https://doi.org/10.1093/cybsec/tyad021

Research paper

Efficient collective action for tackling time-critical cybersecurity threats

Sébastien Gillard [1,2,*], Dimitri Percia David [1,3,4], Alain Mermoud [3], Thomas Maillart [1,5]

[1] Information Science Institute, Geneva School of Economics & Management, University of Geneva , Boulevard du Pont-d'Arve 40, 1211 Geneva, Switzerland

[2] Chair of Defense Economics, Military Academy at ETH Zurich, Kaserne Reppischtal, 8903 Birmensdorf, Switzerland

[3] Cyber-Defence Campus, armasuisse Science and Technology, Feuerwerkstrasse 39, 3602 Thun, Switzerland

[4] Institute of Entrepreneurship & Management, University of Applied Sciences of Western Switzerland (HES-SO Valais-Wallis), Techno-Pôle 1, Le Foyer, 3960 Sierre, Switzerland

[5] Citizen Cyber Lab, University of Geneva, Avenue de Sécheron 15, 1202 Geneva, Switzerland

[*] Correspondence address: Kaserne Reppischtal, CH-8903 Birmensdorf. E-mail: sebastien.gillard@etu.unige.ch, sebgillard@gmail.com

Received 2 September 2022; revised 9 June 2023; accepted 12 September 2023

Abstract

The latency reduction between the discovery of vulnerabilities, the build-up, and the dissemination of cyberattacks 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 contribution dynamics on Malware Information Sharing Platform (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)

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 gathering, integrating, and sharing 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 an MISP[1] 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) its modularity increase performance.

The remainder of this article is organized as follows. The *Background* section covers the literature from the perspectives of social dilemma, productivity, and information integration in collective action in general and for cybersecurity. The *Data* section introduces MISP and presents the data. We then introduce the *Theoretical framework* followed by the research *Hypotheses* and *Methods*. We then present our *Results* and *Discuss* them before *Concluding*.

Background

Knowledge sharing in cybersecurity has been considered a crucial way to overcome a number of vulnerabilities [16] and threats [1]. It is, however, contingent to limiting factors, such as social dilemma on the one hand, and on the other hand, to enhancing return-on-scale effects. 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.

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 for larger groups, minor interests will be over-represented and diffuse majority interests trumped, due to a free-rider problem [18,19]. This free-riding effect is generally stronger for larger groups [20]. For instance, while Dejean et al. [21] 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. Beyond the increase of the community, due to the selfish behavior of the community members, the community efficiency to produce public goods depreciates [22]. 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 [21,23], Wikipedia [24], Stack Overflow [25], and communities of open-source software developers [26,27]. The Dejean et al. paradox can at least partially be resolved by considering that (i) the distribution of effort is highly skewed, with few contributors providing most effort, and (ii) the dynamics of contributions are highly nonlinear [27–29]. Taken together, these phenomena are associated with positive return-on-scale of production [27], which may be hindered by coordination costs [30]. Super-linear productivity has been debated at length in the organization and management sciences. Investigations of how the number of members and temporal dynamics of events generated can positively influence outputs in a 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 [31–34] by fine-tuning complementary mechanisms within the organization [35], which also foster innovation [36].

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 [37], so that knowledge appears as an utmost strategic resource [38]. For instance, there is growing evidence in science that greater teams create more impacting knowledge [39]. 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 [38,40]. With the emergence of virtual exchanges, firms are increasingly seen as distributed knowledge systems [41]. Yet, new interaction methods present various new constraints in terms 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 [42]. These platforms promote knowledge transfer by generating modular collaborative units [43]. 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 and Debreu [44], differentiation and integration have been a focal point in optimizing the structure of organizations [45,46]. 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 [40,47]. Recently, Engel and Malone used the theory of consciousness as information integration [48] to measure information integration computer systems and on collaborative platforms [47].

Collective action and information integration for cybersecurity

As early as 20 years ago, the first CERT and ISACs have been established as a central resource for sharing information on cybersecurity threats to critical infrastructures [49]. Nowadays, threat intelligence platforms help organizations aggregate, correlate, and analyze threat data from multiple sources in almost real-time to support defensive actions [50]. Further, open-source solutions have been proposed as a counterweight to cybercriminals successfully working together [5]. The swift evolution of cyber threats has forced organizations and governments to develop new strategies [51] in order to reduce the risks of security breaches [42]. 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 [52]. As such, information sharing can be considered as a marketplace on which transactions occur and knowledge is transferred [53]. However, human beings have a tendency to not optimize organizational goals [54] in the absence of selective incentives [55] and—in the case of collective action—might adopt a selfish behavior that is not conducive to the overall goal of sharing information [6]. As a consequence, cybersecurity professionals share probably less information than what would be socially desirable, leading to a knowl-

[1] MISP stands for "Malware Information Sharing Platform."

Table 1. Contributions of the most productive organizations.

Rank	Org ID	#users	#events contributed	Percentage of total events
1	1 092	8	7 682	19.38%
2	1 395	2	5 637	14.22%
3	1 960	3	3 214	8.11%
4	2	31	2 939	7.41%
5	1 857	3	1 411	3.56%
6	201	8	1 247	3.15%
7	1 713	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%

A total of 10 of 1908 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.

edge 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 when the situations get extraordinarily difficult, the behaviors tend to become unselfish, leading to an increase of contributions [56]. In this context, 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 [53]. 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.

Data

To understand the nuts and bolts of cybersecurity information sharing, we resort to *MISP Project*,[2] a popular open-source platform, which is used, e.g. by the North Atlantic Treaty Organization (NATO).[3] 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 at broad on which people can share, store, and collaborate on all sorts of incidents (e.g. COVID-19 MISP community),[4] 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. A detailed description of MISP is provided in Appendix A.

