Are You at Risk ?
Profiling Organizations and Individuals Subject to Targeted Attacks

Olivier Thonnard, Leyla Bilge, Anand Kashyap, and Martin Lee

Symantec Research Lab,
{Olivier_Thonnard,Leylya_Yumer,Anand_Kashyap,Martin_Lee}@symantec.com

Abstract. Targeted attacks consist of sophisticated malware developed by attackers having the resources and motivation to research targets in depth. Although rare, such attacks are particularly difficult to defend against and can be extremely harmful. We show in this work that data relating to the *profiles* of organisations and individuals subject to targeted attacks is amenable to study using epidemiological techniques. Considering the taxonomy of Standard Industry Classification (SIC) codes, the organization sizes and the public profiles of individuals as potential risk factors, we design case-control studies to calculate *odds ratios* reflecting the degree of association between the identified risk factors and the receipt of targeted attack. We perform an experimental validation with a large corpus of targeted attacks blocked by a large security company's mail scanning service during 2013-2014, revealing that certain industry sectors and larger organizations –as well as specific individual profiles – are statistically at elevated risk compared with others. Considering targeted attacks as akin to a public health issue and adapting techniques from epidemiology may allow the proactive identification of those at increased risk of attack. Our approach is a first step towards developing a predictive framework for the analysis of targeted threats, and may be leveraged for the development of cyber insurance schemes based on accurate risk assessments.

Keywords: Targeted attacks, epidemiology, risk analysis, cyber insurance

1 Introduction

In recent years, we observe a dramatic increase on targeted attacks [31]. Publicised attacks, such as Shamoon among many others, show how such attacks may cause considerable disruption and financial harm to Internet users. Unfortunately, the traditional malware defense mechanisms are not adequate to detect such attacks. Therefore, organisations need to remain vigilant for the presence of such malware within their systems. However, targeted attacks remain rare. Many organisations may not need to expend significant resources in attempting to detect threats to which they may never be exposed. Similarly, some organisations may be in imminent danger of being attacked, yet have little security infrastructure in place to detect and reorganisations [2].

Anecdotal evidence from publicised targeted attacks hints that certain industry sectors and certain employee profiles may be at heightened risk of attack. For instance, the Nitro campaign was associated with the chemical industry [8], Luckycat affected with the shipping and defence industries, among others [32, 35]. Most infamously of all, Stuxnet [11, 30] targeted a specific industrial control system operating within the energy sector.

It may be intuitive that critical industries such as defence and chemical industrial sectors are more prone to targeted attacks than other sectors. However, this is not sufficient to assess the *level of risk* a targeted cyber attack may pose to a given organization. Identifying the specific industrial sectors and the specific user profiles which may be at heightened risk requires more than intuition and assumption.

One method of identifying high risk sectors and employees is to consider targeted attacks as akin to a *public health issue*. Epidemiological science has developed various statistical techniques for discovering associations between lifestyle or genetic factors, and adverse health outcomes. Once predisposing factors for diseases have been discovered, campaigns can be instigated to educate those affected of their particular risk and how this risk can be mitigated.

Case-control studies are commonly used within health-care research to identify *risk factors* within a population that are associated with developing a disease. A risk factor is a binary variable that can be observed within members of a population to test if the risk factor is associated with a health outcome. Such factors may be lifestyle factors or the prior exposure to an environmental pollutant. An advantage of case-control studies is that they can be *retrospective* by design, and used to investigate groups already affected by an issue. In such a study the incidence of many potential risk factors within the members of a subject group known to by afflicted by a disease (the cases) are compared with those of a second similar group that does not have the disease (the controls). Risk factors can then be identified through their statistical association with the disease using a well characterised methodology.

In this paper, we show that it is possible to conduct a rigorous case-control study in which the detection of being sent a targeted attack is considered as the *outcome*. Such a study can identify the potential risk factors, such as the activity sector and size of an organisation or job characteristics of an employee, that might be associated with being subject to a cyber attack. The identification of these risk factors allows organisations to assess their risk level and take proactive measures to mitigate or at least to control this risk. Moreover, it could be also beneficial for cyber insurance systems that suffer from elaborated risk assessment methodologies for assigning accurate insurance ratings to the organizations or individuals.

By applying this approach to a large corpus of targeted attacks blocked by e-mail scanning service of a large security company, we show that larger organizations and specific industry sectors, such as *National Security and International Affairs*, or the *Energy* and *Mining* sectors, are strongly associated with the risk of receiving targeted attacks and hence can be considered of being at higher risk than other industry sectors. Furthermore, incorporating data obtained from LinkedIn about the employees that were targeted in these companies, we have found that not only Directors or high-level executives are likely to be targeted, but other specific job roles such as Personal Assistants are even more at risk of targeted attack compared to others.

The rest of this paper is organized as follows. In Section 2 we discuss related works and position our contribution. Section 3 gives some background on epidemiology concepts used in this work and describes the design of our case-control study. We present and discuss our experimental results obtained with a large corpus of targeted attacks in Section 4. Section 5 concludes the paper.

2 Related work

The use of epidemiology concepts in computer security is not novel. However, we note that previous work has mainly focused on malware epidemics and computer worm epidemiology, *i.e.*, developing analytical models for computer virus propagation and worm outbreaks within vulnerable populations in the Internet.

In the years 1991 to 1993, pioneering work by Kephart *et al.* extended classical epidemiological models with directed-graphs to model the behavior of computer viruses and determine the conditions under which epidemics are more likely to occur [16, 15, 17]. Follow-up work relied mostly on the classical Susceptible → Infected → Recovered (SIR) epidemiology model – developed by Kermack-McKendrick for modeling infectious disease epidemics [12, 10] – to measure the total infected population over time during an Internet worm outbreak. Examples of such studies include various analyses of significant worm outbreaks such as CodeRed [39, 29, 22] and Slammer epidemics [21]. In [38] the authors examined other types of propagation like email worms (*e.g.*, the Witty worm, also studied by Shannon and Moore in [28]).

Another closely related research area has looked more specifically at *response* technologies for computer virus propagation and Internet worm epidemics. In early work Wang *et al.* investigated the impact of immunization defenses on worm propagation [36]. Subsequently Zou *et al.* developed a more accurate *two-factor* worm model that includes the dynamic aspects of human countermeasures and the variable infection rate. Then Moore *et al.* investigated methods for Internet quarantine and have set up in [23] requirements for containing self-propagating code. Later, Zou *et al.* proposed a dynamic quarantine defense method inspired by methods used in epidemic disease control and evaluated the approach through simulation of three Internet worm propagation models [40].

