Empirical analysis of system-level vulnerability metrics through actual attacks

(In IEEE Transaction on Dependable and Secure Computing)

Mathias Ekstedt, Associate Professor
Industrial Information and Control Systems
KTH Royal Institute of Technology, Sweden

Hannes Holm, KTH Royal Institute of Technology
Dennis Andersson, Swedish Defence Research Agency

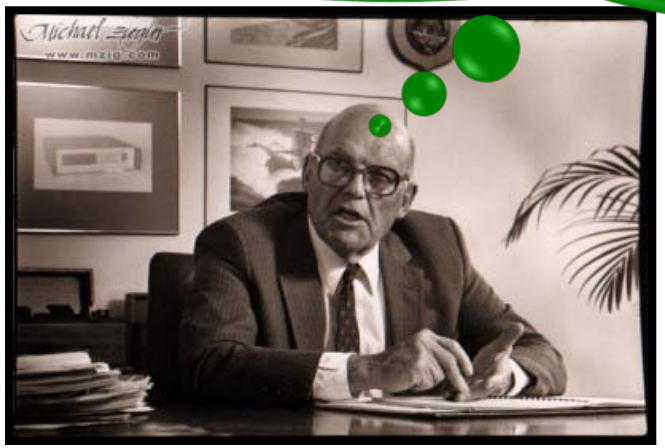
ESP-CSP Workshop - ISRCS 2012

Cyber Security Research at Industrial Information and Control Systems

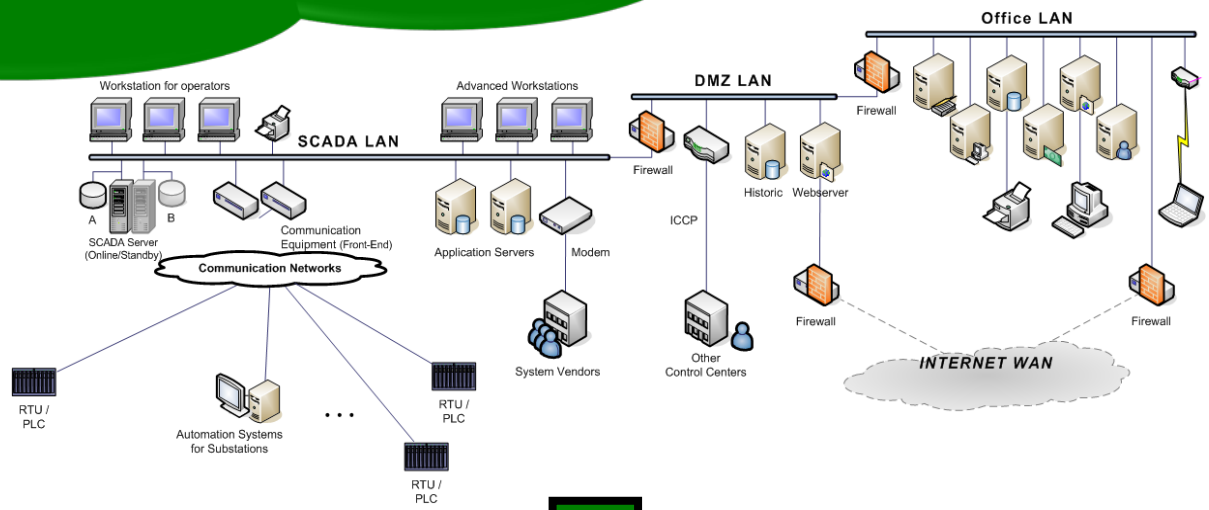
- Research areas
  - Security analysis of enterprise-level information systems architectures (user/customer-side system architectures)
  - In particular for power utilites (i.e. SCADA and substation automation systems, and smart grid architectures)
  - Information Security Managment (security governance and organization)
- Methodolodigal approach
  - Information systems architecture modelling
  - Attack/defense graphs
  - Probabilistic analyses
- Projects/financing
  - EU FP7 project VIKING
  - Swedish Center of Excellence in Electrical Power Engineering
  - European Institute of Innovation and Technology/InnoEnergy
  - Swedish Grid/ Swedish Defence Research Agency