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 an MISP instance maintained by the Computer Incident Response Center Luxembourg (CIRCL), i.e. the Luxembourg CERT.

As of 8 February 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 10 November 2008. Table 1 shows the 10 most involved organizations. The number of events contributed by organizations is highly skewed. Indeed, Fig. 1a shows that the complementary cumulative distribution function (CCDF) 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 subgroups, which cannot 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 Fig. 2, follow power laws with exponents, respectively, $\mu_A = 0.64(1)$ (with an upper cut-off around $A_{\mathrm{upper}} = 10^5$) 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 (see Appendix A for an explanation), suggesting that either threat analysis is complicated or users tend to forget to formally close resolved events. The cumulative number of tags $N_{T,\mathrm{cum}} = 116\,407$ used is bigger than the amount of unique tags $N_{T_U} = 3\,786$. Thus, there is a massive reuse of already existing tags.

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 14 September 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 can see that the contribution effort is highly heterogeneous. It is also worth noting that event contributions started on 10 November 2008, long before the first organizations joined MISP CIRCL instance. This can be explained in the following way: organizations first ran their MISP instance locally before joining the MISP CIRCL community and sharing 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 vetted joining process, controlled by CIRCL.

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 Fig. 3b.We find that $\Delta\ t_C(t)$ follows an exponential decrease in phase. By applying a monthly binning and computing the mean value $\overline{\Delta t_C}$ for each bin, we see a first phase that extends from 2011 to Γ (i.e. the transition between the two phases at the end of 2019), which decreases slower than the second phase from Γ to today. By applying the linear regression on the data, according

[2] https://www.misp-project.org/

[3] https://misp.ncirc.nato.int

[4] https://covid-19.iglocska.eu

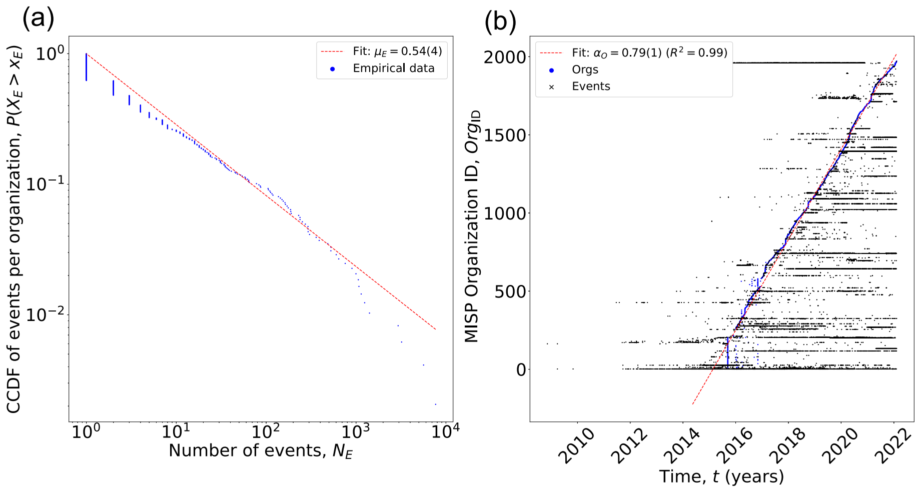

Figure 1. (a) CCDF of events per contributing organization, which is best described by a power law distribution $P(X_E > x_E) \sim 1/x_E{}^{\mu_E}$ with $\mu_E = 0.54(4)$. The fit and the goodness-of-fit, provided by the Kolmogorov–Smirnov statistics test, are obtained with the *Python* library plfit. (b) Curve of the joining organizations (in blue) has followed, after 14 September 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) and the distribution shows the heterogeneity of organizations efforts.

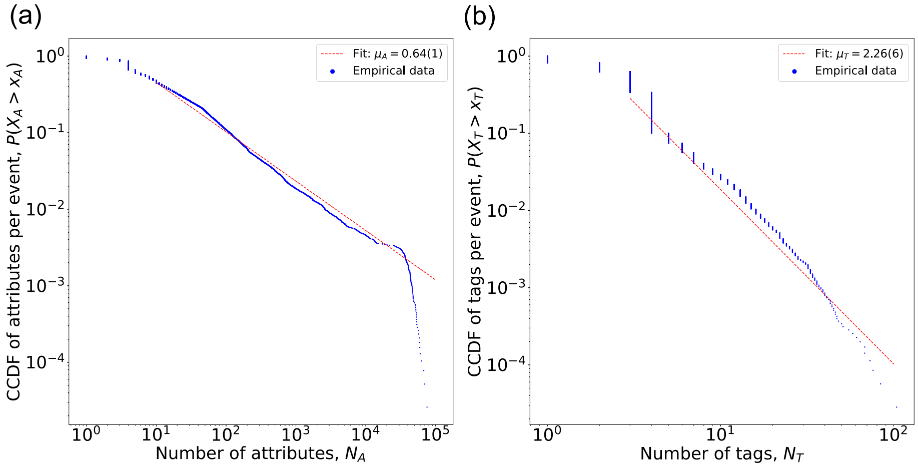

Figure 2. (a) CCDF of attributes encapsulated in an event, which is best described by a power law distribution $P(X_A > x_A) \sim 1/x_A{}^{\mu_A}$ with $\mu_A = 0.64(1)$. (b) CCDF of tags attached to an event, which is best described by a power law distribution $P(X_T > x_T) \sim 1/x_T^{\mu_T}$ with $\mu_T = 2.26(6)$. 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.

to the equation ($B4$), we confirm that Δt_C exhibits an exponential decrease:

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

where

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

The fit from the linear is of high quality as 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.