Follow-up work by Porras *et al.* studied a hybrid quarantine defense approach by looking at potential synergies of two complementary worm quarantine defense strategies under various worm attack profiles [26]. Finally, Dagon *et al.* extended the classical SIR model and created a diurnal model which incorporates the impact of time zones on botnet propagation to capture regional variations in online vulnerable populations [9].

The analysis of the current state of the art in computer epidemiology reveals clearly a lack of research in the field of developing predictive analytics for more advanced threats, such as *targeted attacks*. Our study is a first step towards considering such attacks as a public health issue amenable to epidemiological studies. However, the techniques required for modeling targeted threats are different from those used previously in computer worm epidemics. Targeted trojans differ from other common forms of malware in that the attacker researches and selects potential targets to which the attacks are directed. It is not necessarily the behavior of the individual or the vulnerable status of a system that leads to exposure to malware, but rather something specific to the individual (or the organization he belongs to) that leads them to come to the attention of attackers.

Closer to our research is the work done by Carlinet *et al.* in [7], where the authors have used epidemiological techniques to identify risk factors for ADSL users to generate malicious traffic. The study identified that the use of web and streaming applications and use of the Windows operating system were risk factors for apparent malware infection. Recently, Bossler and Holt conducted a similar study looking at factors associated with malware infection, finding that media piracy was positively correlated with infection, as was "associating with friends who view online pornography", being employed and being female [6]. In [18], the author conducted a preliminary case-control study on academic malware recipients, using the HESA JACS coding of academic subjects to investigate the relationship between research interests and the receipt of targeted attacks. While the methodology used in [18] was similar as the one used in this paper, the study was performed on a limited scale (with only academic recipients) and at the level of individuals instead of organizations. A recent study by Levesque et. al [19] analyzes the interactions between users, AV software and malware leveraging studies widely adopted in clinical experiements. Finally, in [33] the authors provided an in-depth analysis of targeted email attacks and the associated malware campaigns as orchestrated by various teams of attackers.

The main contribution of this paper is to show how statistical techniques borrowed from the public health community may be effectively used to derive putative *risk factors* associated with the profiles of organizations likely to be at an increased risk of attack because, *e.g.*, of their activity sector or organizational size. further extended to develop a predictive framework in which the degree of risk of being attacked could be evaluated even more precisely by combining an extended set of relevant factors pertaining to the profile of organisations or the individuals belonging to them.

3 Methodology

3.1 Epidemiology concepts

In epidemiology, a commonly used method for determining if a factor is associated with a disease consists in performing a retrospective *case-control* study [20] in which a population known to be afflicted with a disease is compared to a similar population that is unafflicted. For example, the risk of tobacco use on lung cancer is assessed by comparing the volume of tobacco use of the population that is afflicted with lung cancer(1) with the disease-free (0) population [1]. Note that, while a case control study can be effective at identifying *risk factors*, it cannot impart information about the likelihood of an outcome, since we are pre-selecting an afflicted group rather than searching for the affliction in a random population [27].

If we now substitute "afflicted with a disease" with "encountered a targeted attack", we can use these same epidemiology techniques to identify risk factors that are associated with targeted attacks, and leverage this knowledge to identify the characteristics of risky organisations and individuals.

To interpret the results of a case control study, we need to calculate the *odds ratio* (OR) that is a measure of the degree of association between a putative risk factor and an outcome – the stronger the association, the higher the odds ratio [4]. Suppose that p_{11} is the probability of afflicted entities possessing the risk factor and p_{01} is the probability of afflicted entities not possessing the risk factor. Similarly, p_{10} is the probability of unafflicted individuals within the control group also possessing the risk factor, and p_{00} is the probability of unafflicted individuals in the control group not possessing the risk factor. The odds ratio (OR) is then calculated as:

$$OR = \frac{p_{11} \times p_{00}}{p_{10} \times p_{01}}$$

Empirical measurements that sample populations have an inherent rate of error. To reach the test of being in excess of 95% certain that any risk factor that we have identified is an actual risk factor and not an artefact of our test, we need to calculate the standard error associated with our sampling using:

$$SE(\log_e OR) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}$$

where n_{11} is the number of afflicted entities possessing the risk factor, n_{10} is the number of afflicted entities without the risk factor, n_{01} is the number of control unafflicted entities with the risk factor, and n_{00} is the number of control unafflicted entities without the risk factor. The upper and lower 95% confidence values (W,X) for the natural logarithm of the odds ratio are then calculated as:

$$\begin{cases} W = \log_e OR - (1.96\, SE(\log_e OR)) \\ X = \log_e OR + (1.96\, SE(\log_e OR)) \end{cases}$$

The 95% confidence interval for the odds ratio is the exponential of W and X, e^W to e^X. In order for a putative risk factor to be positively associated with an outcome with greater than 95% probability, both e^W and e^X should be greater than 1.0. For the risk factor to be negatively correlated with the outcome, both e^W and e^X should be less than 1.0 [24].

3.2 Case-control study design

As our main goal is to discover risk factors for being victims of targeted attacks, our case-control study consists in analyzing organizations and individuals that encountered e-mail based targeted attacks, and compare them with the ones that did not. Note that there exists other means of exploitation to compromise the targets. Nevertheless, the data we use for this study comprises of only attacks that spread through e-mails and therefore, we focus on finding the risk factors for e-mail based targeted attacks (also referred to as *spear-phishing* emails).

Organization level. For this study, the *afflicted* population is composed of 3,183 organisations that was identified by a large security company's mail scanning service as being victims of at least one e-mail based targeted attacks. Note that the process of finding the victims involves careful manual effort.

In case-control studies one crucial step is to prepare the *control* group selectively. Ideally, the control group should be as large as possible, to increase the number of subjects in the study to act to reduce the calculated standard error values and increase the power of the study. However, this also acts to increase the resources necessary to conduct the study. Typically the size of the control group should be in the order of at least four times larger than the afflicted group [14]. Therefore, we constructed our control group from 15,915 organisations through *random* selection from 37,213 organisations that received traditional malware attacks during 2013. It is worth noting that random sampling is usually considered as the best sampling approach in order to avoid any bias in the representativeness of the control population [5].