Our research goal:
Cyber security decision-making support

What should I do to increase security?
(to a good enough level, and
as cost-efficient as possible)



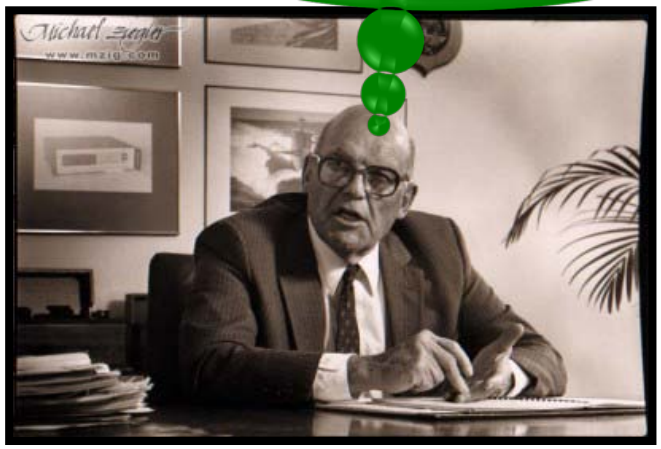
System owner



Cyber Security/Resilience = ?

Security metrics to the rescue?

What about scanning the system for vulnerabilities and use a vulnerability-based metric for governing my decisions?



System owner

Let us find out!

- Study a cyber defense exercise; the Baltic Cyber Shield

- But what is security..?
  - We take Time-To-Compromise (TTC) as the definition
    (Calendar time from start of an attack until successful intrusion.)

- Vulnerability metrics
  - Common Vulnerability Scoring System (CVSS) – Base score
    - (Subjective assessment of) Criticality of ***individual*** vulnerabilities
    - Sub metrics
      - Impact score – ”loss” in terms of CIA
      - Exploitability – ”difficulty” of attacking

BaseScore = (0**,** 6*Impact + 0,4*Exploitability -1,5) * f(Impact)
Impact = 10**,**41*(1-(1- ConfImpact)*(1-IntegImpact)*(1-AvailImpact))
Exploitability = 20*AccessComplexity*Authentication*AccessVector
f(Impact) = 0 if Impact = 0 ; 1,176 otherwise

Metrics aggregating vulnerability to component/system level 1(2)

- Weakest link metrics
  - Single most ”high-scoring” vulnerability
    - Per *system*, measured in Base, Impact, Exploitability CVSS scores
    - Per *service*, measured in Base, Impact, Exploitability CVSS scores and summed up for all services per system
- Amount of vulnerabilities
  - Total number of vulnerabilities per system
    - # CVSS High base score, #Medium, # Low, # in Total
    - # High, #Medium, # Low, # in Total – with available known exploit
- VEA-bility [Tupper and Zincir-Heywood 2008]
  - Vulnerability (V), Exploitability (E), Attckability(A) dimensions
  - Own aggregation per system based on CVSS Impact, Temporal, and Explotability Scores and present network- and attack paths to the system

ROYAL INSTITUTE OF TECHNOLOGY

$$\text{VEA-bility} = 10 - \frac{(V + E + A)}{3}$$

$$\text{V}(system) = \min\left(10, \ln \sum e^{S(v)}\right)$$

$$\text{S}(v) = \frac{\text{Impact Score}(v) + \text{Temporal Score}(v)}{2}$$

$$\text{E}(system) = \frac{\min\left(10, \ln \sum e^{\text{Exploitability Score}(v)}\right) \times hs}{ns}$$

$$\text{A}(system) = \frac{10 \times \sum \text{attack paths}}{\sum \text{network paths}}$$

Metrics aggregating vulnerability to component/system level 2(2)