Theoretical framework

Collective action is thought to be a fundamental tool to overcome sprawling and possibly increasingly sophisticated time-critical cybersecurity threats [57–59]. Yet, despite numerous studies of online platforms fostering collective action [60,61], very little evidence has been uncovered linking the organization of collective action with group performance as an output. By investigating the MISP threat management platform run by the CIRCL, we have a unique opportunity to better understand how collective action is organized to tackle time-critical cybersecurity threats.

We posit that the performance of collective intelligence 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 in turn are associated with increased performance. We further posit

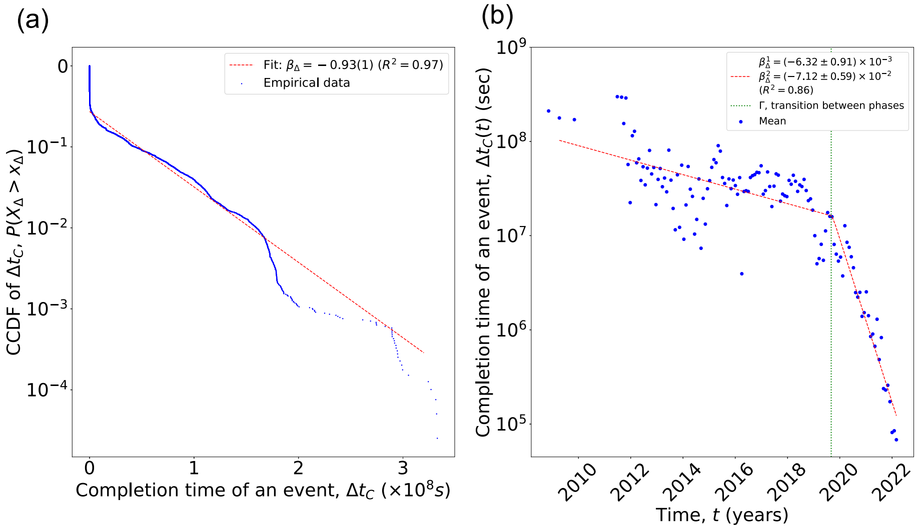

Figure 3. (a) CCDF of the completion time Δt_C, which is best described by a decreasing exponential distribution $P(X_\Delta < x_\Delta) \sim 10^{\beta_\Delta}$ with $\beta_\Delta = -0.93(1)$. (b) Completion time Δt_C of events over the time. The data (blue dots) represent the mean value of Δt_C binned monthly. The data depict an exponential decrease in two phases, changing at Γ (green-dotted line), fitted by linear regression (dashed red line), $\Delta t_C(t) \sim (-6.32 \pm 0.91) \times 10^{-2}$ for $t \in [2011, \Gamma[$ and $\Delta t_C(t) \sim (-7.12 \pm 0.59) \times 10^{-2}$ for $t \in [\Gamma, 2022]$ ($R^2 = 0.86$, 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.

that, in order to offset decreasing return-of-scale due to increased groups size and coordination costs [30], the organization of collective action must adapt in a modular way [62], as it has already been witnessed in several open-source projects [63,64].

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. *Hypotheses*). 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. *Methods* and *Results*).

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. We then define *event complexity*, the *overall collective action environment*, and how *knowledge is integrated*. We hypothesize that these three factors significantly influence collective action performance, in terms of improved completion time in characterizing threat events.

Event complexity hinders performance (H1)

First, events are not all equal: while many are fairly simple and require limited input in terms of attributes and of categorization with tags, others are more complex and require more effort. For each event, the information gathering process involves adding attributes or tags associated with an event both by the event creator and by other users (i.e. submission validated by the event creator). Attributes and tags may not exist in the MISP instance, and shall therefore be created by users (c.f. description of MISP in Appendix A), hence, requiring highly variable time and effort. Updated content shall then be shared with other users. The more complex, i.e. the more attributes and tags encapsulated in the corresponding event, the longer it takes to complete it. Figure 2a and 2b shows that the distributions of, respectively, attributes and tags are heavy tailed: while a majority of events have a limited number of attributes (respectively tags), some carry a large numbers of attributes (respectively tags), presumably affecting the time required to complete the characterization of an event. We therefore state Hypothesis 1 as follows:

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

To summarize plausible causality relationships between complexity and performance, we produce the causal diagram at Figure 4 although we don't intend to actually test causality.

Collective action improves performance (H2)

We consider how collective action at scale positively or negatively affects 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 [17,27–30]. 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. Figure 1 shows that, over time, organizations join the CIRCL MISP instance following a Poisson process. Upon joining, these organizations immediately benefit from the knowledge accumulated and shared by other organizations, which contributed early on and gained expertise. Also, similar or partially similar threats can be treated more efficiently over time, representing economies of scale. Conversely, for new users, learning and familiarizing with MISP may reduce the performance on the short term [8], while bringing long-term positive marginal gains. Finally, having more participants should bring marginally increasing performance. We therefore test the following hypothesis:

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

Yet, as already shown in reference [65], increased workload is likely to affect negatively performance, and thus, increase the expected completion time. When several threats occur, respectively, are

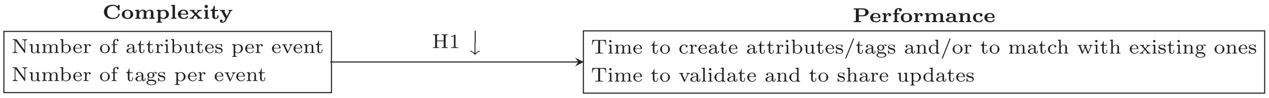
Complexity
Number of attributes per event
Number of tags per event
H1 ↓
Performance
Time to create attributes/tags and/or to match with existing ones
Time to validate and to share updates

Figure 4. Causal diagram relating complexity (left) and performance factors (right) ↓ indicates a negative effect of the former on the latter.