We performed a case-control study with two different organization-level features to understand whether they could be one of the risk factors for targeted attacks. Motivating by the fact that a majority of notable targeted attacks seem to be launched against organizations operating in specific sectors, we chose first to investigate the industry sector of the organizations that are part of our customer base. We identify the sector of the organizations in our control group by leveraging both internal data sources (*e.g.*, marketing and customer data) as well as publicly available sources providing the Standard Industry Classification (SIC) (such as `www.leadferret.com` and `www.companycheck.co.uk`) for customers and organisations lacking such detailed information. For this study, we restrict ourselves to the primary Standard Industry Classification (SIC) *2-digit* code [25], and leave the analysis of the more detailed SIC 4-digit classification as future work.

The second feature we used in our case-control study is the size of the organization in terms of number of employees. Organisations were divided into 4 size groups according to the number of employees that used the large security company's mail scanning service. Therefore, the size we estimated for the organizations might be smaller that organization's actual size. Nevertheless, these numbers should reflect quite accurately the organisation sizes, and more importantly, the relative differences in size among different organisations.

Individual level. In addition to the organizational-based risk factors, we conducted a case control study to investigate individual-based risk factors that are associated with targeted attacks. While the afflicted group consists of the individuals that received e-mail based targeted attacks, the control group is composed of individuals that are in the same organizations and never received targeted attacks. The individual-based features are computed from information that can be obtained from the corresponding LinkedIn profiles of the individuals.

From the 3,183 afflicted organizations that we studied in the previous section we selected organizations that allow us focus only on organizations that have enough data (at least 100 afflicted and 300 unafflicated employees) for accurate statistical inference and that have the appropriate mailing convension (`<firstname>(.|_)<lastname>@<copanydomain>` or `<lastname>(.|_)<firstname>@<copanydomain>`) for their employees such that it is possible to collect her/his LinkedIn profiles information using the LinkedIn search API. Following these two criteria, we were able to obtain LinkedIn profiles of 4150 afflicted individuals and 12031 unafflicated individuals from 82 organizations.

The most insightful features we were able to extract from the LinkedIn profiles of the users are as follows:

- *Job Level:* The job level indicates an employee's position in an organization's hierarchy. We have considered 7 job levels: *Intern, Temporary Workers, Support Staff, Individual Contributors, Managers, Directors, and Executives.*
- *Job Type:* The job type indicates the job function performed by an employee in an organization's hierarchy. We have considered 9 job types: *Operations, Engineering, IT, Sales and Marketing, HR, Finance, Legal, QA, and Research.*
- *Location:* The location field in LinkedIn is typically free form text (e.g., San Francisco Bay Area, Greater Mumbai Area, etc.), and may not contain the name of a country. We look up the name of the country by searching the location string on Google and Wikipedia.
- *Number of LinkedIn Connections:* We divide the number of LinkedIn connections into four groups: *0, 1-250, 250-500, and 500+.*

3.3 Validation with Chi-square test

To validate the odds ratio results, we performed a *chi-square* test, which is commonly used in statistics to test the significance of any association in a contingency table containing frequencies for different variables. More specifically, chi-square allows to test the *null hypothesis* that there is no significant association between two (or more) variables, the alternative being that there is indeed an association of any kind [13, 5].

In this case, we apply the chi-square test to measure the association between the variables *afflicted* versus *unafflicted* on one hand, and *has factor 'x'* versus *don't have factor 'x'* on the other hand. For example, *SIC code 'x'* versus *other sector.* The same test can be performed using any other risk factor as variable, instead of the SIC code. The test consists then in comparing the *observed* frequencies (O) with the *expected* frequencies (E) obtained by using the marginal totals for rows and columns. If the two variables are not associated, the expected and observed frequencies should be close to each other and we should not observe any significant difference between the two, any discrepancy being due to merely random variation.

The chi-square test allows us to evaluate the difference between expected and observed frequencies: we just need to calculate the sum of the squared differences between the observed and expected values (*i.e.*, $\sum(O-E)^2/E$), and then compare the final value to the distribution of the chi-square statistic with $(r-1)(c-1)$ degrees of freedom, where r is the number of rows and c the number of columns (*i.e.*, in this case we have only 1 degree of freedom). As a result, we obtain a probability value p that allows us to accept or reject the null hypothesis with a certain confidence level. In most cases, we consider $p < 0.05$ as a significant probability to safely reject the null hypothesis, and thus conclude that there is good evidence of a relationship between the two variables.

By repeating this statistical test for each risk factor under test, we calculate the chi-square p-value to evaluate the significance of any association between a specific

Table 1: Odds ratios (OR) for the sectors that the highest and lowest association with targeted attacks.

SIC2	SIC2 Description	Odds ratio	Confidence interval	χ^2 p-val
97	National Security and International Affairs	22.55	4.87 - 55.56	< .001
40	Railroad Transportation	11.26	1.25 - 44.93	0.011
14	Mining and Quarrying of Nonmetallic Minerals, Except Fuels	5.01	1.51 - 24.80	0.033
96	Administration of Human Resource Programs	4.69	1.68 - 23.31	< .001
10	Metal Mining	4.10	1.69 - 9.90	0.001
44	Water Transportation	3.77	1.61 - 8.95	0.001
92	Justice, Public Order, And Safety	3.75	2.02 - 53.72	< .001
96	Administration Of Economic Programs	3.64	1.52 - 45.49	0.003
28	Chemicals and Allied Products	2.92	2.14 - 3.98	< .001
29	Petroleum Refining and Related Industries	3.12	1.62 - 9.57	0.040
13	Oil And Gas Extraction	2.87	1.55 - 6.59	0.001
60	Depository Institutions	2.74	1.98 - 3.80	< .001
37	Transportation Equipment	2.17	1.40 - 3.37	0.001
49	Electric, Gas, And Sanitary Services	2.12	1.43 - 3.24	< .001
48	Communications	1.58	1.10 - 2.27	0.019
27	Printing, Publishing, And Allied Industries	1.50	1.12 - 2.01	< .001
65	Real Estate	0.75	0.58 - 0.97	0.020
64	Insurance Agents, Brokers and Service	0.62	0.39 - 0.98	0.031
81	Legal Services	0.58	0.43 - 0.77	< .001
99	Nonclassifiable Establishments	0.34	0.31 - 0.38	< .001
17	Construction - Special Trade Contractors	0.24	0.14 - 0.41	< .001
07	Agricultural Services	0.18	0.04 - 0.75	0.007

factor and the fact of receiving targeted attacks within the selected population. As shown in our experimental results (Section 4), it enables us to validate the statistical significance of Odds Ratios for any association discovered between a risk factor and the receipt of targeted attacks. Note, however, that chi-square is not an index of the *strength* of the association between the tested variables. Also, certain categories may be excluded from the test because of a too small sample size. The conventional criterion for a chi-square test to be valid is that at least 80% of the expected frequencies exceed 5 and all the expected frequencies exceed 1 [13, 5].

