 215–225. ISSN: 02724332. `https://onlinelibrary.wiley.com/doi/10.1111/risa.12878` (2022) (Feb. 2018).

50. Meier, R., Scherrer, C., Gugelmann, D., Lenders, V. & Vanbever, L. *FeedRank: A tamper- resistant method for the ranking of cyber threat intelligence feeds* in *2018 10th International Conference on Cyber Conflict (CyCon)* ISSN: 2325-5374 (May 2018), 321–344.

51. Mermoud, A., Keupp, M. M., Huguenin, K., Palmié, M. & Percia David, D. To share or not to share: a behavioral perspective on human participation in security information sharing. *Journal of Cybersecurity* **5,** tyz006. ISSN: 2057-2085. https://doi.org/10.1093/cybsec/tyz006 (2022) (Jan. 2019).

52. Percia David, D., Keupp, M. M. & Mermoud, A. Knowledge absorption for cybersecurity: The role of human beliefs. *Computers in Human Behavior* **106,** 106255 (May 2020).

53. Mermoud, A., Keupp, M. M. & Percia David, D. *Governance Models Preferences for Security Information Sharing: An Institutional Economics Perspective for Critical Infrastructure Protection* in *Critical Information Infrastructures Security* (eds Luiijf, E., utautait, I. & Hämmerli, B. M.) (2019).

54. Mermoud, A. *Three articles on the behavioral economics of security information sharing: A theoretical framework, an empirical test, and policy recommendations* PhD Thesis (Université de Lausanne, Faculté des hautes études commerciales, 2019).

55. Bouwman, X. Governance of cybersecurity communities: Understanding threat intelligence sharing as a collective action problem through incentivization of the National Detection Network. en. https://repository.tudelft.nl/islandora/object/uuid%3A3134a935-1156-409d-a58c-47cf06a88dad (2022) (2018).

56. Wagner, T. D., Mahbub, K., Palomar, E. & Abdallah, A. E. Cyber threat intelligence sharing: Survey and research directions. en. *Computers & Security* **87,** 101589. ISSN: 0167-4048. https://www.sciencedirect.com/science/article/pii/S016740481830467X (2022) (Nov. 2019).

57. Bouwman, X. *et al.* Helping hands: Measuring the impact of a large threat intelligence sharing community. en, 17 (2022).

58. McColl, R. C., Ediger, D., Poovey, J., Campbell, D. & Bader, D. A. *A performance evaluation of open source graph databases* in *Proceedings of the first workshop on Parallel programming for analytics applications* (Association for Computing Machinery, New York, NY, USA, 2014), 11–18. ISBN: 978-1-4503-2654-4. https://doi.org/10.1145/2567634.2567638 (2022).

59. Baldwin, C. Y. & Clark, K. B. The Architecture of Participation: Does Code Architecture Mitigate Free Riding in the Open Source Development Model? *Management Science* **52.** Publisher: INFORMS, 1116–1127. ISSN: 0025-1909. https://pubsonline.informs.org/doi/abs/10.1287/mnsc.1060.0546 (2022) (July 2006).

60. Narduzzo, A. & Rossi, A. *The Role of Modularity in Free/Open Source Software Development* en. Chap. ISBN: 9781591403692 Pages: 84-102 Publisher: IGI Global. 2005. https://www.igi-global.com/chapter/role-modularity-free-open-source/www.igi-global.com/chapter/role-modularity-free-open-source/18721 (2022).

61. Langlois, R. N. & Garzarelli, G. Of Hackers and Hairdressers: Modularity and the Organizational Economics of Opensource Collaboration. en. *Industry and Innovation* **15,** 125–143. ISSN: 1366-2716, 1469-8390. https://www.tandfonline.com/doi/full/10.1080/13662710801954559 (2022) (Apr. 2008).

62. Kuypers, M. & Maillart, T. *Designing Organizations for Cyber Security Resilience* in *Proceedings of the 2018 The Workshop on the Economics of Information Security (WEIS), Innsbruck, Austria* (2018), 18–19.

63. Oizumi, M., Albantakis, L. & Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology* **10** (ed Sporns, O.) e1003588 (May 2014).