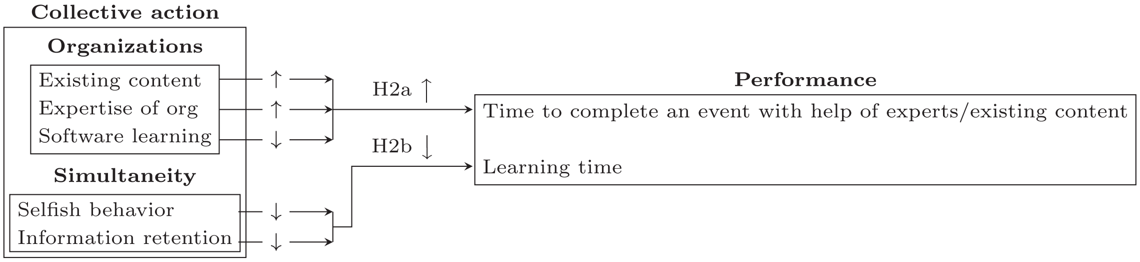

Figure 5. Causal diagram between the factors of collective action on the left-hand side and the factors of performance on the right-hand side. ↑ means an increase of performance, while ↓ means a decrease.

open, we posit that completion time of a focal event will necessarily be delayed as a result of priority queuing [66]. We could further hypothesize that given task overload, organizations may resort to increasingly selfish behavior, by focusing only on their threats and possibly by reducing their information sharing, hence, decreasing collective performance [6,22]. Therefore, our second hypothesis states:

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

The hypothesized causality between collective action and performance is shown in Fig. 5.

Knowledge integration increases performance (H3)

Having more contributors does not necessarily imply economies of scale [30]. Economies of scale may rather be generated by “the whole is more than the sum of its parts” mechanisms [27], which may stem from productive integration of information [47,67,68] as a single entity [27] or through the efficient communication of several modular subsystems [69,70], which, in turn, may even mitigate free riding [62]. Here, we recognize that the first form on knowledge integration occurs through (i) experience as *learning*, (ii) regular software use, (iii) repeated resolution processes of numerous events, and (iv) interactions with other participating organizations and their users within organizations [71]. 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 contribute to faster event resolution.*

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, 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: *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 [48,62,63]. As MISP clusters of events or attributes, called *Galaxies* , have been progressively introduced and developed on MISP CIRCL, we have an opportunity to test for modularity. Indeed, events or attributes can be attached to one or several Galaxies according to key values (e.g. their type, tags, category, distribution level, and/or threat level) associated with a given level of granularity, which is proportional to its prevalence in the MISP ecosystem (c.f. Appendix A). Therefore, a higher granularity refers to higher specificity, which, in turn, goes against performance. Conversely, a key value that would be too general, would not provide discriminate information, and therefore would go against performance [72]. Modularity provides a good balance between too fine-grained and too coarse-grained. We therefore formulate the following hypothesis:

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

Figure 6 shows the causality relationships between knowledge integration factors and 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. Figure 7 illustrates the expected influence of event complexity, collective action, and knowledge integration on the time needed to complete the characterization of threats events.

Methods

We proceed to validate our theory through the testing of three hypotheses, divided in six subhypotheses (c.f. *Hypotheses* section). 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. *Theoretical framework* section).

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. For each event, we define $\Delta t_{C,e}$ the completion

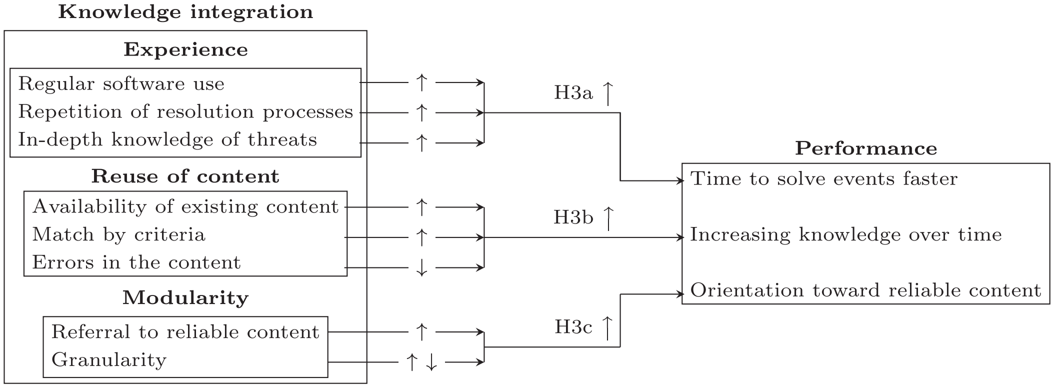
Knowledge integration
Experience
Regular software use
Repetition of resolution processes
In-depth knowledge of threats
↑
↑
↑
H3a ↑
Reuse of content
Availability of existing content
Match by criteria
Errors in the content
↑
↑
↓
H3b ↑
Modularity
Referral to reliable content
Granularity
↑
↑↓
H3c ↑
Performance
Time to solve events faster
Increasing knowledge over time
Orientation toward reliable content

Figure 6. Causal diagram between the factors of knowledge integration on the left-hand side and the factors of performance on the right-hand side. ↑ means an increase of performance, while ↓ means a decrease.