4 Experimental Results

4.1 Organization Risk Factors

The SIC 1987 taxonomy contains 83 distinct *major group* codes denoted by the first 2 digits of the SIC classification. Of these, 78 were represented in the classifications of organisations studied. Table 1 presents the results of the case-control study we performed on the sector of the organizations. Because of the space limitations, we only provide the results of the sectors that have the highest and lowest assossiation with targeted attacks. Note that to get solid statistica results higher confidence, every test was repeated five times, and we consider the median value as final outcome, excluding outliers that might result as an artefact of the random sampling.

Positive statistical significance was taken to be if the lower value of the 95% confidence interval was greater than 1.0; negative statistical significance was taken to be if the upper value of the 95% confidence interval was less than 1.0. Using these definitions, 37 of the major group classifications were found to be significantly associated with the set of organisations in the *afflicted* group, with the major group *National Security and International Affairs* showing the strongest association with the targeted attacks. A further 8 major group classifications, as well as the additional group of *Nonclassifiable Establishments* (99) were significantly negatively associated with the *afflicted* group. These categories, which include sectors such as Real Estate, Legal Services, Construction or Agricultural Services, seem even protected from receiving targeted cyber attacks. Yet, it does not mean that organizations in these sectors will never see any targeted attack, however it is much less likely, and if this happens, it is unlikely to be due to their business activity but rather to some other factor.

To make it easier to further process the OR results, we have normalized them using the customary normalization method: $OR_{norm} = (OR - 1)/(OR + 1)$. By doing so, we normalize all OR values in the range $[-1, 1]$, with 0 as neutral value (corresponding to $OR = 1$). The OR_{norm} results for SIC2 sectors are visualized in Fig. 1 along with their respective confidence ranges.

As mentioned earlier, the second organization-based feature we analyze is the size of the organizations. We also wanted to evaluate whether the *organisational size* may be statistically associated with the receipt of targeted attacks. The results of this case control study is visualized in Fig. 2, which shows the normalized OR values for the various size groups along with their respective 95% confidence range. The results indicate that as the common sense suggests the size of the organisation is highly correlated with being at risk to targeted attacks.

While certain results might look intuitive, others can be more surprising. For example, major SIC groups 73 (Business Services) and 15 (Construction) were ranked in our data among the *most frequently* targeted sectors (in terms of absolute numbers), however it does not appear to be significantly at higher risk of attack compared to other categories. This might be due to the size of these categories which may comprise a relatively larger proportion of organizations. Conversely, other categories corresponding to apparently *less targeted* sectors (like the Mining sector) now appear to have very high odds ratio, and may be thus at increased risk of attack. The same holds for the size groups, where smaller organizations (1-250) are by far more numerous and might thus appear as more frequently targeted, however the associated Odds Ratio shows that they are at significantly reduced risk of attack compared to very large companies (5000+).

4.2 Individual Risk Factors

The results for the case-control study of the four individual risk factors are presented in Figure 3 and Figure 4. Some of the results are intuitive; for example, the directors and managers in an organization are at higher risk of being targeted than individual contributors. While the results for number of LinkedIn connections is fairly interesting, the results we obtain with geographical location based features are confusing. The odds-ratio calculation of LinkedIn connections numbers feature shows that employees who have between 1 and 500 connections are at significantly higher risk of being targeted when compared to people who have more than 500 connections. Based on the organizations that we have analyzed, employees based in US, Brazil, and India are at significantly reduced risk of being targeted, however, employees in China, Europe, and Australia are at high risk of being targeted. This is quite surprising. While it is hard to make any reasoning without deeper investigation, the reason for obtaining such results for the location-based feature might be due to the nature of our data collection methodology for the individuals. Note that the analysis we performed on individuals strongly depends on the number of LinkenIn profields we were able to find using the simple heuristic we explained earlier.

4.3 Combined Results

While individual OR and OR_{norm} results provide interesting insights into which risk factors might be associated with targeted attacks, in this Section we propose a

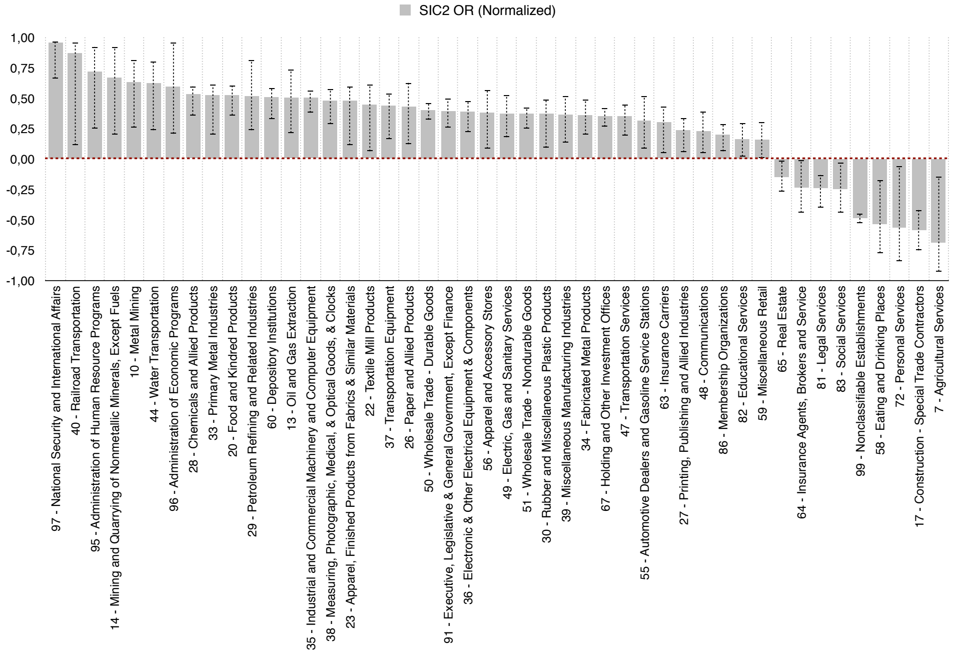


Fig. 1: Normalized Odds Ratios for the major SIC (2 digits) categories. Values above 0.0 refer to industry sectors that are at higher risk of receiving targeted attacks (the higher, the more at risk). Sectors associated with normalized OR lower than 0.0 are protected from such attacks.