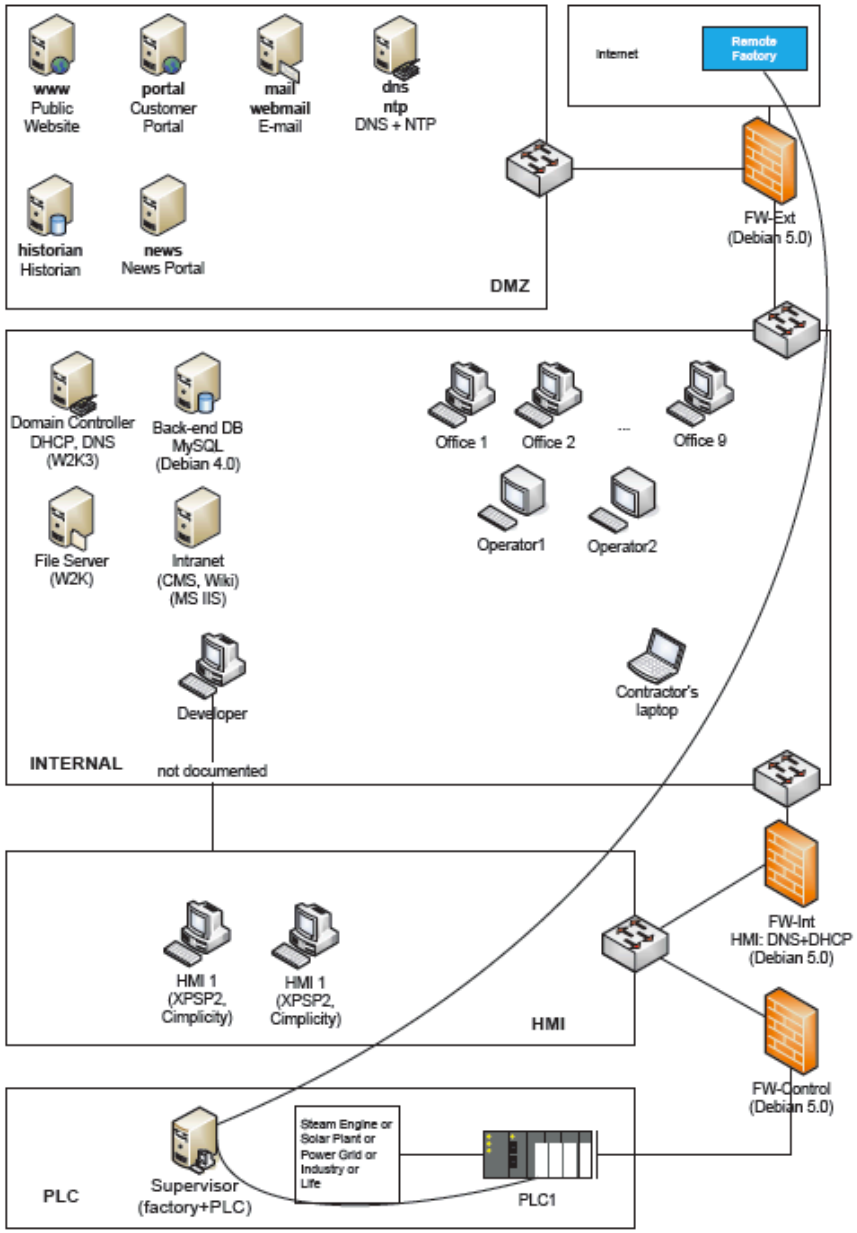
- Vulnerability Exposure [Boyer and McQueen 2008]
  - Aggregated # ”vulnerability days” per system
    - Disclosure dates of CVSS High base score at US NVD.
- Model by Lai and Hsia [2007]
  - Based on an aggregate of CVSS Base score for all vulnerabilities on a system. Weighted by (subjective) threat and asset weight indices
    - Threat and asset weights were set to equal values in the study
- TTC estimation model by McQueen et al. [2006]
  - A random process (probability) model of attack success that depends sub-processes associated with attacker actions aimed at the exploitation of vulnerabilities.
  - Depends on estimations of # vulnerabilities, value of successful attack, attacker skills, # available exploits (in metasploit)