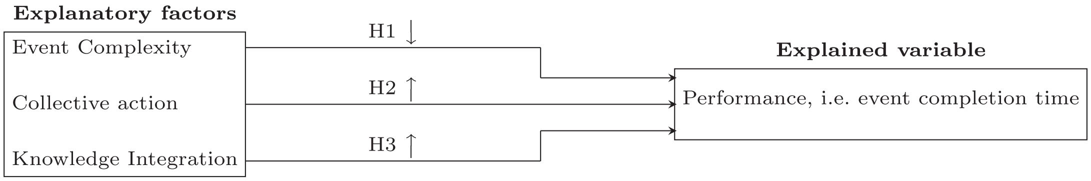

Figure 7. Causal diagram between the explanatory factors on the left-hand side and the explained variable of performance, namely event completion time on the right-hand side. ↑ means an increase of performance, while ↓ means a decrease.

time of events as ,

$$\Delta t_{C,e} = t_{f,e} - t_{c,e}, \qquad (3)$$

with $t_{c,e}$ is the event creation date and $t_{f,\,e}$ is 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 [73]. 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:

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

with:

(i) Δt_{C_e}: time completion for event e,
(ii) ζ: constant,
(iii) N_k: number of explanatory variables,
(iv) κ_k: autoregressor parameter corresponding to $Z_{k,\,e}$,
(v) $Z_{k,\,e}$: k-th explanatory variable for event e,
(vi) 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. It is therefore easy to consider all chosen independent variables. However, due to the heavy-tailed behavior of the variables and their difference of magnitude (see *Data*), we take the logarithm of the variables [74]. The results are indicated as a percentage change of $\Delta t_{C,\,e}$ when $Z_{k,\,e}$ varies by a certain percentage [74].

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

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

(i) $N_{O,\,e}$ 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**).
(ii) $E_{\text{sim},\,e}$ 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**) [65,66].

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

(i) $E_{C,\,e}$ 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**).

(ii) $I_{\%A,e}$ 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**).

(iii) $N_{G,e}$ 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**).

(iv) $N_{E_G,e}$ 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).

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

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

Model validation is performed as follows. When handling a multivariate regression, one must pay particular attention to multicollinearity 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 $\mathrm{VIF}_k = 1/(1 - {R_k}^2)$ and must be <10 [73]. 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 $\alpha = 0.05$ [73]. 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.

Results

In order to establish evidence of collective action as an efficient way for tackling time-critical cybersecurity threats, we have resorted to data from the MISP instance, which is run by the 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.

The regression results are shown in Table 3. Overall, the regression model is robust and explains 41.3% 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 multicollinearity 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 multicollinearity of the explanatory variables. We computed the VIF for each explanatory variables, which happens to be all smaller than 10. This implies that there is no evidence of multicollinearity 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 test. 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.

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. Organizations shall resort to collective action to gather information and integrate knowledge as two pillars of threat event characterization not only as attacks unravel, but also before attacks unravel [59]. 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 the objective function for 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 in information sharing on the studied MISP instance, 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 [48] as the collective action process at work to generate the "the whole is more than the sum of its parts" [27]. With Hypothesis 3, we tested and verified organizational learning, knowledge integration, and modularity as being 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. To the exception of SimEvents, i.e. the number of simultaneously open events on MISP CIRCL at the creation, which had to be excluded from the model, all our explanatory variables are signifi-

Table 2. Correlation matrix of dependent and explanatory variables.

	$\log(\Delta t_C)$	$\log(N_{A,e})$	$\log(I_{\%A,e})$	$\log(N_{T,e})$	$\log(E_{\mathrm{sim},e})$	$\log(N_{O,e})$	$\log(E_{C,e})$	$\log(N_{G,e})$	$\log(N_{E_G,e})$
$\log(\Delta t_C)$	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					
$\log(E_{\mathrm{sim},e})$	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 3. Results of the OLS regression.

Dep. variable		Completion time	
Method	OLS	F-stat.	2.251×10^3
No. observations	22 423	Prob (F-stat.)	0.00
R-squared	0.413	Log-likelihood	-5.030×10^4
		coeff	std error
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

The OLS regression is performed 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-value $< .1$: (*), $<.05$: (**), or $< .01$: (***) and the goodness-of-fit by the R-squared. The other added information are not necessary for the evaluation of the model.

cant. This 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. references [27–29], may 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 [66]. Indeed, when tackling large amounts of time-critical tasks, such as cybersecurity threats or incidents, contingencies necessarily appear [65], which may affect coordination between contributors, and performance as a result, 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 subcommunities 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, one may investigate kinship between threats, as many events 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. For that purpose, establishing causality between complexity, collective action, modularity factors, and performance would certainly help refine the entangled determinants of performance. Further, our results open 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 the best use of time as an increasingly critically scarce resource, especially in face of a looming tsunami of cybersecurity threats. Consequently, the increasing adoption of MISP, or equivalent information-sharing platforms by more and more of critical infrastructures and of organizations, as evident in our data, further emphasizes their relevance and, in turn, the positive externalities associated with more organizations joining. Notably, MISP's effectiveness in catering to the needs of small and medium businesses adds to the value proposition [75], even though the efficiency of information-sharing platforms for organizations remains to be tested against their size. By merging and modularizing diverse sources of information, such as different communities or instances, we anticipate an enhancement in time performance due to the improved situational awareness, ultimately optimizing information-sharing efficiency, and hence making information-sharing platforms increasingly attractive.

Conclusion

Information sharing in cybersecurity has become an increasingly common collective action practice. Yet, its benefits have so far remained unclear. We have investigated MISP, a commonly used open-

Table 4. Computation of the VIF for the explanatory variables of the econometric model.

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

The values of the VIF allows to detect the presence of multicollinearity between the considered variables. As all values VIF $<$ 10, there is no evidence of multicollinearity between the explanatory variables. These results validate the econometric model.

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.