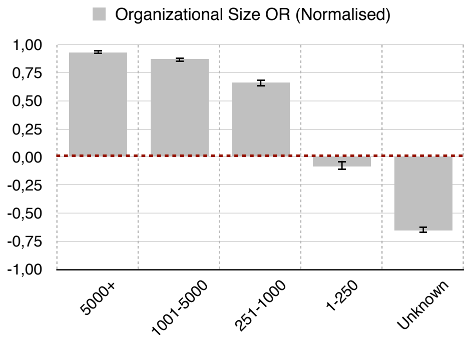


Fig. 2: Normalized Odds Ratios for organization size groups. The risk of receiving targeted attacks increases significantly with the size of the organization.

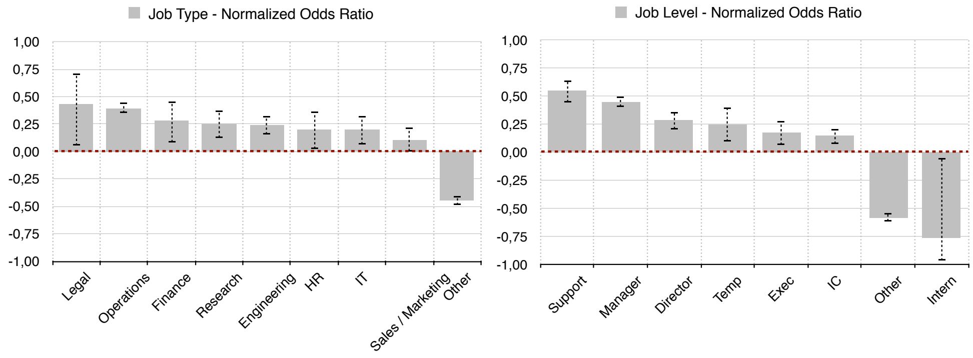


Fig. 3: Normalized Odds Ratios for individual job types and job levels.

straightforward yet powerful technique to combine all odds ratios previously found with respect to individual features.

A simple way to combine all normalized OR values would be to take their average. However, this method has many drawbacks, *e.g.*, it does not take into account the relative importance of each factor, nor their interrelationships. Hence, a smarter and more flexible way of aggregating multiple scores consists in using Multi-Criteria Decision Analysis (MCDA), which provides mathematical tools to define advanced *aggregation models* matching a set of complex requirements (The details of the methodology could be find in the Appendix). In this case, we wanted to assign relative importances to individual OR scores, as well as a fuzzy decision threshold on the amount of high scores required to obtain a global score accurately reflecting a significant high risk of becoming victim of a targeted attack in the near future. For these reasons, we decided to combine all normalized OR values using the *Weighted OWA* (WOWA) operator [34], which can aggregate an input vector by taking into account both the reliability of the information sources (as the weighted mean does),

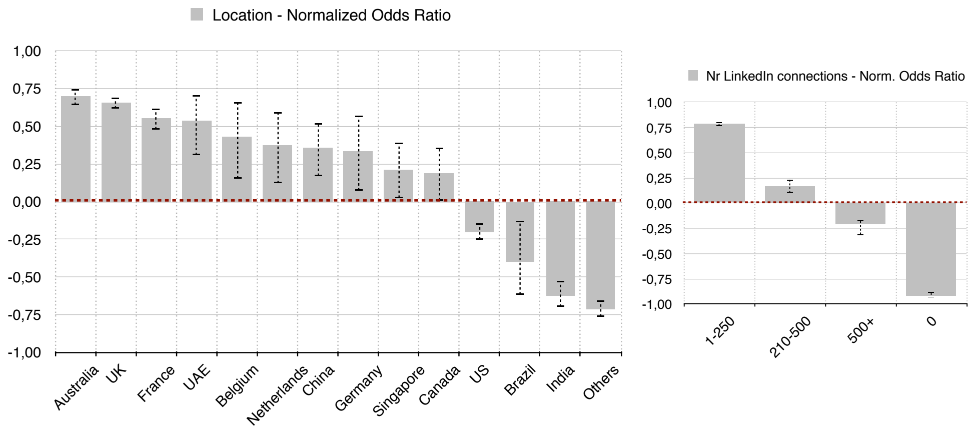


Fig. 4: Normalized Odds Ratios for individual locations and number of LinkedIn connections.

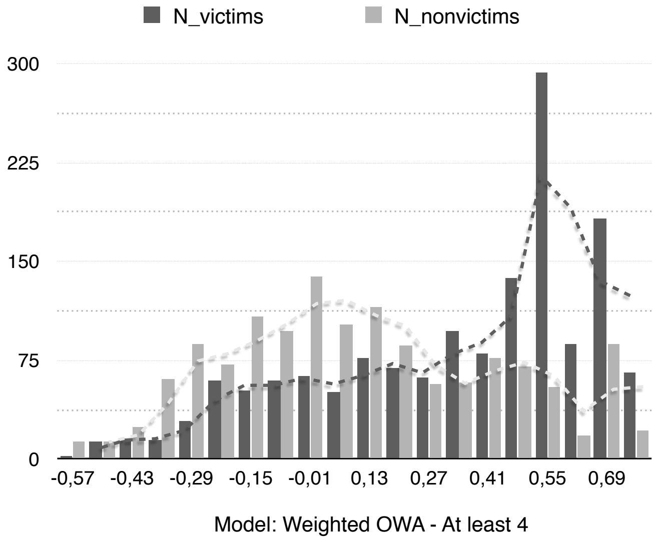


Fig. 5: Combined risk factor distribution for targeted vs non-targeted individuals (Model: Weighted OWA, with at least 4 high risk factors).

and at the same time, by weighting the values in relation to their relative ordering (as the OWA operator).