- Only CVSS High base score vulnerabilities were studied
- Attackers were assumed to be “experts”

The Baltic Cyber Shield Exercise

- 10-11th of May 2010
- Managed and run by:
  - Cooperative Cyber Defence Centre of Excellence, Estonia
  - Swedish National Defence College
  - Swedish Defence Research Agency
- Teams
  - 6 blue teams (defenders) with 6-10 persons/ team
  - 1 red team (attackers) with 16 persons
  - 1 white team (exercise management/judges)
  - 1 green team (infrastructure operators)
- Mimic an environment of critical infrastructure operations
- Game scoring according to performance of assigned tasks
- Purpose
  - Train blue/red teams in control system security operations
  - Train green/white teams in running exercises
  - ….In an international setting

Blue teams



- Pre-built networks with 20 physical PC servers running a total of 28 virtual machines / blue team
- Four VLAN segments
  - DMZ, INTERNAL, HMI, PLC
- Set-up as ”somewhat” insecure
- Blue team mission
  - Defend against attacks
  - While keeping business functions operating (e.g. communicating outwards) and “living” by other delimitations (e.g. some systems were not allowed to be patched)

DMZ:

System	Operating System	Services
DNS + NTP	Debian 4.0r1	DNS, FTP, NTP, SSH
E-mail	Debian 4.0r1	HTTP, IMAP, POP3, SMTP, SSH
Customer Portal	Debian 4.0etch/oldstable	HTTP, HTTPS, SMTP, SSH
Public Website	Win Server 2003	FTP, HTTP, HTTPS, RFB, SSH
Historian	Win Server 2000 SP4	FTP, HTTP, RFB, SMTP, SSH
News	CentOS 5.4	HTTP, SSH

Red team

- Professional network penetration testers
- Ended up in four sub-teams depending on competence
  - Client-side
  - Fuzzing
  - Web app
  - Remote
- “Game story”
  - extremist environmental organization demanding green power
- Game/exercise phases
  - “Declaration of war” (publish message on “company” web site)
  - “Breaching the castle wall” (Compromise DMZ)
  - “Owning the infrastructure” (Compromise HMI and PLC)
  - “Wanton destruction” (Go crazy…)