Acknowledgments

The authors would like to thank Christophe Vandeplas and the Computer Incident Response Center Luxembourg (CIRCL) for their technical support and access to the data. The authors thank the WEIS’2021, the WEIS’2022 anonymous reviewers, as well as the participants of the WEIS’2022 conference in Tulsa, USA, for their comments. The authors also acknowledge useful comments from reviewers and the *Journal of Cybersecurity* editor.

Conflict of interest statement: None declared.

Funding

One of the authors (S.G.) acknowledges financial support from the Military Academy at ETH Zurich. Two authors (D.P.D. and T.M.) acknowledge financial support from the Cyber-Defence Campus (armasuisse Science and Technology).

Author contributions

Sébastien Gillard (Conceptualization, Data curation, Formal analysis, Investigation, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing), Dimitri Percia David (Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing), Alain Mermoud (Conceptualization, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing), and Thomas Maillart (Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing)

Data availability

The data underlying this article cannot be shared publicly due to compliant reasons and the limited access to the MISP CIRCL community. A request for access can be made to MISP CIRCL. The code is published open source at the following URL: https://github.com/technometrics-lab/9-Cyber-threats-Intelligence.

References

1. Sridhar K, Householder A, Spring J,. *et al.* Cybersecurity information sharing: analysing an email corpus of coordinated vulnerability disclosure. In: *The 20th Annual Workshop on the Economics of Information Security*, Online, 2021. https://weis2021.econinfosec.org/wp-content/uploads/sites/9/2021/06/weis21-sridhar.pdf.
2. Gal-Or E, Ghose A. The economic incentives for sharing security information. *Inf Syst Res* 2005;**16**:186–208.
3. Maillart T, Zhao M, Grossklaggs J. *et al.* Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *J Cybersecur* 2017;**3**:81–90.
4. Sridhar K, Ng M. Hacking for good: leveraging HackerOne data to develop an economic model of bug bounties. *J Cybersecur* 2021;**7**: 1–9.
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*, Vienna, Austria, 2016, 1–2.
6. Laube S, Böhme R. Strategic aspects of cyber risk information sharing. *ACM Comput Surv* 2017;**50**:1–36.
7. Brady RM, Anderson RJ, Ball RC. *Murphy’s Law, the Fitness of Evolving Species, and the Limits of Software Reliability*. Cambridge:University of Cambridge, Computer Laboratory, 1999. https://doi.org/10.48456/tr-471.
8. Stojkovski B, Lenzini G, Koenig V. *et al.* What’s in a cyber threat intelligence sharing platform?: a mixed-methods user experience investigation of MISP. In: *Annual Computer Security Applications Conference*, Virtual Event, 2021, 385–98.
9. Levy S. *Hackers: Heroes of the Computer Revolution*. New York, NY: Anchor Press/Doubleday, 1984.
10. Benkler Y. *The Penguin and the Leviathan: How Cooperation Triumphs over Self-Interest*. Random House Digital, Inc.: Currency, 2011.
11. Benkler Y. *The wealth of networks: how social production transforms markets and freedom*. New Heaven, CT: Yale University Press, 2006.
12. Katz ML, Shapiro C. Network externalities, competition, and compatibility. *Am Econ Rev* 1985;**75**:424–40.
13. Shapiro C, Varian HR. *Information Rules: A Strategic Guide to the Network Economy*. Boston: Harvard Business School Press, 1999.
14. Wagner C, Delaunoy A, Wagener G. *et al.* 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*, New York, NY: Association for Computing Machinery, 2016, 49–56.
15. Zibak A, Simpson A. Cyber threat information sharing: perceived benefits and Barriers. In: *Proceedings of the 14th International Conference on Availability, Reliability and Security*. Canterbury, CA: ACM, 2019. 1–9.
16. Mell P, Scarfone K, Romanosky S. Common vulnerability scoring system. *IEEE Secur Priv* 2006;**4**:85–9.
17. Olson M. *The Logic of Collective Action: Public Goods and the Theory of Groups, with a New Preface and Appendix*. Cambridge: Harvard University Press, 1971.