WOWA makes use of two different weighting vectors: a vector p, which quantifies the relative importances of the different features, and a vector w, which weights the values in relation to their relative *ordering* and allows us to emphasize different combinations of largest, smallest or mid-range values. To define these vectors, we use both our expertise and domain knowledge gained through an in-depth analysis of victim versus non-victim profiles, as well as the characteristics of various statistical distributions of our dataset. For w we computed for every employee the number of odds ratios higher than 1.0, and then compared the distribution of this counting measure for victims and non-victims in our population. It turns out that starting at a count of 4 odds ratio greater than 1.0, the two distributions cross each other, with the number of victims largely exceeding the number of unafflicted customers. Hence, we have set vector w such that it models an aggregation of "at least 4" high scores to obtain a high combined score.

Similarly, by investigating the importance and prevalence of individual risk factors in our population, we have set the components of vector p to the following values:

$$p = [0.32, 0.08, 0.08, 0.12, 0.16, 0.24]$$

with the respective weights corresponding to the following list of features:

[SIC2, org_size, job_type, job_level, location, nr_linkedin_conn]

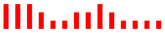
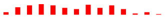
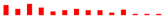
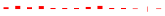
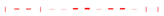
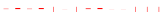
The results of this combined analysis are displayed in Fig. 5, which represents the distribution of combined risk scores for victims and non-victims. We only considered here individuals having complete profiles and belonging to the SIC sectors for which we could obtain statistically significant results. Fig. 5 shows interesting and very promising results, as we can see a clear difference in the distributions in particular starting at combined risk scores above 0.27. By identifying additional features that could be used as risk factors, this combined risk model would probably further improve our capability to truly assess cyber risk, and thus to proactively identify who is at increased risk of attack in the near future based on his/her intrinsic characteristics. Just like for health insurance models, our combined risk model could thus be used to design cyber insurance schemes that accurately reflect real-world risks in cyber space.

4.4 Follow-up study

A case-control study is not designed to test the power of the identified risk factors for predicting future attacks, as this would require instead a full *cohort* study, which requires a significant amount of resources and is beyond the scope of this paper. However, to evaluate the predictive nature of our case-control study, we performed a limited follow-up study examining subsequent attacks in the first Quarter (Jan-Mar) of 2014 by taking the organisational size and a limited set of SIC categories as the only risk factors under consideration. In this follow-up study, we have observed the proportion of targeted organisations (expressed as "1 in x" ratios) among our sample population, the proportion of *newly targeted* organisations that previously belonged to our control group (referred to as the *renewal rate*), and the targeted organisations ratios as observed on a *weekly* basis in 2014-Q1.

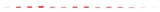
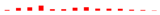
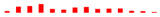
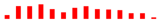
]

Table 2: Follow-up study in 2014 (Q1) on a subset of SIC codes (2 digits)

SIC2	**Category**	**Targeted** **(1 in x)**	**Renewal** **(1 in x)**	**Org./week (%)**
97	National Security and International Affairs	2.4	12.0	
60	Depository Institutions	3.3	8.1	
13	Oil and Gas Extraction	3.4	9.8	
64	Insurance Agents, Brokers and Service	6.9	20.0	
81	Legal Services	17.7	26.5	
65	Real Estate	18.9	54.6	

In Table 3, we note that the observed incidence of targeted attack during 2014-Q1, segmented by organisational size, is consistent with the predictive model. The odds ratios calculated from 2013 data suggest the risk of attack increases with the size of the organisation, and new statistics for 2014 seem to follow the very same trend. Furthermore, the trend line showing the weekly rate of targeted organizations is well-aligned with the predictive model calculated in 2013. Only the renewal rate for size group *5000+* seems to be somehow an outlier (the number of newly targeted companies in this group seems to be significantly smaller), and may thus indicate that attackers have initiated a change in their tactics by targeting more heavily smaller organisations, instead of large multinational companies.

Table 3: Follow-up study in 2014 (Q1) on the Organisation size

Org. size	**Targeted** **(1 in x)**	**Renewal** **(1 in x)**	**Org./week (%)**
1-250	8.2	12.8	
251-1,000	2.8	6.6	
1,001-5,000	1.9	9.0	
5,000+	1.4	16.8	

Finally, Table 2 shows the incidence of targeted attacks in 2014-Q1 for a subset of SIC codes (2-digits). Here too, we observe the predictive model is consistent with subsequent observations: SIC codes identified as being at higher risk of attack in 2013 exhibit much higher proportions of organizations afflicted by new waves of targeted attacks in 2014. Conversely, for SIC categories that had a strong negative statistical significance in 2013 (Table 1), these particular sectors of activity seem to have a protective effect for those organizations, as only a few of them encounter targeted attacks on a weekly basis (which may happen merely by accident, or due to other circumstances perhaps).

5 Conclusion

As demonstrated by recent high-profile and highly publicised attacks against governments and large industries, cyber criminals seem to rely increasingly on more sophisticated malware and targeted threats as an effective means for industrial espionage. While the high profile identification of those threats may be effective in raising awareness of the danger, it does not necessarily help in determining the level of risk that targeted malware may really pose to an organisation. It is thus important to develop tools for security practitioners to assess rigorously the true level of risk to which their organization might be exposed to, *e.g.*, because of the sector of activity, the profitability of the industry, its geographical location, or possibly any other profile characteristic susceptible of being a significant risk factor.

In this paper, we show that these *risk factors* can be effectively determined for different organizations by adapting appropriate techniques from epidemiology. Considering the taxonomy of standard industry classification codes and the organizational size as potential risk factors, we have designed *case-control* studies to calculate odds ratios reflecting the degree of association with the receipt of targeted attack. A validation with a large corpus of targeted attacks blocked by [company name] mail scanning service during the whole year 2013 revealed that certain industry sectors – such as *National Security* and the *Energy* sectors, among others – are statistically at elevated risk compared with others. Similarly, we found that the risk of receiving targeted attacks increases significantly with the organizational size.

The epidemiology techniques used in this study may be further extended to allow the proactive identification of those at increased risk of attack. We believe our study is a first step towards developing a predictive framework for the analysis of targeted threats, where the degree of risk of being attacked may be calculated from a more comprehensive set of relevant factors pertaining to the profile of an organisation, or of the individuals belonging to it. A precise quantification of these risk factors – and more importantly, the combination hereof – will strengthen the epidemiological model and its capability for predicting which specific individuals or companies are the most at risk of being attacked in the near future. This, in turn, will enable organizations to take proactive measures to mitigate or at least control this risk by investing the appropriate level of resources.