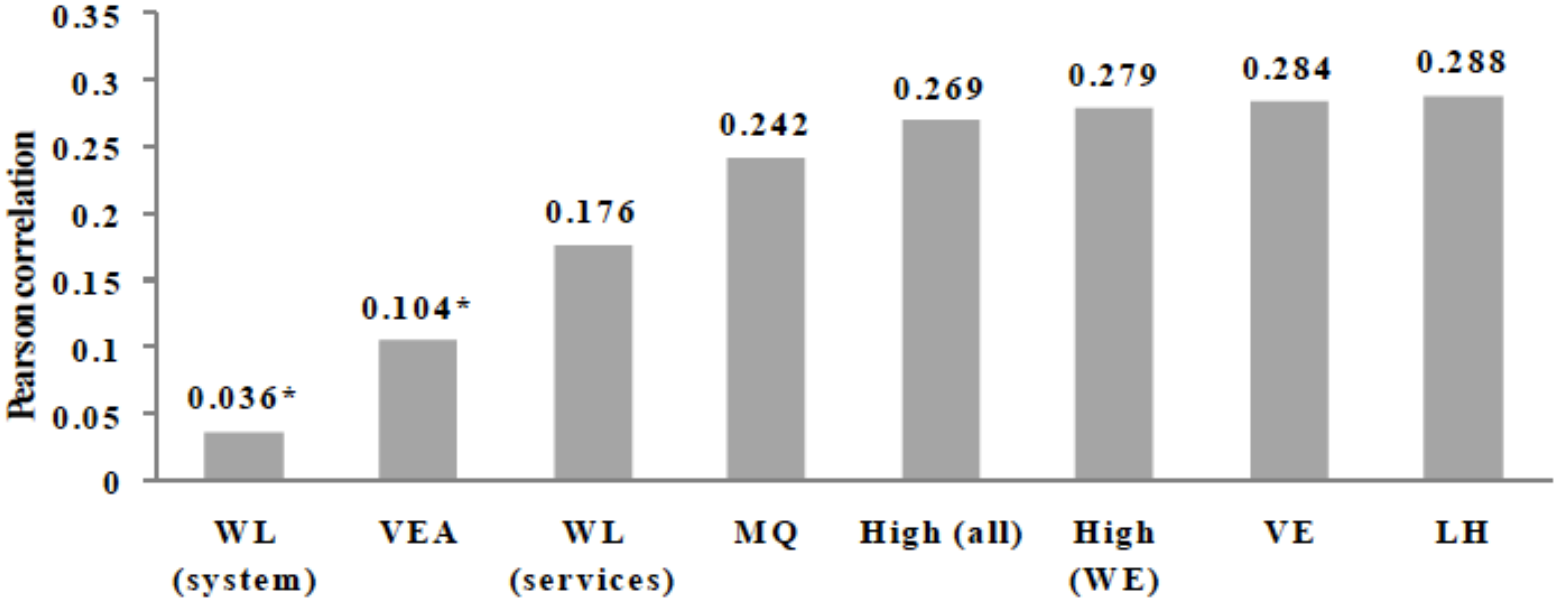
Data Collection

- Data sources: network data, red team attack logs, observer logs, and vulnerabilities
- Data were used from game phases 1, 2, and 3.
- One blue team misunderstood the game rules → no data
- Attacks
  - Only successful compromises were studied
  - All network traffic captured in .pcap files that were run ex-post in Snort. This provided ”attack attempts”
    - → Approx. 3 million alarms…
    - → Only data from attacker MAC-addresses were considered (thus excluding attacks originating from compromised systems)
  - Attack success identified from attacker and observer logs
  - TTC = calendar time from first alarm to compromise
- Vulnerabilities
  - Only data from DMZ compromises (poor blue team mitigation logs and vulnerabilities in other segments were not stable enough)
  - Unauthenticated scan with Nessus provided vulnerability data
    - →remotely reachable vulnerabilities (preferred method for several metrics)
- After ”cleaning”: 34 successful attacks provided the data points for the study

Analysis

- Hypotheses:
  - System security estimated through [METRIC] is positively/negatively correlated with TTC
- Pearson correlation and two-tailed hypothesis testing with SPSS
- Assumptions
  - TTC and [METRIC] are normally distributed random variables
  - TTC values are independent, and [METRIC] scores are independent
  - There is a linear relationship between TTC and [METRIC]
  - TTC and [METRIC] have the properties of a bivariate normal distribution

Results



WL (system): Weakest link on a system level (CVSS base score)
VEA: VEA-bility
WL (services): Weakest link on a service level (CVSS base score)
MQ: Model by McQueen et al.
High (all): All High vulnerabilities
High (WE): High vulnerabilities with available known exploits
VE: Vulnerability exposure, public disclosure dates
LH: Lai and Hsia's model

* Correlation not in the hypothesised direction

Validity and Reliability issues 1(4)

- Exercise does not reflect reality
  - Network was more vulnerable than reality – but that shouldn't stop a good metric from correlating(?)
    VEA-bility does not nuance very vulnerable systems
  - TTC is dependent on attackers' skills – Red team was picked through peer review so they can be viewed as skilled. One week of preparation. – This is obviously not universally generalizable, but neither obviously non-representative.
- Only 6 systems and 34 compromises
  - Again no universally generalizable conclusion, but the systems and compromises are neither obviously non-representative
  - But more data is obviously needed… (all results random variation..?)
- Calendar time and not working time was measured
  - The attacker logs and observer logs suggest that most successful attacks were of high effort so there is probably moderate deviation