18. Anesi V. Moral hazard and free riding in collective action. *Soc Choice Welf* 2009;**32**:197–219.
19. Varian HR. System reliability and free riding. In: *Economics of Information Security*. Boston: Springer, 2004, 1–15.
20. Esteban J, Ray D. Collective action and the group size paradox. *Am Political Sci Rev* 2001;**95**:663–72.
21. Dejean S, Pénard T, Suire R. *Olson's Paradox revisited: an empirical analysis of incentives to contribute in P2P file-sharing communities*. Elsevier 100 South Clinton Avenue Suite 2407 Rochester, NY 14604: SSRN. 2010. https://doi.org/10.2139/ssrn.1299190.
22. Koutsoupias E, Papadimitriou C. Worst-case equilibria. In: C Meinel, S Tison (ed.), *STACS 1999*. Berlin: Springer, 1999, 404–13.
23. Asvanund A, Clay K, Krishnan R. *et al.* Empirical analysis of network externalities in peer-to-peer music-sharing networks. *Inf Syst Res* 2004;**15**:155–74.
24. 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, Vancouver BC*. New York, NY: Association for Computing Machinery, 2015, 1106–15.
25. Wang S, Lo D, Jiang L. An empirical study on developer interactions in Stack-Overflow. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing, Coimbra*. New York, NY: Association for Computing Machinery, 2013, 1019–24.
26. Hippel EV, Krogh GV. Open source software and the "private-collective" innovation model: issues for Organization Science. *Organ Sci* 2003;**14**:209–23.
27. 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* 2014;**9**:e103023. https://doi.org/10.1371/journal.pone.0103023.
28. Maillart T, Sornette D. Aristotle vs. Ringelmann: on superlinear production in open source software. *Physica A Stat Mech Appl* 2019;**523**:964–72.
29. Murić G, Abeliuk A, Lerman K. *et al.* Collaboration drives individual productivity. *Proc ACM Hum-Comput Interact* 2019;**3**:1–24.
30. Scholtes I, Mavrodiev P, Schweizer F. From Aristotle to Ringelmann: a large-scale analysis of team productivity and coordination in open source software projects. *Empir Softw Eng* 2016;**21**:642–83.
31. Tziner A, Eden D. Effects of crew composition on crew performance: does the whole equal the sum of its parts? *J Appl Psychol* 1985;**70**:85–93.
32. Sundstorm E, De Meuse KP, Futrell D. Work teams: applications and effectiveness. *Am Psychol* 1990;**45**:120–33.
33. Cohen SG, Bailey DE. What makes teams work: group effectiveness research from the shop floor to the executive suite. *J Manage* 1997;**23**:239–90.
34. Neumann GA, Wright J. Team effectiveness: beyond skills and cognitive ability. *J Appl Psychol* 1999;**84**:376.
35. 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. *J Manage* 2010;**36**:207–33.
36. Sacramento CA, Chang MWS, West MA. Team innovation through collaboration. In: MM Beyerlein, ST Beyerlein, FA Kennedy (eds.), *Innovation Through Collaboration*. Bingley: Emerald Group Publishing Limited, 2006, 81–112.
37. Nonaka I. A dynamic theory of organizational knowledge creation. *Organ Sci* 1994;**5**:14–37.
38. Grant RM. Prospering in dynamically-competitive environments: organizational capability as knowledge integration. *Organ Sci* 1996;**7**:375–87.
39. Wuchty S, Jones BF, Uzzi B. The increasing dominance of teams in production of knowledge. *Science* 2007;**316**:1036–9.
40. Lawrence PR, Lorsch JW. Differentiation and integration in complex organizations. *Adm Sci Q* 1967;**12**:1–47.
41. Majchrzak A, Griffith TL, Reetz DK. *et al.* Catalyst organizations as a new organization design for innovation: the case of hyperloop transportation technologies. *Acad Manag Discov* 2018;**4**:472–96.
42. Safa NS, Von Solms R. An information security knowledge sharing model in organizations. *Comput Hum Behav* 2016;**57**:442–51.
43. Mockus A, Fielding RT, Herbsleb J. A case study of open source software development: the Apache server. In: *Proceedings of the 22nd international conference on Software engineering*, Limerick Ireland, New York, NY: Association for Computing Machinery, 2000, 263–72.
44. Arrow KJ, Debreu G. Existence of an equilibrium for a competitive economy. *Econometrica* 1954;**22**:265–90.
45. Ravasi D, Verona G. Organising the process of knowledge integration: the benefits of structural ambiguity. *Scand J Manag* 2001;**17**:41–66.
46. Huang JC, Newell S. Knowledge integration processes and dynamics within the context of cross-functional projects. *Int J Proj Manag* 2003;**21**:167–76.
47. Engel D, Malone TW. Integrated information as a metric for group interaction. *PLoS One* 2018;**13**:e0205335. https://doi.org/10.1371/journal.pone.0205335.
48. Tononi G. Consciousness and complexity. *Science* 1998;**282**:1846–51.
49. Zheng DE, James A. *Cyber Threat Information Sharing: Recommendations for Congress and the Administration*. Washington DC: Center for Strategic & International Studies, 2015
50. He M, Devine L, Zhuang J. Perspectives on cybersecurity information sharing among multiple stakeholders using a decision-theoretic approach: cybersecurity information sharing. *Risk Anal* 2018;**38**:215–25.
51. Meier R, Scherrer C, Gugelmann D. *et al.* FeedRank: a tamper-resistant method for the ranking of cyber threat intelligence feeds. In: *10th International Conference on Cyber Conflict (CyCon), Tallinn*. Tallinn, Estonia: NATO Cooperative Cyber Defence Centre of Excellence, 2018, 321–44.
52. Mermoud A, Keupp MM, Huguenin K. *et al.* To share or not to share: a behavioral perspective on human participation in security information sharing. *J Cybersecur* 2019;**5**:1–13.
53. Percia David D, Keupp MM, Mermoud A. Knowledge absorption for cybersecurity: the role of human beliefs. *Comput Hum Behav* 2020;**106**:106255. https://doi.org/10.1016/j.chb.2020.106255.
54. Mermoud A, Keupp MM, Percia David D. Governance models preferences for security information sharing: an institutional economics perspective for critical infrastructure protection. In: E Luiijf, I Žutautaitė, B Hämmerli (eds.), *Critical Information Infrastructures Security (CRITIS) 2018*. Cham: Springer, 2019, 179–90.
55. Oliver P. Rewards and punishments as selective incentives for collective action: theoretical investigations. *Am Journal of Sociol* 1980;**85**:1356–75.
56. Hirshleifer J. From weakest-link to best-shot: the voluntary provision of public goods. *Public Choice* 1983;**41**:371–86.
57. Mermoud A. Three articles on the behavioral economics of security information sharing: a theoretical framework, an empirical test, and policy recommendations. Ph.D. Thesis, University of Lausanne, Faculty of Business and Economics, 2019.
58. Bouwman X. Governance of cybersecurity communities: understanding threat intelligence sharing as a collective action problem through incentivization of the National Detection Network. *Master Thesis*, Delft University of Technology Policy and Management, 2018.
59. Wagner TD, Mahbub K, Palomar E. *et al.* Cyber threat intelligence sharing: survey and research directions. *Comput Secur* 2019;**87**:101589. https://doi.org/10.1016/j.cose.2019.101589.
60. Bouwman X, Le Pochat V, Foremski P. *et al.* Helping hands: measuring the impact of a large threat intelligence sharing community. In: *31st USENIX Security Symposium (USENIX Security 22), Boston*. Berkeley, CA: USENIX Association, 2022, 1149–65.
61. McColl RC, Ediger D, Poovey J. *et al.* A performance evaluation of open source graph databases. In: *Proceedings of the First Workshop on Parallel Programming for Analytics Applications, Orlando*. New York, NY: Association for Computing Machinery, 2014, 11–8.
62. Baldwin CY, Clark KB. The architecture of participation: does code architecture mitigate free riding in the open source development model? *Manage Sci* 2006;**52**:1116–27.
63. Narduzzo A, Rossi A. The role of modularity in free/open source software development. In: S Koch (ed.), *Free/Open Source Software Development*. Hershey: IGI Global, 2005, 84–102.
64. Langlois RN, Garzarelli G. Of hackers and hairdressers: modularity and the organizational economics of open-source collaboration. *Ind Innov* 2008;**15**:125–43.