References

1. A. J. Alberg, J. G. Ford, and J. M. Samet. Epidemiology of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest*, 132(3 Suppl):29S–55S, 2007.
2. BBC News. Shamoon virus targets energy sector infrastructure. `http://www.bbc.co.uk/news/technology-19293797`, aug 2012.

3. G. Beliakov, A. Pradera, and T. Calvo. *Aggregation Functions: A Guide for Practitioners*. Springer, Berlin, New York, 2007.
4. J. M. Bland and D. G. Altman. Statistics notes: The odds ratio. *BMJ*, 320(7247):1468, May 2000.
5. M. Bland. *An Introduction to Medical Statistics (Oxford Medical Publications)*. Oxford University Press, USA, 2000.
6. A. Bossler and T. Holt. On-line Activities, Guardianship, and Malware Infection: An Examination of Routine Activities Theory. *International Journal of Cyber Criminology*, 3(1):400–420, January-June 2009.
7. Y. Carlinet, L. Mé, H. Debar, and Y. Gourhant. Analysis of Computer Infection Risk Factors Based on Customer Network Usage. In *Proceedings of the 2008 Second International Conference on Emerging Security Information, Systems and Technologies*, SECURWARE '08, pages 317–325, Washington, DC, USA, 2008. IEEE Computer Society.
8. E. Chien and G. O'Gorman. The Nitro Attacks, Stealing Secrets from the Chemical Industry. Symantec Security Response, `http://bit.ly/tDd3Jo`.
9. D. Dagon, C. Zou, and W. Lee. Modeling Botnet Propagation Using Time Zones. In *In Proceedings of the 13th Network and Distributed System Security Symposium (NDSS)*, 2006.
10. D. J. Daley and J. M. Gani. *Epidemic Modeling: An Introduction*. Cambridge University Press, 1999.
11. N. Falliere, L. O. Murchu, and E. Chien. W32.Stuxnet Dossier. `http://www.symantec.com/security_response/whitepapers.jsp`, Feb 2011.
12. J. C. Frauenthal. *Mathematical Modeling in Epidemiology*. Springer-Verlag, 1980.
13. P. E. Greenwood and M. S. Nikulin. *A Guide to Chi-Squared Testing (Wiley Series in Probability and Statistics)*. Wiley, 1996.
14. D. A. Grimes and K. F. Schulz. Compared to what? Finding controls for case-control studies. *Lancet*, 365(9468):1429–1433, 2005.
15. J. Kephart, S. White, and D. Chess. Computers and epidemiology. *Spectrum, IEEE*, 30(5):20 –26, may 1993.
16. J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. In *IEEE Symposium on Security and Privacy*, pages 343–361, 1991.
17. J. O. Kephart and S. R. White. Measuring and modeling computer virus prevalence. In *Proceedings of the 1993 IEEE Symposium on Security and Privacy*, SP '93, pages 2–, Washington, DC, USA, 1993. IEEE Computer Society.
18. M. Lee. Who's Next? Identifying Risk Factors for Subjects of Targeted Attacks. In *22nd Virus Bulletin International Conference*, pages 301–306, Sep 2012.
19. F. L. Levesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji. A clinical study of risk factors related to malware infections. In *Proceedings of the 2013 ACM SIGSAC conference on Computer and communications security (CCS '13)*, 2013.
20. C. J. Mann. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J*, 20(1):54–60, Jan 2003.
21. D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver. Inside the slammer worm. *IEEE Security and Privacy*, 1(4):33–39, July 2003.
22. D. Moore, C. Shannon, and J. Brown. Code-Red: a case study on the spread and victims of an Internet worm. In *Internet Measurement Workshop (IMW) 2002*, pages 273–284, Marseille, France, Nov 2002. ACM SIGCOMM/USENIX Internet Measurement Workshop.
23. D. Moore, C. Shannon, G. Voelker, and S. Savage. Internet quarantine: requirements for containing self-propagating code. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 3, pages 1901—1910, March-April 2003.

24. J. A. Morris and M. J. Gardner. Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J (Clin Res Ed)*, 296(6632):1313–1316, May 1988.
25. Occupational Safety & Health Administration. SIC Manual. http://www.osha.gov/pls/imis/sic_manual.html.
26. P. Porras, L. Briesemeister, K. Skinner, K. Levitt, J. Rowe, and Y.-C. A. Ting. A hybrid quarantine defense. In *Proceedings of the 2004 ACM workshop on Rapid malcode*, WORM '04, pages 73–82, New York, NY, USA, 2004. ACM.
27. K. F. Schulz and D. A. Grimes. Case-control studies: research in reverse. *Lancet*, 359(9304):431–434, Feb 2002.
28. C. Shannon and D. Moore. The spread of the witty worm. *Security Privacy, IEEE*, 2(4):46–50, July-Aug. 2004.
29. S. Staniford, V. Paxson, and N. Weaver. How to Own the Internet in Your Spare Time. In *Proceedings of the 11th USENIX Security Symposium*, pages 149–167, Berkeley, CA, USA, 2002. USENIX Association.
30. Symantec. Stuxnet 0.5:The Missing Link. http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/stuxnet_0_5_the_missing_link.pdf, Feb 2013.
31. Symantec. Internet Security Threat Report Vol. 19. http://www.symantec.com/threatreport/, April 2014.
32. Symantec Security Response. The Luckycat Hackers, White paper. http://www.symantec.com/security_response/whitepapers.jsp.
33. O. Thonnard, L. Bilge, G. O'Gorman, S. Kiernan, and M. Lee. Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In *RAID*, pages 64–85, 2012.
34. V. Torra. The weighted OWA operator. *Int. Journal of Intelligent Systems*, 12(2):153–166, 1997.
35. Trend Micro. Luckycat redux, Inside an APT Campaign with Multiple Targets in India and Japan. Trend Micro Research Paper, http://www.trendmicro.co.uk/media/wp/luckycat-redux-whitepaper-en.pdf, 2012.
36. C. Wang, J. C. Knight, and M. C. Elder. On computer viral infection and the effect of immunization. In *Proceedings of the 16th Annual Computer Security Applications Conference*, ACSAC '00, pages 246–, Washington, DC, USA, 2000. IEEE Computer Society.
37. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Trans. Syst. Man Cybern.*, 18(1):183–190, 1988.
38. C. Zou, D. Towsley, and W. Gong. Email worm modeling and defense. In *Proceedings of the 13th International Conference on Computer Communications and Networks (ICCCN 2004)*, pages 409–414, October 2004.
39. C. C. Zou, W. Gong, and D. Towsley. Code red worm propagation modeling and analysis. In *Proceedings of the 9th ACM conference on Computer and communications security*, CCS '02, pages 138–147, New York, NY, USA, 2002. ACM.
40. C. C. Zou, W. Gong, and D. Towsley. Worm propagation modeling and analysis under dynamic quarantine defense. In *Proceedings of the 2003 ACM workshop on Rapid malcode*, WORM '03, pages 51–60, New York, NY, USA, 2003. ACM.