64. Malone, T. W. *Superminds: how hyperconnectivity is changing the way we solve problems* English. OCLC: 1130848208. ISBN: 978-1-78607-568-0 (Oneworld Publicatins, London, 2019).

65. Barrett, A. B. & Seth, A. K. Practical Measures of Integrated Information for Time-Series Data. en. *PLoS Computational Biology* **7** (ed Sporns, O.) e1001052. ISSN: 1553-7358. `https://dx.plos.org/10.1371/journal.pcbi.1001052` (2021) (Jan. 2011).

66. Baldwin, C. & Clark, K. *Design Rules: The Power of Modularity* (MIT Press, Cambridge, 2000).

67. Argote, L. & Miron-Spektor, E. Organizational Learning: From Experience to Knowledge. *Organization Science* **22.** Publisher: INFORMS, 1123–1137. ISSN: 1047-7039. `https://pubsonline.informs.org/doi/abs/10.1287/orsc.1100.0621` (2022) (Oct. 2011).

68. Asteriou, D. & Hall, S. G. *Applied Econometrics* en. Google-Books-ID: eOEdCwAAQBAJ. ISBN: 978-1-137-41547-9 (Macmillan International Higher Education, Oct. 2015).

69. Benoit, K. Linear Regression Models with Logarithmic Transformations. en, 8 (2011).

70. Maillart, T., Sornette, D., Frei, S., Duebendorfer, T. & Saichev, A. Quantification of deviations from rationality with heavy tails in human dynamics. *Physical Review E* **83,** 056101 (May 2011).

71. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review* **51,** 661–703 (Feb. 2009).

Appendices

A MISP: Description and Data Retrieval

A.1 Detailed Description of MISP

MISP is a partially de-centralized system of communities (e.g., NATO MISP, CIRCL MISP). interacting more or less together across MISP instances. A MISP instance consists in the installation of the MISP software and the community database in which community members share and collect data. Similarly to *GIT*,[4] **organizations** work on their own instance and synchronize with remote instances. According to their sharing setting (i.e., your organization only, community only, connected communities, all communities or defined sharing group), community members have access to a certain amount of data.

Based on investigation needs or reports found in the newspapers or on specialized websites, the user creates an **event** to contextualize and encapsulate the related **attributes** (i.e., IoCs) and their properties (e.g., an IP address). All events have some general properties of the event, such *creation date*, aforementioned sharing level, *threat level* (i.e., 1: High, 2: Medium, 3: Low, 4: Undefined), **analysis level** (i.e., 0: Initial, 1: Ongoing, 2:

[4] `https://git-scm.com/`

Complete) and a general description. The creator of an event can choose if this event is published on the remote instance or remains internal to the organization. Then, when the event is created, some attributes are added to populate this event. The event attributes refer to intrusion artifacts or methods used by attackers. These attributes provide details and they are characterized by their **type** (e.g., filename|md5, sha256, etc.) and their belonging to a **category** (e.g., Antivirus detection, Targeting data, etc.), putting them in the context and justify then its attribution to its corresponding event. To add an attribute related to an event, global information such as its category, its type and its distribution, either the same as for the event or its own rule, is required, as well as two important text fields: **value** and **contextual comment**. The "value" field stores the data we want to add, e.g. an url leading to a report, while the "comment" field allows complementary information about the attribute. Moreover, it is possible to allocate one **tag** or more to an event in order to simplify the read and the classification of this event. These tags can follow the MISP taxonomy, i.e. a fixed machine-tag vocabulary, or be created by the users according to their needs.

On the platform, events, attributes, organizations and tags are associated to their own identification (ID) number and their creation are timestamped, as well as the publication and the last update of an event.

As an open-source platform, MISP relies on voluntary action. On the one hand, its members can create or exchange content. On the other hand, these same actors can obtain new insights or possible response elements from the community regarding cyber-threats of interest. To organize interactions and to create information-sharing incentives for the participants, MISP offers several aforementioned sharing levels through a comprehensive sharing model. Users can select to whom they want to share information among the following levels from the most restrictive to the most open. Regardless of access and to guarantee the quality of the shared data, only organizations that created an event have the permission to modify this event. However, each user has the possibility to submit his own suggestions to change an event created by others, who can then accept or reject the proposal.