ROYAL INSTITUTE OF TECHNOLOGY

Validity and Reliability issues 2(4)

- Logs can be of poor quality
  - Observers and attackers were the same throughout the exercise
  - Observer and attacker logs concur
- Results from the vulnerability scanner contain false negatives and false positives
  - True, they are not without flaws…
  - But in practice there is little room for alternative data collection methods, so it won't help the CVSS and the metrics as such…
- There might have been zero-days (which are not possible to catch with CVSS)
  - At least the vast majority of the exploits run were for CVSS-scored vulnerabilities

Validity and Reliability issues 3(4)

- Bias due to the assumption of normal distribution
  - Other distributions are certainly possible for TTC. But the study covered multiple systems of various vulnerability levels. So, even if single systems have other distributions the set of systems could be normal distributed.
- Bias due to the assumption of independence.
  - How TTC for different combinations of vulnerabilities have not been studied. However, we are not aware of any dependencies
- Bias due to the assumption of linearity
  - Maybe. But no one suggests otherwise…
  - Specifically questionable for the Vulnerability Exposure and VEA-bility
- Bias due to the assumption of a bivariate normal distribution
  - Maybe. But Pearson correlation is still informative about the dergree of dependence

Validity and Reliability issues 4(4)

- Biggest uncertainty source for the study is that we don't have good documentation on what happened in detail.
  - Maybe compromises had nothing to do with vulnerabilities in the particular system but in the neighboring system…? (independence assumption)

ROYAL INSTITUTE OF TECHNOLOGY

Conclusions

- None of the hypotheses could be accepted/corroborated
- If the study can be considered to be “fair enough”
  - Is CVSS a poor data source for system security prediction?
  - Are the studied metrics poor?
    - And what do system metrics really mean in a networked context..? Attacks are not typically single step, but quite complicated
  - (Vulnerabilities are obviously related to TTC)

- A more detailed observation is that metrics using more CVSS information performed better (weakest link was worst, for instance)
  - Scanners typically produce results based vulnerability magnitude, so if attackers are influenced by scanner results they may be ”attracted” to systems with many vulnerabilities...?

ROYAL INSTITUTE OF TECHNOLOGY

Some personal reflections

- The BCS exercise was not at all intended for research. If also research is actively considered as a purpose, much better results can be achieved.
- Much research infrastructure is needed however…
  - logging functionality, sorting/aggregation, visualization,…

- What is a representative attacker in vulnerability-based metrics research?
- This study was only considering “single-system” security, what about “system-of-systems” level security?

ROYAL INSTITUTE
OF TECHNOLOGY

Questions?

ROYAL INSTITUTE OF TECHNOLOGY

References

The study:

- Holm et al., “Empirical analysis of system-level vulnerability metrics through actual attacks” In IEEE Transaction on Dependable and Secure Computing, Preprint, DOI:http://doi.ieeecomputersociety.org/10.1109/TDSC.2012.66

Baltic Cyber Shield description:

- Geers, Kenneth (2010) "Live Fire Exercise: Preparing for Cyber War," Journal of Homeland Security and Emergency Management : Vol. 7: Iss. 1, Article 74. DOI: 10.2202/1547-7355.1780

The metrics:

- W. Boyer and M. McQueen, “Ideal based cyber security technical metrics, for control systems,” Critical Information Infrastructures Security, pp.246–260, 2008.
- M. Tupper and A. Zincir-Heywood, “Vea-bility security metric: A network security analysis tool,” in Availability, Reliability and Security, ARES 08. Third International Conference on. IEEE, 2008, pp.950–957.
- Y. Lai and P. Hsia, “Using the vulnerability information of computer systems to improve the network security,” Computer Communications, vol. 30, no. 9, pp. 2032–2047, 2007.
- M. McQueen, W. Boyer, M. Flynn, and G. Beitel, “Time-to-compromise model for cyber risk reduction estimation,” Quality of Protection, pp. 49–64, 2006.