Appendix A: Detailed Odds Ratio (OR) Results

Organisation size	**Odds**	**Confidence**	χ^2
	ratio	**interval**	**p-value**
5,000+	27.12	20.59 - 35.72	< .001
1,001-5,000	14.13	12.45 - 17.03	< .001
251-1,000	4.90	4.39 - 5.46	< .001
1-250	0.85	0.79 - 0.91	< .001
UNK	0.21	0.18 - 0.23	< .001

Table 4: OR calculated as per *Organisational size*

Job level	**Odds**	**Confidence**	χ^2
	ratio	**interval**	**p-value**
Support Staff	3.46	2.62 - 4.56	< .001
Managers	2.63	2.35 - 2.94	< .001
Directors	1.79	1.51 - 2.13	< .001
Temporary Workers	1.74	1.27 - 2.39	0.007
Executives	1.45	1.16 - 1.82	0.013
Individual Contributors	1.29	1.13 - 1.47	0.003
Others	0.27	0.25 - 0.30	< 0.001
Interns	0.16	0.03 - 0.84	0.099

Job Type	**Odds**	**Confidence**	χ^2
	ratio	**interval**	**p-value**
Legal	2.36	1.08 - 5.16	0.178
Operations	2.23	2.00 - 2.48	< .001
Finance	1.81	1.22 - 2.70	0.033
Research	1.66	1.27 - 2.17	0.002
Engineering	1.61	1.35 - 1.93	< .001
HR	1.69	1.19 - 2.41	0.031
IT	1.47	1.13 - 1.93	0.041
Sales & Marketing	1.25	1.01 - 1.54	0.231
Others	0.38	0.34 - 0.42	< .001

Table 5: OR calculated as per individual *job type* and *job level*.

Location	**Odds**	**Confidence**	χ^2
	ratio	**interval**	**p-value**
Germany	1.91	1.10 - 3.30	0.137
Netherlands	2.27	1.31 - 3.93	0.059
UAE	2.83	1.57 - 5.10	0.004
India	0.23	0.18 - 0.31	< .001
France	3.53	2.90 - 4.28	< .001
China	2.19	1.48 - 3.24	0.001
USA	0.67	0.61 - 0.75	< .001
Brazil	0.48	0.27 - 0.86	0.095
Australia	5.75	4.59 - 7.19	< 0.001
UK	4.74	4.14 - 5.43	< 0.001

Linkedin	**Odds**	**Confidence**	χ^2
connections	**ratio**	**interval**	**p-value**
1-250	8.73	7.83 - 9.73	< .001
251-500	1.40	1.23 - 1.60	< .001
500+	0.62	0.53 - 0.73	< .001
0	0.05	0.04 - 0.06	< .001

Table 6: OR calculated as per individual *location* and *Linkedin connections*.

Appendix B: Combining Odds Ratios using Multi-Criteria Decision Analysis

We use *Multi-Criteria Decision Analysis* (MCDA) to design an aggregation model for the calculation of combined risk scores, taking as input all odds ratio associated with the individual features. A typical MCDA problem consists to evaluate a set of alternatives w.r.t. different criteria using an *aggregation function* [3]. The outcome of this evaluation is a global score obtained with a well-defined aggregation model that incorporates a set of constraints reflecting the preferences and expectations of the decision-maker.

An aggregation function is defined as a monotonically increasing function of n arguments ($n > 1$): $f_{aggr} : [0,1]^n \longrightarrow [0,1]$.

In the family of *averaging* aggregation functions, the *Ordered Weighted Average* (OWA) operator extends these functions by combining two characteristics: (i) a weighting vector (like in a classical weighted mean), and (ii) *sorting* the inputs (usually in descending order). OWA is defined as [37]:

$$OWA_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{n} w_i x_{(i)} =< \mathbf{w}, \mathbf{x}_{\searrow} >$$

where $\mathbf{x}_{\searrow}$ is used to represent the vector $\mathbf{x}$ arranged in decreasing order: $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(n)}$. This allows a decision-maker to design more complex decision modeling schemes, in which we can ensure that only a portion of criteria is satisfied without any preference on which ones precisely (*e.g.*, "at least" k criteria satisfied out of n). OWA differs from a classical weighted means in that the weights are not associated with particular inputs, but rather with their *magnitude*. It can thus emphasize a subset of largest, smallest or mid-range values.

It might be useful sometimes to also take into account the *reliability* of each information source in the aggregation model, like in Weighted Mean (WM). Torra [34] proposed thus a generalization of OWA, called *Weighted OWA* (WOWA). This aggregation function quantifies the *reliability* of the information sources with a vector $\mathbf{p}$ (as the weighted mean does), and at the same time, allows to weight the values in relation to their relative *ordering* with a second vector $\mathbf{w}$ (as the OWA operator). It is defined by [34]:

$$WOWA_{\mathbf{w},\mathbf{p}}(\mathbf{x}) = \sum_{i=1}^{n} u_i x_{(i)},$$

where $x_{(i)}$ is the i^{th} largest component of $\mathbf{x}$ and the weights u_i are defined as

$$u_i = G\left(\sum_{j \in H_i} p_j\right) - G\left(\sum_{j \in H_{i-1}} p_j\right)$$

where the set $H_i = \{j | x_j \geq x_i\}$ is the set of indices of the i largest elements of $\mathbf{x}$, and G is a monotone non-decreasing function that interpolates the points $(i/n, \sum_{j \leq i} w_j)$ together with the point $(0,0)$. Moreover, G is required to have the two following properties:

1. $G(i/n) = \sum_{j \leq i} w_j$, $i = 0, \ldots, n$;
2. G is linear if the points $(i/n, \sum_{j \leq i} w_j)$ lie on a straight line.