Moreover, the experience of older MISP versions has shown that the time to fill the fields and a complicated web interface introduce some frictions. For this purpose, a free text importer has been deployed, so that data can be copied and pasted into the intended field. Further, MISP implements a heuristics-based algorithm, which helps users to match events or event attributes with events or attributes from events already in the data base. However, let us added that the matching is never performed automatically, and goes through human supervision.

A.2 Data Retrieval

To investigate our hypotheses, we have to curate the main dataset by considering only the closed events, i.e. the events with an analysis level equal to 2, meaning "complete".

To retrieve the data, we have followed the user guide[5] provided by the MISP CIRCL instance. We used the PyMISP module to download data in `.json` format file. The main dataset contains one file per event. These event files contain the attributes (see MISP core format[6]), as well as the name and the ID of the concerned organizations.

[5] `https://www.circl.lu/doc/misp/book.pdf`

[6] `https://www.misp-standard.org/rfc/misp-standard-core.html`

However, due to the policy of the MISP CIRCL instance, we cannot disclose the names of these organizations and present no interest and have no influence on the obtained results.

B Exploratory Analysis of the Data Set

B.1 Probabilistic Distributions

In order to understand the mechanisms handling on the MISP platform, we want to investigate the distribution of our data, we have to present the selected variables and explore the distribution associated with these. In some cases, we are able to investigate the probabilities distribution. Hence, if we consider a random variable X with a probability density function (PDF) $f_X(x)$, the cumulative distribution function (CDF), $F_X(x)$ is given by:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt. \tag{6}$$

Then, thanks to the formula (6), the complementary cumulative distribution function (CCDF) $\bar{F_X}(x)$ can be written as follow:

$$\bar{F_X}(x) = 1 - F_X(x) = P(X > x). \tag{7}$$

This CCDF provides a rank ordering of the selected variables.

B.2 Fit of the Data

Before we start fitting our data, a visual analysis can be performed. Then, in any case, by varying the scale of axis – double linear, linear-logarithmic or double logarithmic – depicting our data, we are able, if our data follow approximately a straight line in one of cases presented below, to fit the data. The logarithmic scales are considered in base 10.

B.2.1 Double Linear Scales

By considering two vectors of data $\overrightarrow{x}$ and $\overrightarrow{y}$ and plotting the data contained in $\overrightarrow{y}$ (y-axis) in function of the data in $\overrightarrow{x}$ (x-axis) in linear scale for the axes x and y. If the displayed data shows an approximate straight line, that means that each element y_i of the vector $\overrightarrow{y}$ is given by the relation:

$$y_i = a \cdot x_i + b, \tag{8}$$

where a is the slope of the straight line and b, its intercept. Thanks to the relation (8), we are able to compute the estimated $\hat{y_i}$, a and b by applying a least-square linear regression. To validate the parameter obtained from the linear regression, we need to establish the goodness-of-fit with these parameters. For this type of simple linear regression, we use the Pearson's coefficient of determination R^2 and, to reinforce the results of R^2, we perform a Wald test with a chosen level $\alpha = 0.05$ to define if these two samples are significantly identical or not. Then a value $|R^2| \approx 1$ implies a strong correlation between $\overrightarrow{x}$ and $\overrightarrow{y}$, while a p-value $< \alpha$ for the Wald test allows us to affirm that the parameters of the fit are good and the estimated $\hat{\overrightarrow{y}}$ are significant according to $\overrightarrow{y}$. With these indicators, we can thus say that our data have a linear behaviour which follow a straight line with slope a. a is the most important parameter for our analysis, then b

can be neglected To produce the linear regression on our data and to compute R^2 and the p-value < 0.05 for the Wald test, we use the `Python` library `scipy.stats.linregress`.

B.2.2 Linear-Logarithmic Scales

Following the same process as above, excepted that we put the y-axis in logarithmic scale. If data $\overrightarrow{y}$ in function of $\overrightarrow{x}$ depict a straight line, we can write the relation as:

$$\log(y_i) = a \cdot x_i + b, \text{derived from} \tag{9}$$

$$y_i = 10^{(a \cdot x)} \cdot 10^b, \tag{10}$$

where, a is the slope or the increasing factor and b the intercept or an additive constant depending on the relations (9) and (10). In this case, the data describe an exponential shape. As this process is not used in this article, we don't develop completely this, it remains nevertheless important to pursue with the last case.