Time-to-compromise data

System	Samples	Mean	Time-to-compromise		
DNS + NTP	1	02:50:41	02:50:41		
E-mail	6	03:04:06	00:33:23	01:20:19	01:52:04
			02:32:48	04:44:15	07:21:44
Customer Portal	7	02:57:49	01:09:48	01:22:12	02:27:28
			02:51:34	03:24:44	03:50:15
			05:38:41		
Public Website	10	01:36:13	00:00:00*	00:00:11	00:51:26
			01:07:26	01:41:26	01:54:52
			02:20:42	02:36:13	02:41:31
			02:48:20		
Historian	6	02:02:07	00:28:44	00:41:36	00:45:36
			02:22:15	03:40:16	04:14:12
News	4	01:30:23	00:11:29	00:40:19	02:24:28
			02:45:15		

* First alarm by Snort at the time of the successful attack.

ROYAL INSTITUTE OF TECHNOLOGY

Weakest Link Data

System	Vulnerability	WL System B	WL System E	WL System I
DNS + NTP	CVE-2008-0166	7.8	10	6.9
E-mail	CVE-2008-0166	7.8	10	6.9
Customer Portal	CVE-2008-0166	7.8	10	6.9
Public Website	CVE-2006-3439	10	10	10
Historian	CVE-2008-4834	10	10	10
News	*	0	0	0

* No CVE available

System	WL Service B	WL Service E	WL Service I
DNS + NTP	7.8	10	6.9
E-mail	7.8	10	6.9
Customer Portal	7.8	10	6.9
Public Website	37.5	40	36.4
Historian	74.3	78.6	72.8
News	0	0	0

Metric	Correlation	p (two-tailed)
WL System B	0.036	0.838
WL System E	0.166	0.349
WL System I	-0.020	0.913
WL Service B	-0.176	0.320
WL Service E	-0.168	0.342
WL Service I	-0.178	0.312

Vulnerability Exposure data

System	Vulnerability Exposure
DNS + NTP	727
E-mail	727
Customer Portal	727
Public Website	62544
Historian	52299
News	0

Metric	Correlation	p (two-tailed)
Vulnerability exposure	-0.284	0.104

Number of vulnerabilities data

System	High		Medium		Low		Total	
	All	WE	All	WE	All	WE	All	WE
DNS + NTP	2	1	1	0	27	0	30	1
E-mail	4	2	2	1	46	3	52	6
Customer Portal	2	1	6	2	40	1	48	4
Public Website	27	25	31	14	55	4	113	43
Historian	22	16	6	2	85	3	113	21
News	0	0	1	0	18	0	19	0

Metric	Correlation	p (two-tailed)
High vulnerabilities	-0.269	0.124
Medium vulnerabilities	-0.240	0.172
Low vulnerabilities	-0.061	0.733
All vulnerabilities	-0.199	0.260
High vulnerabilities with exploits	-0.279	0.110
Medium vulnerabilities with exploits	-0.240	0.171
Low vulnerabilities with exploits	-0.117	0.509
All vulnerabilities with exploits	-0.264	0.132

VEA-bility data

System	VEA-bility
DNS + NTP	0.73
E-mail	0.73
Customer Portal	0.73
Public Website	0
Historian	0
News	6.67