65. 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. 2018, 18–9. https://weis2018.econinfosec.org/wp-content/uploads/sites/5/2016/09/WEIS_2018_paper_50.pdf.
66. Maillart T, Sornette D, Frei S. *et al.* Quantification of deviations from rationality with heavy tails in human dynamics. *Phys Rev E Stat Nonlin Soft Matter Phys* 2011;**83**:056101. https://doi.org/10.1103/PhysRevE.83.056101.
67. Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol* 2014;**10**:e1003588. https://doi.org/10.1371/journal.pcbi.1003588.
68. Malone TW. *Superminds: How Hyperconnectivity is Changing the Way We Solve Problems*. New York, NY: Simon and Schuster, 2018.
69. Barrett AB, Seth AK. Practical measures of integrated information for time-series data. *PLoS Comput Biol* 2011;**7**:e1001052. https://doi.org/10.1371/journal.pcbi.1001052.
70. Baldwin C, Clark K. *Design Rules: the Power of Modularity (Vol. 1)*. Cambridge: The MIT Press, 2000.
71. Argote L, Miron-Spektor E. Organizational learning: from experience to knowledge. *Organ Sci* 2011;**22**:1123–37.
72. West S, Ali H. On the impact of granularity in extracting knowledge from bioinformatics data. In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - BIOINFORMATICS*, Rome, Italy: SciTePress, 2016;**3**:92–103.
73. Asteriou D, Hall SG. *Applied Econometrics*, 4th edn. London: Bloomsbury Publishing, 2021
74. Benoit K. Linear regression models with logarithmic transformations. *London School of Economics* 2011;**22**:23–36.
75. Van Haastrecht M, Golpur G, Tzismadia G. *et al.* A shared cyber threat intelligence solution for SMEs. *Electronics* 2021,**10**:2913
76. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. *SIAM Rev* 2009;**51**:661–703.

Appendix 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. An 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*,[5] 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, and 4: Undefined), analysis level (i.e. 0: Initial, 1: Ongoing, and 2: 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. These events or attributes can be attached to one or more clusters named "Galaxies" according to their key values (e.g. their type, tags, category, distribution level, and/or threat level).

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 database. However, let us add 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 two, meaning "complete."

To retrieve the data, we have followed the user guide[6] 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[7]), as well as the name and the ID of the concerned organizations. 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.

Appendix B. Exploratory analysis of the dataset

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:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt. \tag{B1}$$

Then, thanks to the formula ($B1$), the complementary cumulative distribution function (CCDF) $\overline{F_X}(x)$ can be written as follow:

$$\overline{F_X}(x) = 1 - F_X(x) = P(X > x). \tag{B2}$$

This CCDF provides a rank ordering of the selected variables.

[5] https://git-scm.com/

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

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

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 show an approximate straight line, that means that each element y_i of the vector $\overrightarrow{y}$ is given by the relation:

$$y_i = a \cdot x_i + b, \tag{B3}$$

where a is the slope of the straight line and b, its intercept. Thanks to the relation ($B3$), 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 behavior, 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 $< .05$ 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:

$$\log(y_i) = a \cdot x_i + b, \text{ derived from} \tag{B4}$$

$$y_i = 10^{(a \cdot x)} \cdot 10^b, \tag{B5}$$

where a is the slope or the increasing factor and b the intercept or an additive constant depending on the relations ($B4$) and ($B5$). 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:

$$\log(y) = a \cdot \log(x) + b, \text{ derived from} \tag{B6}$$

$$y = x^a \cdot 10^b, \tag{B7}$$

where a is the slope or the exponent and b is the intercept or a multiplicative constant according to the equations ($B6$) and ($B7$). From the relation ($B6$), we can determine the estimated values for elements $\hat{y}_i$, a, and b.

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

$$\begin{cases} a \geq 0 \text{ or} \\ a < 0. \end{cases} \tag{B8}$$

In the case of $a \geq 0$, we treat a power function given by the equation ($B7$). 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 $< .05$ 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 ($B6$), 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 [76]. 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 [76]. 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 Clauset et al. and done with *Python* libraries such that plfit for the powerlaw and implemented according to the works of Clauset et al. for the other distributions [76].

B.2.4. Goodness-of-fits summary

The results for the fits presented in this article (i.e. Figs 1, 2, and 3), as well as their goodness of are detailed in the below Table B1.

Table B1. Goodness-of-fits summary.

Fig.	Model	Estimated parameter(s)	Goodness-of-fit	p-value
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]	$\beta^1_{\Delta} = (-6.32 \pm 0.91) \times 10^{-2}$ $\beta^2_{\Delta} = (-7.12 \pm 0.59) \times 10^{-2}$	0.86	$<10^{-3}$

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^{2a} 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].