B.2.3 Double Logarithmic Scales

Considering the same method than the two aforementioned cases, we plot the data contained in $\overrightarrow{y}$ versus $\overrightarrow{x}$ on logarithmic x- and y-axis. In the case where the data behave itselves like a straight line we are then able to deduce the relation:

$$\log(y) = a \cdot \log(x) + b, \text{derived from} \tag{11}$$

$$y = x^a \cdot 10^b, \tag{12}$$

where a is the slope or the exponent and b is the intercept or a multiplicative constant according to the equations (11) and (12). From the relation (11), we can determine the estimated values for elements $\hat{y}_i$, a and b.

From here, we have to distinguish the two following cases:

$$\begin{cases} a \geq 0 \text{ or} \\ a < 0 \end{cases} \tag{13}$$

In the case of $a \geq 0$, we treat a power function given by the equation (12). The fit can be, as for the double linear case, obtained by performing the least-square linear regression. Then, the goodness-of-fit is given by the Pearson's coefficient of determination R^2 and the p-value < 0.05 for the Wald test. The results are computed the `Python` library `scipy.stats.linregress`.

In the case of $a < 0$, we are in presence of a power law. Due to the presence of the logarithm on both sides of (11), we cannot apply a least-square linear regression, because this method and the similar ones return systematic errors for common conditions. For this reason, it is impossible to trust the results [71]. Instead of this method, we estimate the parameters a with the method of maximum likelihood after a quadratic approximation to the log-likelihood to deal with our discrete values. In our analysis, the parameter b is not relevant and we don't need to estimate this. To determine if it really handles of a power law, we proceed to a Kolmogorov-Smirnov test, attempting to minimize the distance between the estimated parameters and our data. If the p-value from the Kolgomorov-Smirnov is smaller than the chosen threshold $\alpha = 0.05$, we can affirm that our data follow a power law [71]. Sometimes, the fits don't fit very well with a power law distribution that is why we have to investigate other heavy-tailed distributions like the

log-normal (L) or the Weibull (W) (i.e., stretched-exponential) distributions, for which we can define the goodness-of-fit with the previous Kolmogorov-Smirnov test and its p-value. However, with approximately same results, the power law is privileged because it is determined by one parameter instead of two parameters for the two aforementioned distributions.
The computations in this part have been widely inspired from the works of A. Clauset & al. and done with `Python` libraries such that `plfit` for the powerlaw and implemented according to the works of A. Clauset & al. for the other distributions [71].

B.2.4 Goodness-of-fits Summary

The results for the fits presented in this article (i.e., Figure 1, 2 and 3), as well as their goodness of are detailed in the below Table 5.

Fig	Model	Estimated Parameter(s)	Goodness-of-fit	p-value	Quality
1A	PL [a]	$\mu_{\text{att}} = 0.64(1)$	6.43×10^{-2}	$< 10^{-2}$	(***)
1B	PL [a]	$\mu_{\text{tags}} = 2.26(6)$	1.52×10^{-1}	$< 10^{-2}$	(***)
2A	PL [a]	$\mu_{\text{events}} = 0.54(4)$	1.51×10^{-1}	$< 10^{-2}$	(***)
2B	LR [b]	$\beta_{\text{orgs}} = 0.79(1)$	0.99	$< 10^{-2}$	(***)
3A	LR [b]	$\beta_{\Delta} = -0.93(1)$	0.97	$< 10^{-2}$	(***)
3B	LR [b]	$\beta^1_{\Delta} = (-6.32 \pm 0.91) \times 10^{-2}$	0.86	$< 10^{-3}$	(***)
		$\beta^2_{\Delta} = (-7.12 \pm 0.59) \times 10^{-2}$			

Table 5: Goodness-of-fits summary. The fits are generated by the Power Law [a] and ordinary least squares (OLS) Linear Regression [b] models. The goodness-of-fit are obtained with the Pearson's coefficient R^2 [a] and the p-value of a Wald test for the Linear Regression [a] model and with the Kolmogorov-Smirnov statistic test, also providing the p-value, for the Power Law [b] model. The results are computed with the `Python` libraries `scipy.stats.linregress`[a] and `plfit` [b].