Correlation	p (two-tailed)
-0.104	0.557

Lai and Hsia's model data

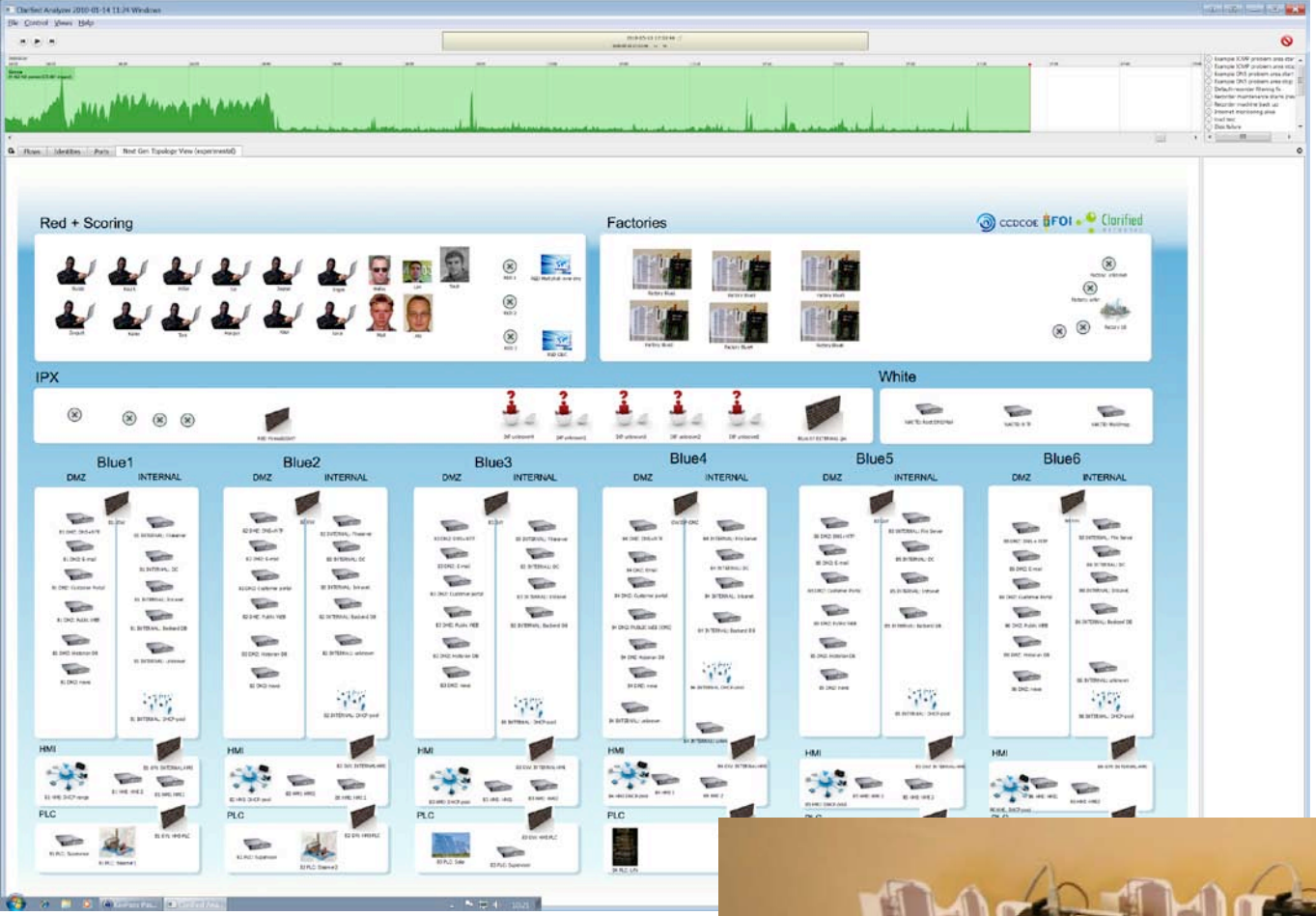
System	LH
DNS + NTP	7.8
E-mail	7.8
Customer Portal	7.8
Public Website	470.7
Historian	254.3
News	0

Pearson	p (two-tailed)
-0.288	0.099

Estimated TTC [McQueen et al.] data

System	Estimated TTC
DNS + NTP	41:31:08
E-mail	39:18:59
Customer Portal	41:31:08
Public Website	14:07:39
Historian	16:36:28
News	46:24:00

Pearson	p (two-tailed)
0.242	0.167


Red + Scoring
Factories
CCDCOE
FOI
Clarified
IPX
White
Blue1
Blue2
Blue3
Blue4
Blue5
Blue6
DMZ
INTERNAL
HMI
PLC


CIE - H10